Memory & Cognition
https://doi.org/10.3758/s13421-024-01538-3

Distributional analyses reveal the polymorphic nature of the Stroop interference effect: It's about (response) time

Léa M. Martinon[1,2,5] · Ludovic Ferrand[2] · Mariana Burca[1,2] · Nabil Hasshim[3] · Dounia Lakhzoum[2] · Benjamin A. Parris[4] · Laetitia Silvert[2] · Maria Augustinova[1,6]

Accepted: 9 February 2024
© The Psychonomic Society, Inc. 2024

Abstract
The study addressed the still-open issue of whether semantic (in addition to response) conflict does indeed contribute to Stroop interference (which along with facilitation contributes to the overall Stroop effect also known as Congruency effect). To this end, semantic conflict was examined across the entire response time (RT) distribution (as opposed to mean RTs). Three (out of four) reported experiments, along with cross-experimental analyses, revealed that semantic conflict was absent in the participants' faster responses. This result characterizes Stroop interference as a unitary phenomenon (i.e., driven uniquely by response conflict). When the same participants' responses were slower, Stroop interference became a composite phenomenon with an additional contribution of semantic conflict that was statistically independent of both response conflict and facilitation. While the present findings allow us to account for the fact that semantic conflict has not been consistently found in past studies, further empirical and theoretical efforts are still needed to explain why exactly it is restricted to longer responses. Indeed, since neither unitary nor composite models can account for this polymorphic nature of Stroop interference on their own, the implications for the current state of theory are outlined.

Keywords Stroop effect · Distributional analyses · Response speed · Response conflict · Semantic conflict

Introduction

In the *Stroop task* (Stroop, 1935), it is particularly difficult to identify the ink color of a word when it denotes a different color (e.g., "BLUE" presented in green ink, hereafter $BLUE_{green}$). In line with Stroop's original idea, this difficulty is largely attributed to the fact that the unintentional reading of *color-incongruent* trials interferes with the identification of their ink color. Despite this consensus, a disagreement persists regarding the processing level(s) at which this *interference* occurs (see, e.g., Henik et al., 2018, and Parris et al., 2019a, 2019b, 2022, for the most recent discussions). Therefore, the present paper aimed to shed some additional light on the locus of this *interference*, considered – along with *facilitation* – as the driving force behind *the overall Stroop effect* (also known as the Congruency effect, see Fig. 1).

Locus versus loci of Stroop interference

The vast majority of models assume that Stroop interference occurs during response selection (e.g., Cohen et al., 1990; Glaser & Glaser, 1989; Logan, 1980; Melara & Algom, 2003;

✉ Léa M. Martinon
lea.martinon@gmail.com

✉ Maria Augustinova
maria.augustinova@univ-rouen.fr

1 Normandie Université, UNIROUEN, CRFDP, Rouen, France

2 LAPSCO, Université Clermont Auvergne, CNRS, Clermont-Ferrand, France

3 Department of Psychology, University of Salford, Salford, UK

4 Department of Psychology, Bournemouth University, Poole, UK

5 Université de Rouen Normandie, Le Centre de Recherches sur les Fonctionnements et Dysfonctionnements Psychologiques (CRFDP, EA 7475) and Université Clermont Auvergne, Laboratoire de Psychologie Sociale et Cognitive (LAPSCO, CNRS UMR 6024), UFR PSSSE - 34 avenue Carnot - TSA 60401, 63001 Clermont-Ferrand Cedex, France

6 Université de Rouen Normandie, Le Centre de Recherches sur les Fonctionnements et Dysfonctionnements Psychologiques (CRFDP, EA 7475), UFR SHS – Bâtiment Freinet, Place Emile Blondel, 76821 Mont Saint Aignan Cedex, France

Published online: 11 March 2024

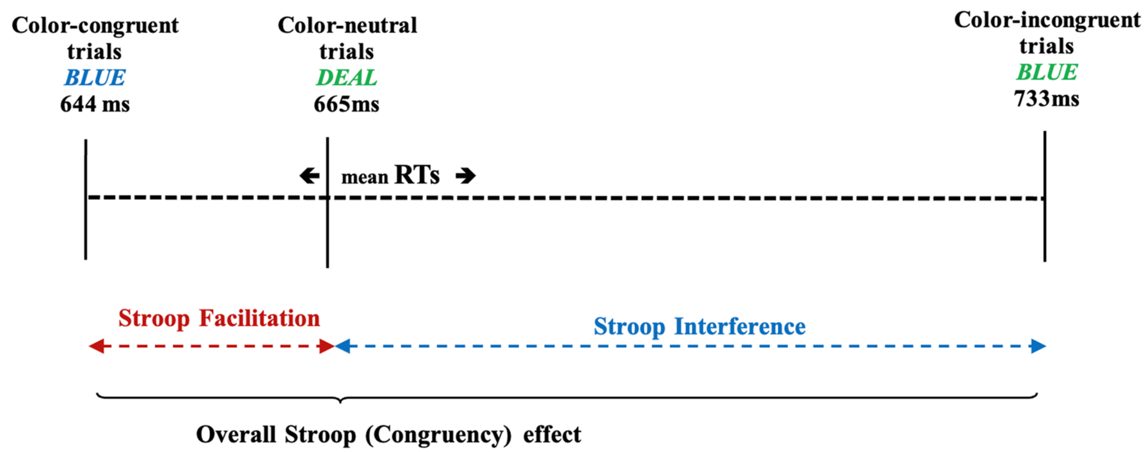

Fig. 1 Diagrammatic representation of how the contribution of Stroop facilitation (i.e., faster mean response times (RTs) to color-congruent than to color-neutral trials) vs. Stroop interference (i.e., slower mean RTs to color-incongruent than to color-neutral trials) to the overall Stroop (Congruency) effect is typically derived from mean RTs (data from Experiment 2 of Augustinova et al., 2019)

Phaf et al., 1990; Roelofs, 2003). Indeed, whenever the irrelevant word dimension of Stroop trials prompts a response that is included in the response set (e.g., blue for $BLUE_{green}$), a conflict between two eligible responses occurs (e.g., blue vs. green for $BLUE_{green}$). As a result of this *response conflict*, the selection of a correct color response for color-incongruent trials is significantly delayed and more prone to errors than the one for color-neutral trials (e.g., $DEAL_{green}$).

Only one model (Zhang & Kornblum, 1998; Zhang et al., 1999) proposes that the unintentional reading of color-incongruent words causes an additional conflict to occur at the level of the stimulus. The assignment of two color-responses to the same response-key (e.g., responses to blue- and red-colored stimuli to 'F' key and those to green- and yellow-colored stimuli to 'J' key) allowed De Houwer (2003) to test this alternative in a particularly elegant way. Indeed, this stimulus-response mapping causes some color-incongruent trials (e.g., $BLUE_{green}$) to prompt eligible responses toward two different response keys (hence termed *different-response trials*), thereby causing response conflict. Other color-incongruent trials (e.g., $BLUE_{red}$), however, do not generate this type of conflict, as the two eligible responses converge toward the same response-key ('F' here). Therefore, according to one-conflict (or *unitary*) models, these *same-response trials* should not produce any interference. Conversely, Zhang and colleagues' two-conflict model still expects these same-response trials to interfere, since it assumes *stimulus* or *semantic conflict* to occur whenever two closely related semantic representations are simultaneously active (i.e., two color-concepts here).

In line with the idea that different-response trials generate both semantic and response conflict, whereas same-response trials only generate semantic conflict, the participants in De Houwer's (2003) study responded more slowly to different- than to same-response trials (e.g., reaction times (RTs) to $BLUE_{green}$ > RTs to $BLUE_{red}$). They also responded more slowly to both of these color-incongruent trials than to baseline trials (e.g., RTs to $BLUE_{red}$ > RTs to $BLUE_{blue}$). Consequently, the *two-to-one Stroop paradigm* has become a popular way of distinguishing the contribution of semantic conflict from the one of response conflict (e.g., Chen et al., 2011, 2013; Chen et al., 2013; Hershman & Henik, 2019, 2020; Jiang et al., 2015; Šaban & Schmidt, 2021; Schmidt et al., 2018; Schmidt & Cheesman, 2005; Shichel & Tzelgov, 2018; Van Veen & Carter, 2005).

It is, however, important to understand that in all of these studies – including De Houwer's (2003) original study – the interference induced by same-response trials was measured against *color-congruent* (or identity) trials (e.g., $BLUE_{blue}$) that are known to produce *facilitation* (i.e., faster responses to color-congruent than to *color-neutral* trials, e.g., $DEAL_{green}$; see, e.g., Dalrymple-Alford & Budayr, 1966, for the first demonstration; see Figs. 1 and 2 for graphical representations). Since the one-conflict (or unitary) models often consider facilitation to be the flip side of interference,[1] they can easily explain the abovementioned positive difference in RTs between same-response and color-congruent items as resulting from *facilitation* on color-congruent trials rather than from *interference* on same-response trials (i.e., prima facie evidence for semantic conflict).

To test this alternative interpretation, Hasshim and Parris (2014) used an additional baseline consisting of color-neutral

[1] This is because facilitation and interference are considered as underpinned by a common mechanism (i.e., converging vs. diverging information from the word and its color; see, e.g., Cohen et al., 1990; Roelofs, 2003; but see, e.g., T. L. Brown, 2011, for arguments against this idea; see also Parris et al., 2022, for a discussion).

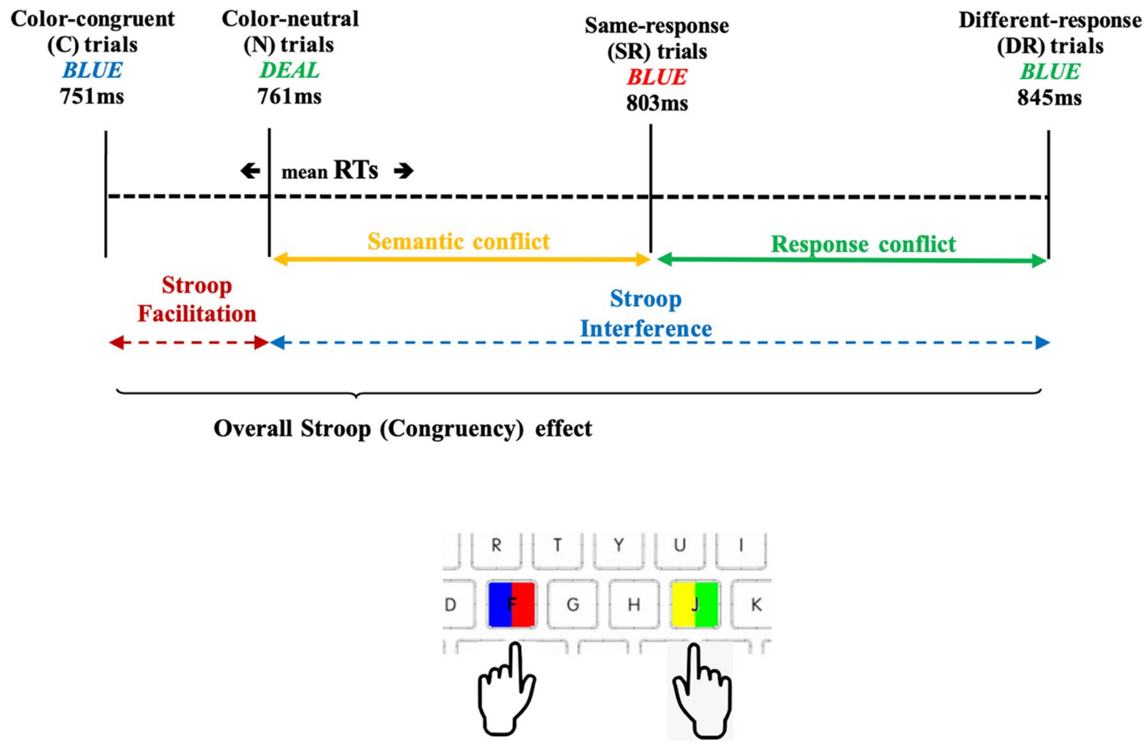

Fig. 2 Diagrammatic representation of how the contribution of semantic conflict (i.e., faster mean response times (RTs) to same-response than to color-neutral trials) to Stroop interference effect – contribution that is independent of both response conflict (i.e., faster mean RTs to different-response than to same-response trials) and of Stroop facilitation – is derived from mean RTs collected in the two-to-one Stroop paradigm with an additional color-neutral baseline (data from Burca et al., 2022, collapsed across older and younger participants)

word trials (e.g., $\textit{DEAL}_{\text{green}}$) that are free of facilitation (e.g., T. L. Brown, 2011; MacLeod, 1991). They reported longer RTs for same-response trials than for color-congruent trials, but no difference between same-response trials and additional color-neutral trials, which clearly runs counter to the contribution of semantic conflict in the two-to-one Stroop paradigm (see also RTs from pupillometric studies of Hasshim & Parris, 2015, and of Hershman & Henik, 2020). As a result of this line of research and because of difficulties associated with alternative ways to induce semantic conflict (see *Present study* section below), Parris et al. (2022) argued that it has not been shown to be an independent form of conflict in the Stroop task.

However, this conclusion does not consider results of Burca et al. (2022) published later that year. Indeed, in complete contrast to Hasshim and Parris (2014, see also Hasshim & Parris, 2015; Hershman & Henik, 2020), they were able to isolate a robust contribution of semantic conflict (same-response – color-neutral trials; e.g., RTs to $\textit{BLUE}_{\text{red}}$ > RTs to $\textit{DEAL}_{\text{red}}$) to the overall Stroop effect (different-response – color-congruent trials; e.g., $\textit{BLUE}_{\text{green}}$ > RTs to $\textit{BLUE}_{\text{blue}}$). Importantly, this contribution was clearly independent of both response conflict (different-response – same-response trials; e.g., RTs to $\textit{BLUE}_{\text{green}}$ > RTs to $\textit{BLUE}_{\text{red}}$) and of facilitation (color-congruent – color-neutral trials; e.g., RTs to $\textit{BLUE}_{\text{blue}}$ < RTs to $\textit{DEAL}_{\text{blue;}}$; see Fig. 2 for a graphical representation of these results). Since unitary models can account for facilitation produced by color-congruent trials but not interference produced by same-response trials (see above), they seem unable to account for Burca et al.'s (2022) findings.

However, in addition to the fact that these latter findings were observed in a single experiment and therefore need to be replicated, it remains unclear why Burca et al. (2022) observed semantic conflict in the two-to-one Stroop paradigm, when other studies did not (Hasshim & Parris, 2014, see also RTs from pupillometric studies of Hasshim & Parris, 2015; and of Hershman & Henik, 2020). Consequently, the aim of the present study was to address these issues.

Present study

To this end, the present study was designed to inspect the contribution of semantic conflict across the entire range of RTs (as opposed to mean RTs reported in all the studies outlined above). Indeed, distributional analyses have been instrumental in documenting that the magnitude of the overall Stroop effect increases proportionally to the slow-down in response speed (see, e.g., Faust et al., 1999; Jackson & Balota, 2013; Spieler et al., 1996), and that this proportional increase results from growing contributions of both interference and facilitation (Roelofs, 2010). Given this relationship between response speed and magnitudes of the overall Stroop effect and of its components (i.e., interference vs. facilitation), it seems plausible that larger magnitudes of Stroop interference in slower responses could be due to the contribution of an additional (i.e., semantic) conflict.

This is precisely what several – more recent Stroop studies – suggest. In fast responses, Scaltritti and colleagues (2022) found no evidence of semantic conflict induced by semantic-associates (e.g., the word SKY that is associated with blue) presented in an incongruent color (SKY_{green}). Semantic conflict was, however, found in slower responses and increased proportionally as the speed of participants' responses slowed down (see also Labuschagne & Besner, 2015; Sulpizio et al., 2022). Similarly, Hasshim et al. (2019) only reported semantic conflict – induced by non-response set incongruent trials (e.g., $PURPLE_{\text{green}}$) – with the slowest responses, whereas response conflict contributed over a wider range of RTs. While taken together these studies suggest that semantic conflict could indeed be restricted to slower responses, it is important to note that only Hasshim and colleagues (2019) had actually induced and measured both semantic and response conflict (as opposed to Labuschagne & Besner, 2015, Scaltritti et al., 2022, and Sulpizio et al., 2022, who induced and measured semantic conflict alone). More importantly, since it has been induced by semantic-associates (SKY_{green}) or non-response set trials ($PURPLE_{\text{green}}$), neither study has unambiguously demonstrated that what varies across the range of RTs is semantic (and not response) conflict. Indeed, it has been argued – including by ourselves – that since these items prompt irrelevant responses that are not part of the response-set (e.g., sky and purple), they do not generate response conflict (e.g., Augustinova & Ferrand, 2014; Hasshim et al., 2019; Quétard et al., 2023). However, this is not the point of view shared by unitary models. In Roelofs' model (2003), for instance, the amount of a single (i.e., response) conflict is determined by the strength of semantic connections between the irrelevant responses and the response-set colors. Because these connections undeniably exist for both semantic-associates and non-response set trials (unlike for color-neutral ones), they are still expected to interfere at the response (as opposed to stimulus) level. Consequently, the presence of semantic conflict solely in the tail of RT distribution still remains to be demonstrated with trials that are unambiguously free of response conflict – as is the case for same response trials. Indeed, for these trials, the two eligible responses converge toward the same response-key, thereby removing the conflict that arises when selecting between response effectors (De Houwer, 2003; see also above and Fig. 2). If these trials do indeed induce semantic conflict – as anticipated by the two-conflict model (Zhang et al., 1999; Zhang & Kornblum, 1998) – it should be found in slow as opposed fast responses (Hasshim et al., 2019; Labuschagne & Besner, 2015; Scaltritti et al., 2022; Sulpizio et al., 2022). This a priori expected pattern of results – putting the emphasis on the role of response speed – would then explain why semantic conflict induced by same-response trials has not been systematically found in past studies considering *mean* RTs only. Indeed, since distributional analyses provide information that is not necessarily reflected in analyses of mean RTs (see above and see Balota & Yap, 2011, for a general discussion of this issue), the presence of semantic conflict in the tail of the RT distribution is likely to be detected independently of whether it is (Burca et al., 2022; see also Burca et al., 2021) or is not reflected in mean RTs (Hasshim & Parris, 2014; see also RTs from pupillometric studies of Hasshim & Parris, 2015, and Hershman & Henik, 2020).

Experiment 1

Method

To examine the abovementioned hypothesis, the existing data from two unpublished experiments were merged. The first (with 87 participants) used a single within-participant factor (i.e., Stimulus-Type: different-response vs. same-response vs. neutral vs. congruent) for data collection. The initial design also included a within-participants variation in Response Stimuli Intervals (RSIs; 2,000 ms vs. 200 ms) that was administered in a fixed order such that the first block with a 2,000-ms RSI was always presented first. The data from this first block (including randomly intermixed 96 different-response (DR) trials, 48 same-response (SR)[2] trials, 48 color-neutral word (N) trials, and 48 color-congruent (C) trials; see Fig. 2) were merged with those from another experiment conducted under virtually identical conditions (see *Apparatus, stimuli, and procedure* section below) with 82 participants. Together, the data of 169 French-speaking psychology undergraduates (142 females and 27 males; M_{age} = 19.97 years; SD = 4.14) from Université Clermont Auvergne, all volunteers with normal or corrected-to-normal color vision,[3] were analyzed in the present experiment. This

[2] If each color-word is presented in all the incongruent colors an equal number of times (to control for contingency), it generates twice as much DR than SR trials.

[3] To ensure that this was the case, the participants were asked to name the color of four colored stickers, as color-blindness, along with uncorrected normal version and late acquisition of French, constituted the three exclusion criteria.

sample size therefore largely exceeded a total sample size of 13 participants recommended by G*Power (Faul et al., 2009) to detect the 4 (Stimulus-Type: different-response vs. same-response vs. neutral vs. congruent) × 5 (Response-speed: Bin 1–5) interaction in fully within-participants ANOVA (for an effect size of 0.25, power of 0.95, and type 1 error rate of 0.05; i.e., the interaction effect size found by similarly designed studies) that was used to further analyze all the distributions of RTs. Five levels of the latter quasi-experimental variables – resulting from the fact that each participant's RT data were sorted into five quantiles or bins, ranging from their fastest to their slowest responses – were based on Scaltritti et al. (2022). In this work, five bins for each condition were computed – each comprising about 1,296 observations. Merging the two data sets enabled the present study to exceed this number as recommended by Brysbaert and Stevens (2018).[4] Therefore, the analyses conducted in the present experiment can be considered as appropriately powered.

Apparatus, stimuli, and procedure

Participants were seated approximately 50 cm from a 15-in. color monitor. For 87 of the participants, E-prime 2.0 software (Psychology Software Tools, Pittsburgh, PA, USA) was used for data presentation and recording, while for the other 82, PsychoPy software (Pierce, 2007) was used for data presentation and recording. The participants were instructed to identify the color of the stimulus presented in the center of the screen as quickly and accurately as possible by pressing the appropriate color button and ignoring everything else in the display. They were asked to stare at a fixation cross ("+"), which appeared in the center of the screen before each trial and remained there for 2,000ms. The word stimulus then appeared and remained on the screen until the participants responded or until 3,500 ms had elapsed.

Stimuli were presented in lowercase, 18-pt Times New Roman bold font on a black background if displayed using E-prime or on a gray background if displayed using PsychoPy, and subtended an average visual angle of 0.9° high × 3.0° wide. They consisted of four color-words: rouge [red], jaune [yellow], bleu [blue], and vert [green]; and four non-color counterparts: plomb [lead], liste [list], page [page], cave [basement], that were paired in length and frequency via Lexique 3.38 (New et al., 2004). Participants answered manually on an AZERTY keyboard; the "D" key was used for "blue" and "red" responses and the "K" key for "green" and "yellow" responses. Color stickers were placed on the response keys as a reminder of the assigned colors for 87 of the participants but not for the other 82.

Two training blocks were administered to familiarize participants with the response set colors and their locations on the keyboard. The first consisted of 48 strings of asterisks (***) or hashes (###) that were randomly presented in the four response set colors. For 87 of the participants only, this first block was followed by a second block comprising 12 trials consisting of the four color words and three color-neutral words (i.e., balcon [balcony], pont [bridge], and chien [dog]) that were presented in the response set colors in such a way that participants encountered each condition of the Stimulus-Type factor (i.e., different-response vs. same-response vs. neutral vs. congruent) three times. The accuracy rate for all the practice blocks was over 80% for all participants. Altogether, it took about 25 min to complete the entire experiment.

Results and discussion

Considering correct trials only, a cut-off of 3 *SD*s above or below mean latency for each of the four conditions of Stimulus-Type factor in each participant was applied to the RTs, and all RTs under 200 ms (i.e., three trials) were removed (i.e., a total of 2% of the data were excluded). RTs observed for each participant as a function of Stimulus-Type were then ordered from fast to slow trials and subsequently analyzed in a 4 (Stimulus-Type: different-response vs. same-response vs. neutral vs. congruent) × 5 (Response-speed: Bin 1-5) ANOVA (see Table 1 for descriptive statistics and Online Supplementary Materials (OSM) including Table S1 for additional results). These analyses revealed the main effect of Stimulus-Type [$F(3,504) = 116.00$, $p < .001$, $\eta_p^2 = 0.408$, $BF_{10} = 1.09e+6$ (i.e., $BF_{10} > 100$)[5]], Response-speed [$F(4,672) = 1532.91$, $p < .001$, $\eta_p^2 = 0.901$, $BF_{10} = +\infty$], and Stimulus-Type × Response-speed interaction [$F(12,2016) = 75.21$, $p < .001$, $\eta_p^2 = 0.309$, $BF_{incl} > 1$ for the interaction alone (i.e., BF_{10} for the interaction and both main effects divided by BF_{10} of both main effects); see Table S2 in OSM for models generated by JASP 0.14.1.0 ; JASP Team, 2017[6]].

[4] Brysbaert and Stevens (2018) recommend 1,600 trials, which reinforces our decision of using five bins (as in Scaltritti et al., 2022; see also Sulpizio et al., 2022) and not more (see, e.g., Hasshim et al., 2019; Labuschagne & Besner, 2015).

[5] BF_{10} corresponds to the Bayesian probability of the occurrence of a hypothesis (H1) and the likelihood of another null hypothesis (H0). It was calculated with JASP 0.14.1.0 (JASP Team, 2017, default priors were used to this end) and interpreted according to Lee and Wagenmakers (2014, adjusted from Jeffreys, 1961).

[6] One obvious limitation of JASP is that it does not generate all the models. Thus, to obtain a Bayes factor for the interaction alone – so the direct comparison with standard frequentist ANOVA can be established – BF_{10} value for the interaction and both main effects needs to be divided by BF_{10} of both main effects. This additionally implies that whenever the evidence in favor of the main effects is close to infinity, the Bayesian evidence in favor of the Stimulus-Type × Response-speed interaction alone (without main effects) cannot be estimated more precisely than reported above (i.e., $BF_{incl}>1$).

Table 1 Color-identification performance (mean response times, standard errors, and 95% confidence intervals) observed as a function of stimulus- or effect-type and response speed in Experiment 1

	Bin 1		**Bin 2**		**Bin 3**		**Bin 4**		**Bin 5**	
Stimulus-type	M (SE)	CI	M (SE)	CI	M (SE)	CI	M (SE)	CI	M (SE)	CI
Different-Response (DR) Color-Incongruent Items	428 (6)	[417, 440]	527 (8)	[511, 543]	616 (10)	[595, 636]	738 (14)	[711, 766]	1005 (18)	[969, 1041]
Same-Response (SR) Color-Incongruent Items	427 (6)	[415, 438]	513 (7)	[498, 528]	592 (9)	[573, 610]	700 (13)	[675, 725]	946 (17)	[912, 980]
Color-Neutral items	429 (5)	[419, 439]	513 (7)	[500, 526]	585 (9)	[568, 602]	678 (11)	[656, 700]	894 (15)	[865, 923]
Color-Congruent items	419 (5)	[409, 429]	499 (7)	[486, 513]	568 (9)	[551, 586]	658 (11)	[636, 681]	872 (15)	[842, 902]
Magnitudes of effects	Diff. (SE)	CI	Diff. (SE)	CI	Diff. (SE)	CI	Diff. (SE)	CI	Diff. (SE)	CI
Overall Stroop (Congruency) Effect (DR – Congruent)	$10^{\dagger / \dagger\dagger}$ (2)	[6, 14]	$28^{\dagger / \dagger\dagger}$ (3)	[22, 34]	$47^{\dagger / \dagger\dagger}$ (4)	[39, 55]	$80^{\dagger / \dagger\dagger}$ (6)	[69, 91]	$133^{\dagger / \dagger\dagger}$ (8)	[117, 149]
Facilitation Effect (Congruent – Neutral)	$-10^{\dagger / \dagger\dagger}$ (2)	[-14, -6]	$-14^{\dagger / \dagger\dagger}$ (2)	[-18, -9]	$-17^{\dagger / \dagger\dagger}$ (3)	[-22, -11]	$-20^{\dagger / \dagger\dagger}$ (4)	[-28, -11]	$-22^{** / *}$ (8)	[-37, -7]
Stroop Interference Effect (DR – Neutral)	$-1^{ns / ns}$ (2)	[-4, 3]	$14^{\dagger / \dagger\dagger}$ (3)	[9, 19]	$31^{\dagger / \dagger\dagger}$ (3)	[24, 37]	$60^{\dagger / \dagger\dagger}$ (5)	[50, 71]	$111^{\dagger / \dagger\dagger}$ (9)	[94, 128]
Response Conflict (DR – SR)	$2^{ns / ns}$ (2)	[-3, 6]	$14^{\dagger / \dagger\dagger}$ (3)	[8, 20]	$24^{\dagger / \dagger\dagger}$ (3)	[17, 31]	$38^{\dagger / \dagger\dagger}$ (5)	[29, 48]	$59^{\dagger / \dagger\dagger}$ (7)	[46, 72]
Semantic Conflict (SR – Neutral)	$-2^{ns / ns}$ (2)	[-7, 2]	$0^{ns / ns}$ (3)	[-6, 6]	$7^{* / ns*}$ (3)	[0, 13]	$22^{\dagger / \dagger\dagger}$ (5)	[13, 31]	$52^{\dagger / \dagger\dagger}$ (8)	[35, 69]

Standard frequentist inference (presented before the slash used as separator): ^{ns}non-significant; *significant at $p < .05$; **significant at $p < .01$; †significant at $p < .001$; Bayesian inference (presented after the slash used as separator): ^{ns}no evidence of an effect, with BF_{10} value between 0–1; ^{ns*}anecdotal evidence in favor of H1, with BF_{10} value between 1–3; *moderate evidence in favor of H1, with BF_{10} value between 3–10; **strong evidence in favor of H1, with BF_{10} value between 10–30; †very strong evidence in favor of H1, with BF_{10} between 30–100; ††extreme evidence in favor of H1/of an effect, with $BF_{10} < 100$

To decompose this interaction, we first computed the simple main-effect of Stimulus-Type for each Bin (see *Results* in OSM, pp.2–3) and a Bayesian one-tailed t-test for pairwise comparisons was further applied, whenever sensible (see Table 1 for all such comparisons). The overall Stroop effect (i.e., contrast between Different Response and Congruent trials, M_{DR}–M_C) emerged clearly in Bin 1 (+10 ms, $p < .001$, $BF_{10} = 5002.06$) but was entirely driven by facilitation (i.e., contrast between Neutral and Congruent trials, M_N–M_C; -10 ms, $p < .001$, $BF_{10} = 44921.54$, see Table 1). In Bin 2, overall Stroop effect (M_{DR}–M_C $p < .001$, $BF_{10} = 2.27e+13$) (i.e., $BF_{10} > 100$) was driven by both facilitation (-14 ms, p

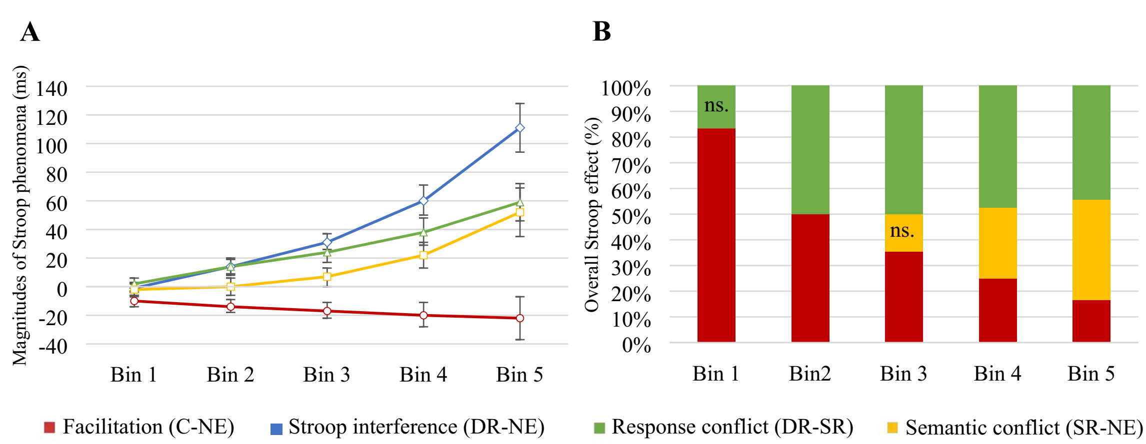

Fig. 3 Magnitudes of each Stroop effect (**panel A**) and composition of the overall Stroop (congruency) effect in percentages (**panel B**) observed as a function of response speed in Experiment 1. Error bars in panel A represent 95% confidence intervals; Sum of Response and Semantic conflicts represented in panel B corresponds to Stroop interference

< .001, BF_{10} = 951117.25) and interference (i.e., contrast between Different Response and Neutral trials, M_{DR}–M_N; +14 ms, $p < .001$, BF_{10} = 65228.90), and this remained the case up to Bin 5. As can be seen in Table 1 and Fig. 3, in Bins 2 and 3, Stroop interference (M_{DR}–M_N) was entirely driven by a unique (i.e., response) conflict (i.e., contrast between Different and Same Response trials, M_{DR}–M_{SR}), as the contribution of semantic conflict (i.e., contrast between Same Response and Neutral trials, M_{SR}–M_N) started to be reliable from Bin 4 (+22 ms, $p < .001$, BF_{10} = 5476.40). Finally, as Table 1 shows (see also Fig. 3), Stroop interference (M_{DR}–M_N) continued to reveal its composite nature in Bin 5. In sum, taking the entire distribution of RTs into account revealed a more complex pattern of results than the one usually reflected in mean RTs. As expected, in slower responses, the significant contribution of semantic conflict was indeed independent of both response conflict and of facilitation, whereas in faster responses, semantic conflict failed to contribute significantly. Still, given that this pattern was observed in a single experiment that additionally merged two pre-existing data sets, the following experiment attempted to replicate this pattern further in a study that was a priori designed to this end.

Experiment 2

Method

One hundred English-speaking participants (65 females and 35 males, M_{age} = 26.07 years, SD = 4.53), including 45 students, were recruited online using Prolific (www.prolific.co) and received £5.50 for their participation. All participants reported having normal or corrected-to-normal color vision, being right-handed, and not having been diagnosed with any language-related disorder. The study used four levels of Stimulus-Type (different-response vs. same-response vs. neutral vs. congruent) for data collection. Again, due to the planned distributional analysis, and as in Experiment 1, the aforementioned sample size largely exceeded the minimum sample size of 13 participants recommended by G*Power (Faul et al., 2009) to detect the 4 (Stimulus-Type) × 5 (Response-speed) interaction in fully within-participants ANOVA (for an effect size of 0.25, power of 0.95, and type 1 error rate of 0.05). Given that this analysis was based on at least 2,880 observations per Stimulus-condition in each of the five bins (see *Results* section), it can be considered as appropriately powered.

Table 2 Color-identification performance (mean response times, standard errors and 95% confidence intervals) observed as a function of stimulus- or effect-type and response speed in Experiment 2

	Bin 1		Bin 2		Bin 3		Bin 4		Bin 5	
Stimulus-type	M (SE)	CI	M (SE)	CI	M (SE)	CI	M (SE)	CI	M (SE)	CI
Different-Response (DR) Color-Incongruent Items	478 (7)	[463, 492]	584 (10)	[564, 604]	673 (13)	[648, 698]	796 (16)	[764, 828]	1107 (23)	[1061, 1153]
Same-Response (SR) Color-Incongruent Items	472 (7)	[458, 487]	565 (9)	[546, 584]	643 (12)	[620, 667]	754 (15)	[723, 784]	1037 (22)	[993, 1082]
Color-Neutral items	473 (7)	[458, 487]	569 (10)	[550, 589]	649 (12)	[625, 673]	756 (15)	[726, 787]	1033 (23)	[988, 1078]
Color-Congruent items	466 (7)	[452, 480]	553 (9)	[535, 571]	628 (11)	[605, 650]	730 (15)	[700, 760]	1012 (23)	[967, 1057]
Magnitudes of effects	Diff. (SE)	CI	Diff. (SE)	CI	Diff. (SE)	CI	Diff. (SE)	CI	Diff. (SE)	CI
Overall Stroop (Congruency) Effect (DR – Congruent)	11†/†† (2)	[8, 15]	31†/†† (3)	[26, 36]	45†/†† (4)	[38, 53]	66†/†† (5)	[56, 77]	94†/†† (8)	[79, 109]
Facilitation Effect (Congruent – Neutral)	-6†/† (2)	[-10, -3]	-16†/†† (2)	[-21, -12]	-21†/†† (3)	[-27, -16]	-27†/†† (5)	[-36, -18]	-21*/* (8)	[-37, -5]
Stroop Interference Effect (DR – Neutral)	5†/† (1)	[2, 8]	15†/†† (2)	[11, 19]	24†/†† (3)	[17, 30]	40†/†† (5)	[29, 51]	73†/†† (9)	[56, 91]
Response Conflict (DR – SR)	5**/** (2)	[2, 8]	19†/†† (2)	[15, 24]	30†/†† (3)	[24, 36]	43†/†† (4)	[34, 52]	69†/†† (8)	[54, 85]
Semantic Conflict (SR – Neutral)	0ns/ns (2)	[-3, 3]	-5ns/ns (2)	[-9, 0]	-6*/ns (3)	[-11, 0]	-3ns/ns (4)	[-11, 5]	4ns/ns (8)	[-12, 20]

Standard frequentist inference (presented before the slash used as a separator): [ns]non-significant;[*]significant at $p < .05$; [**]significant at $p < .01$; [†]significant at $p < .001$; Bayesian inference (presented after the slash): [ns]no evidence of an effect, with BF_{10} value between 0–1; [ns*]anecdotal evidence in favor of H1, with BF_{10} value between 1–3;[*]moderate evidence in favor of H1, with BF_{10} value between 3–10; [**]strong evidence in favor of H1, with BF_{10} value between 10–30; [†]very strong evidence in favor of H1, with BF_{10} between 30–100; [††]extreme evidence in favor of H1/of an effect, with BF_{10} < 100

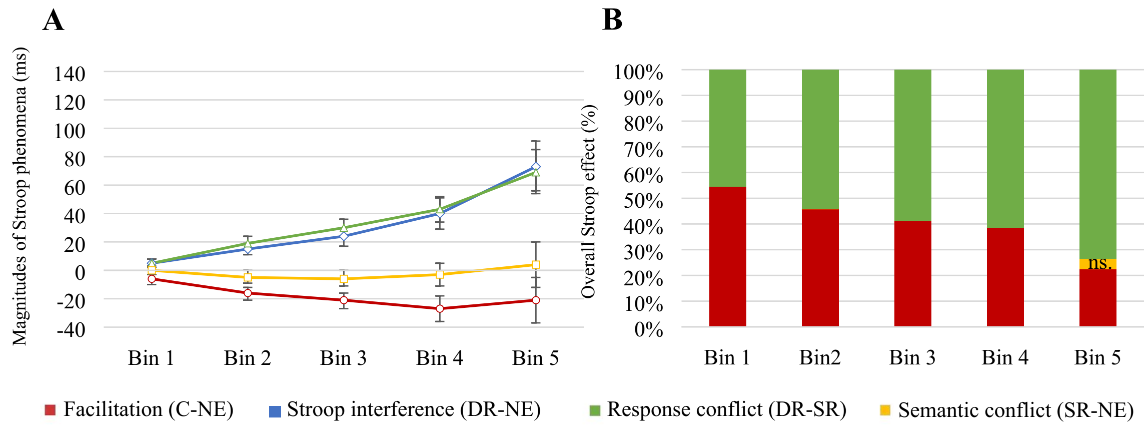

Fig. 4 Magnitudes of each Stroop effect (**panel A**) and composition of the overall Stroop (Congruency) effect in percentages (**panel B**) observed as a function of response speed in Experiment 2. Error bars in panel A represent 95% confidence intervals; Sum of Response and Semantic conflicts represented in panel B corresponds to Stroop interference

Procedure and stimuli

This experiment was completed online. Qualtrics (www.qualtrics.com) software was used to present the information sheet and Pavlovia (www.pavlovia.org) for data presentation and recording. The procedure was generally the same as in Experiment 1, except that Hasshim and Parris's (2014, Exp. 2) stimuli were used (i.e., English color-words: red, yellow, blue, and green; and non-color words: top, along, marvel, and past) and were presented in three blocks of 240 experimental trials (instead of one block as in Experiment 1).

Results and discussion

Applying the same cutoff as in Experiment 1 resulted in the exclusion of 2.2% of the total data, including 14 trials below 200 ms. Remaining RTs were then ordered and analyzed as in Experiment 1 (see OSM for additional results). This analysis revealed the main effect of Stimulus-Type [F(3,297) = 81.25, p < .001, η_p^2 = 0.451]; however, with no Bayesian evidence (BF_{10} = 0.933/BF_{01} = 1.07); the main effect of Response-speed [F(4,396) = 1012.36, p < .001, η_p^2 = 0.911, BF_{10} = +∞] and Stimulus-Type × Response-speed interaction [F(12,1188) = 33.16, p < .001, η_p^2 = 0.251, BF_{incl} > 1 for interaction alone; see Table S4 in OSM for models generated by JASP 0.14.1.0 ; JASP Team, 2017[6]]. The decompositions of this interaction (see Table 2, Fig. 4, and *Results* in OSM) showed that across the entire range of responses, the overall Stroop effect (M_{DR}–M_C) resulted from a significant contribution of both facilitation (M_C–M_N) and interference (M_{DR}–M_N, Roelofs, 2010). However, unlike in Experiment 1, and as in Hasshim and Parris (2014), interference was solely driven by response conflict (M_{DR}–M_{SR}), as the contribution of semantic conflict (M_{SR}–M_N) remained nonsignificant in all bins, and computing ten bins did not change this result.

Experiment 3

So far, Experiment 1 (using stimuli of Burca et al., 2022) revealed semantic conflict in slower responses, whereas semantic conflict was absent across the entire distribution in Experiment 2 (using stimuli of Hasshim & Parris, 2014, Exp. 2). While binned RTs were comparable across the two experiments (see Tables 1 and 2), it should be noted that Experiment 1 was conducted in person, while Experiment 2 was conducted online. Therefore, the present experiment was again conducted online while controlling for the type of stimuli used (French vs. English words as used by Burca et al., 2022, and Hasshim & Parris, 2014, respectively).

Method

One hundred and sixty participants were recruited using Prolific (www.prolific.co) and received £5.50 for their participation. Eighty of them were native French speakers, including 44 students (41 females, 38 males, and one preferred not to say, M_{age} = 24.63 years, SD = 4.26), and 80 were native English speakers, including 31 students (54 females and 26 males, M_{age} = 25.91, SD = 5.06). All participants reported having normal or corrected-to-normal color vision, being right-handed and not having been diagnosed with any language-related disorder. The study used a 4 (Stimulus-Type: different-response vs. same-response vs. neutral vs. congruent) × 2 (Stimuli: French vs. English) design for data collection, with the former factor being within-participant. To detect this within-between interaction, that was not a priori expected, a minimum sample size of 36 participants was recommended by G*Power (Faul et al.,

Table 3 Color-identification performance (mean response times, standard errors and 95% confidence intervals) observed as a function of stimulus- or effect-type and response speed in Experiment 3

	Bin 1		Bin 2		Bin 3		Bin 4		Bin 5	
Stimulus-type	M (SE)	CI	M (SE)	CI	M (SE)	CI	M (SE)	CI	M (SE)	CI
Different-Response (DR) Color-Incongruent Items	476 (6)	[463, 488]	582 (9)	[565, 599]	675 (11)	[654, 696]	804 (14)	[776, 832]	1098 (19)	[1059, 1136]
Same-Response (SR) Color-Incongruent Items	470 (6)	[459, 481]	564 (8)	[549, 580]	646 (10)	[626, 666]	754 (13)	[728, 780]	1013 (18)	[978, 1049]
Color-Neutral items	466 (6)	[455, 477]	557 (7)	[542, 572]	633 (9)	[616, 651]	739 (12)	[716, 762]	988 (16)	[956, 1020]
Color-Congruent items	456 (5)	[446, 466]	544 (7)	[530, 557]	619 (9)	[602, 636]	722 (12)	[699, 745]	980 (18)	[945, 1015]
Magnitudes of effects	Diff. (SE)	CI	Diff. (SE)	CI	Diff. (SE)	CI	Diff. (SE)	CI	Diff. (SE)	CI
Overall Stroop (Congruency) Effect (DR – Congruent)	$20^{\dagger/\dagger\dagger}$ (3)	[14, 25]	$38^{\dagger/\dagger\dagger}$ (3)	[31, 45]	$56^{\dagger/\dagger\dagger}$ (4)	[47, 64]	$82^{\dagger/\dagger\dagger}$ (7)	[68, 95]	$117^{\dagger/\dagger\dagger}$ (11)	[96, 138]
Facilitation Effect (Congruent – Neutral)	$-10^{\dagger/\dagger\dagger}$ (3)	[-16, -4]	$-13^{\dagger/\dagger\dagger}$ (3)	[-18, -8]	$-14^{\dagger/\dagger\dagger}$ (3)	[-21, -8]	$-17^{\dagger/**}$ (6)	[-28, -6]	$-8^{ns/ns}$ (9)	[-26, 11]
Stroop Interference Effect (DR – Neutral)	$10^{\dagger/\dagger}$ (3)	[4, 15]	$25^{\dagger/\dagger\dagger}$ (3)	[19, 32]	$41^{\dagger/\dagger\dagger}$ (4)	[34, 49]	$65^{\dagger/\dagger\dagger}$ (6)	[53, 77]	$109^{\dagger/\dagger\dagger}$ (10)	[90, 129]
Response Conflict (DR – SR)	$6^{*/ns*}$ (2)	[1, 11]	$18^{\dagger/\dagger\dagger}$ (3)	[12, 24]	$29^{\dagger/\dagger\dagger}$ (4)	[21, 36]	$50^{\dagger/\dagger\dagger}$ (6)	[39, 61]	$84^{\dagger/\dagger\dagger}$ (9)	[66, 102]
Semantic Conflict (SR – Neutral)	$4^{ns/ns}$ (3)	[-1, 9]	$7^{*/ns*}$ (3)	[1, 14]	$13^{**/**}$ (4)	[5, 20]	$15^{**/*}$ (6)	[4, 27]	$25^{**/*}$ (9)	[7, 44]

Standard frequentist inference (presented before the slash used as a separator): ^{ns}non-significant; *significant at $p < .05$; **significant at $p < .01$; †significant at $p < .001$; Bayesian inference (presented after the slash): ^{ns}no evidence of an effect, with BF_{10} value between 0–1; ^{ns*}anecdotal evidence in favor of H1, with BF_{10} value between 1–3; *moderate evidence in favor of H1, with BF_{10} value between 3–10; **strong evidence in favor of H1, with BF_{10} value between 10–30; †very strong evidence in favor of H1, with BF_{10} between 30–100; ††extreme evidence in favor of H1/of an effect, with BF_{10} < 100

2009) for an effect size of 0.25, power of 0.95, and type 1 error rate of 0.05. To detect a within-between interaction with an additional within participants five-level factor of Response-speed (that again was not a priori expected), the minimum sample size dropped to 14 participants. As in previous experiments, the sample size was substantially increased in order to conduct planned analysis of the entire RT distribution, namely to detect the a priori expected 4 (Stimulus-Type) × 5 (Response-speed: Bin 1-5) within-participants interaction and analyze it with a substantial number of observations for each level of Stimulus-Type in all five bins.

Procedure and stimuli

Qualtrics (www.qualtrics.com) software was used to present the information sheet and Pavlovia (www.pavlovia.org) for data presentation and recording. The procedure and stimuli were generally the same as in Experiments 1 and 2. The present experiment used one block of 240 randomly presented experimental trials, with 96 different-response (DR), 48 same-response (SR), 48 color-neutral (N), and 48 color-congruent (C) trials. For French-speaking participants, the same stimuli as in Experiment 1 were used. For English-speaking participants, the same stimuli as in Experiment 2 were used.

Results and discussion

Applying the same cutoff as the previous experiments resulted in the exclusion of 2% of the total data, including six trials below 200 ms. Remaining RTs were then analyzed in a 4 (Stimulus-Type: different-response vs. same-response vs. neutral vs. congruent) × 2 (Stimuli: French vs. English) × 5 (Response-speed: bins 1-5) ANOVA. As expected, the only interaction supported by the evidence from both standard ANOVA and Bayesian ANOVA was the Stimulus type × Response-speed interaction [$F(12,1896) = 34.29$, $p < .001$, $\eta_p^2 = 0.178$, $BF_{incl} > 1$ for interaction alone; see Table S6 in OSM for models generated by JASP 0.14.1.0 ; JASP Team, 2017[6] and Table S7 in OSM for descriptive results of a non-significant 4 (Stimulus-Type: different-response vs. same-response vs. neutral vs. congruent) × 2 (Stimuli: French vs. English) interaction]. This interaction – based on at least 1,536 observations per Stimulus-condition in each of the five bins – was further decomposed as in the previous experiments. As can be seen in Table 3 (see also Fig. 5), an overall Stroop effect (M_{DR}–M_C) emerged in Bin 1. This resulted from a significant contribution of both facilitation (M_N–M_C; up to Bin 4) and interference (M_{DR}–M_N, up to Bin 5). The contribution of response conflict (M_{DR}–M_{SR}) to Stroop interference became reliable in Bin 2 (and remained significant up to Bin 5) and that of semantic conflict (M_{SR}–M_N) in Bin 3 (and remained significant

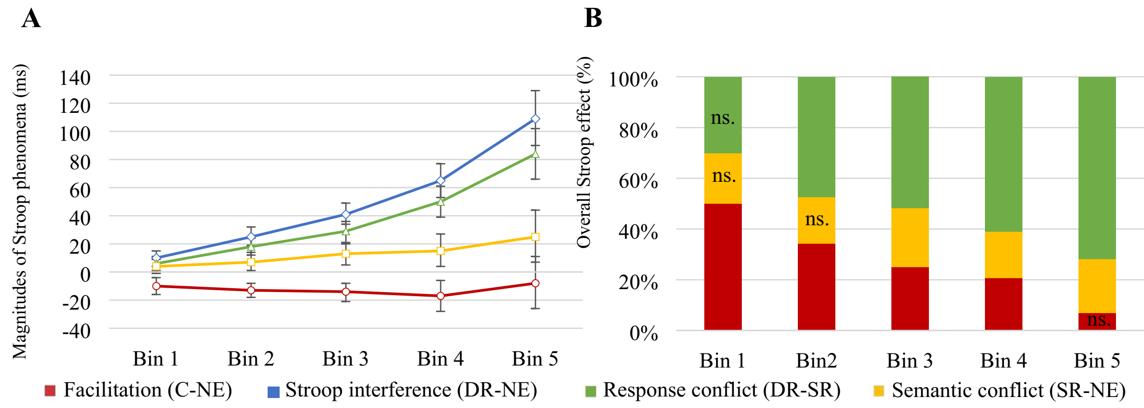

Fig. 5 Magnitudes of each Stroop effect (**panel A**) and composition of the overall Stroop (Congruency) effect in percentages (**panel B**) observed as a function of response speed in Experiment 3. Error bars in panel A represent 95% confidence intervals; Sum of Response and Semantic conflicts represented in panel B corresponds to Stroop interference

up to Bin 5). In sum, overall, the present experiment replicated the results of Experiment 1 (see Figs. 3 and 5).

Experiment 4

So far, two (Experiments 1 and 3) of the three experiments revealed semantic conflict on slow RTs only. However, none of these experiments manipulated response speed. To this end, the following experiment employed a response-stimulus interval (RSI) manipulation that has been shown to modify participants' response speed (Augustinova et al., 2018; De Jong et al., 1999; Jackson & Balota, 2013; Parris et al., 2012; Parris, 2014). Specifically, the 2,000 ms (henceforth long RSI) that elapsed between the individual's response on trial N and the presentation of a new stimulus on trial N+1 in all experiments reported above were shortened to 200 ms (henceforth short RSI) for half of the participants. This latter shortening was expected to increase response speed and subsequently result in the reduced magnitude of the overall Stroop effect compared to that observed with a long RSI (De Jong et al., 1999, for the initial demonstration). These a priori predicted differences in magnitudes were also expected to result from differences in semantic conflict, such that it was predicted to occur with the long but not short RSI.

Method

One hundred and seven psychology undergraduates (93 females and 14 males; M_{age} = 19.37; SD = 3.73) from Université Clermont Auvergne volunteered to take part in this experiment (interrupted by the COVID-19 pandemic). None of them had taken part in Experiment 1. They were all native French-speakers with normal or corrected-to-normal color vision. The study used a 4 (Stimulus-Type: different-response vs. same-response vs. neutral vs. congruent) × 2 (RSI: long vs. short) design for data collection, with the former factor being within-participant. To detect this within-between interaction, a minimum sample size of 36 participants was recommended by G*Power (Faul et al., 2009) for an effect size of 0.25, power of 0.95, and type 1 error rate of 0.05 (i.e., the interaction effect size found by similarly designed studies). In order to additionally conduct distributional analysis as in Experiments 1–3 (for the sake of comparison with these latter experiments), the sample size was again increased substantially such that 53 individuals were randomly assigned to the long-RSI condition and 54 individuals to the short-RSI condition. Beyond detecting the a priori expected 4 (Stimulus-Type) × 5 (Response-speed: Bin 1–5) within-participants interaction, this increase namely allowed us to analyze it with a substantial number of observations for each level of Stimulus-Type in all five bins.

Apparatus, stimuli, and procedure

The Apparatus, stimuli, and procedure were identical to those of Experiment 1 except that the participants used the same custom device as in Burca et al. (2021, 2022) to respond[7] instead of a keyboard, and that the present experiment used eight blocks of 120 experimental trials, resulting in a total of 960 trials per participant (i.e., more than in Experiments 1 and 2). In each block, 48 different-response (DR), 24 same-response (SR), 24 color-neutral (N), and 24 color-congruent (C) trials were randomly presented.

[7] One handle was used for responses to "blue" and "red" and the other for responses to "green" and "yellow." Color stickers were placed on each handle to remind participants of the response mapping. The use of handles by the right versus left hand was counterbalanced across participants.

Table 4 Color-identification performance (mean response times, standard error, 95% confidence intervals and percent errors) observed as a function of stimulus- or effect-type and RSI in Experiment 4

	Long-RSI (2,000 ms)			**Short-RSI** (200 ms)				
Stimulus-type	M (SE)	CI	%ER	M (SE)	CI	%ER	Short RSI effect (RT)	Short RSI effect (ER)
Different-Response (DR) Color-Incongruent Items	587 (12)	[562, 611]	4.3	548 (12)	[524, 572]	5.0	$-39^{*/*}$	$0.7^{ns/ns}$
Same-Response (SR) Color-Incongruent Items	570 (12)	[547, 594]	2.8	536 (12)	[513, 559]	4.2	$-34^{*/ns*}$	$1.3^{**/*}$
Color-Neutral items	561 (11)	[540, 582]	7.7	531 (10)	[511, 552]	8.1	$-30^{*/ns*}$	$0.4^{ns/ns}$
Color-Congruent items	548 (10)	[528, 568]	3.7	523 (10)	[503, 542]	4.8	$-26^{ns/ns*}$	$1.1^{ns/ns*}$
Magnitudes of effects	Diff. (SE)	CI		Diff. (SE)	CI		Short RSI effect (RT)	
Overall Stroop (Congruency) Effect (DR – Congruent)	$+38^{\dagger/\dagger\dagger}$ (4)	[31, 46]		$+25^{\dagger/\dagger\dagger}$ (4)	[18, 33]		$-13^{*/*}$	
Facilitation Effect (Congruent – Neutral)	$-13^{\dagger/\dagger\dagger}$ (3)	[-19, -7]		$-9^{**/\dagger}$ (3)	[-15, -3]		$-4^{ns/ns}$	
Stroop Interference Effect (DR – Neutral)	$+26^{\dagger/\dagger\dagger}$ (3)	[19, 32]		$+17^{\dagger/\dagger\dagger}$ (3)	[10, 23]		$-9^{*/ns*}$	
Response Conflict (DR – SR)	$+16^{\dagger/\dagger\dagger}$ (3)	[10, 22]		$+12^{\dagger/\dagger}$ (3)	[6, 18]		$-4^{ns/ns}$	
Semantic Conflict (SR – Neutral)	$+10^{**/*}$ (3)	[3, 16]		$+5^{ns/ns*}$ (3)	[-2, 11]		$-5^{ns/ns}$	

Standard frequentist inference (presented before the slash used as a separator): ^{ns}non-significant; *significant at $p < .05$; **significant at $p < .01$; †significant at $p < .001$; Bayesian inference (presented after the slash): ^{ns}no evidence of an effect, with BF_{10} value between 0–1; ^{ns*}anecdotal evidence in favor of H1, with BF_{10} value between 1–3; *moderate evidence in favor of H1, with BF_{10} value between 3–10; **strong evidence in favor of H1, with BF_{10} value between 10–30; †very strong evidence in favor of H1, with BF_{10} between 30–100; ††extreme evidence in favor of H1/of an effect, with $BF_{10} < 100$

Results and discussion

Applying the same cutoff as in previous experiments resulted in the exclusion of 2.1% of the total data, including 35 trials below 200 ms. Remaining mean RTs and errors (see Table 4) were then analyzed in a 4 (Stimulus-Type: different-response vs. same-response vs. neutral vs. congruent) × 2 (RSI: long vs. short) ANOVA, with the latter factor being between-participants. This analysis revealed

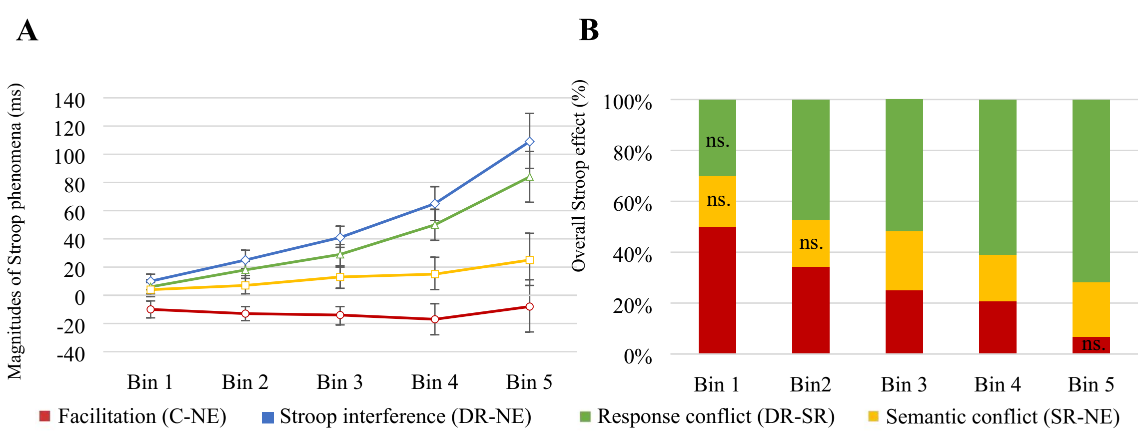

Fig. 6 Magnitudes of each Stroop effect (**panel A**) and composition of the overall Stroop (Congruency) effect in percentages (**panel B**) observed as a function of response speed in Experiment 4. Error bars in panel A represent 95% confidence intervals; Sum of Response and Semantic conflicts represented in panel B corresponds to Stroop interference

Table 5 Color-identification performance (mean response times, standard errors and 95% confidence intervals) observed as a function of stimulus- or effect-type and response speed in Experiment 4

	Bin 1		**Bin 2**		**Bin 3**		**Bin 4**		**Bin 5**	
Stimulus-type	M (SE)	CI	M (SE)	CI	M (SE)	CI	M (SE)	CI	M (SE)	CI
Different-Response (DR) Color-Incongruent Items	364 (5)	[355, 374]	443 (6)	[431, 456]	515 (8)	[499, 531]	618 (11)	[597, 640]	895 (15)	[865, 926]
Same-Response (SR) Color-Incongruent Items	363 (5)	[354, 372]	439 (6)	[428, 451]	506 (7)	[492, 521]	600 (10)	[580, 620]	857 (15)	[828, 887]
Color-Neutral items	365 (5)	[356, 374]	439 (6)	[428, 450]	502 (7)	[489, 515]	591 (9)	[573, 608]	834 (13)	[808, 859]
Color-Congruent items	360 (4)	[351, 369]	429 (5)	[419, 440]	490 (6)	[477, 503]	574 (9)	[557, 591]	824 (13)	[798, 849]
Magnitudes of effects	Diff. (SE)	CI	Diff. (SE)	CI	Diff. (SE)	CI	Diff. (SE)	CI	Diff. (SE)	CI
Overall Stroop (Congruency) Effect (DR – Congruent)	$4^{\dagger/\dagger\dagger}$ (1)	[2, 7]	$14^{\dagger/\dagger\dagger}$ (2)	[10, 18]	$25^{\dagger/\dagger\dagger}$ (3)	[20, 30]	$44^{\dagger/\dagger\dagger}$ (4)	[37, 52]	$72^{\dagger/\dagger\dagger}$ (6)	[61, 83]
Facilitation Effect (Congruent – Neutral)	$-5^{\dagger/\dagger\dagger}$ (1)	[-8, -3]	$-10^{\dagger/\dagger\dagger}$ (1)	[-13, -7]	$-12^{\dagger/\dagger\dagger}$ (2)	[-16, -8]	$-17^{\dagger/\dagger\dagger}$ (3)	[-22, -11]	$-10^{ns/ns*}$ (6)	[-21, 1]
Stroop Interference Effect (DR – Neutral)	$-1^{ns/ns}$ (1)	[-3, 1]	$4^{*/*}$ (2)	[1, 7]	$13^{\dagger/\dagger\dagger}$ (2)	[9, 17]	$28^{\dagger/\dagger\dagger}$ (3)	[21, 34]	$62^{\dagger/\dagger\dagger}$ (6)	[50, 73]
Response Conflict (DR – SR)	$1^{ns/ns}$ (1)	[-1, 3]	$4^{**/**}$ (1)	[1, 7]	$9^{\dagger/\dagger\dagger}$ (2)	[5, 12]	$18^{\dagger/\dagger\dagger}$ (3)	[13, 24]	$38^{\dagger/\dagger\dagger}$ (6)	[27, 49]
Semantic Conflict (SR – Neutral)	$-2^{ns/ns}$ (1)	[-4, 0]	$0^{ns/ns}$ (1)	[-3, 3]	$4^{*/ns*}$ (2)	[0, 8]	$9^{**/*}$ (3)	[3, 15]	$24^{\dagger\dagger/\dagger\dagger}$ (6)	[11, 37]

Standard frequentist inference (presented before the slash used as a separator): ^{ns}non-significant; *significant at $p < .05$; **significant at $p < .01$; †significant at $p < .001$; Bayesian inference (presented after the slash): ^{ns}no evidence of an effect, with BF_{10} value between 0–1; ^{ns*}anecdotal evidence in favor of H1, with BF_{10} value between 1–3; *moderate evidence in favor of H1, with BF_{10} value between 3–10; **strong evidence in favor of H1, with BF_{10} value between 10–30; †very strong evidence in favor of H1, with BF_{10} between 30–100; ††extreme evidence in favor of H1/of an effect, with $BF_{10} < 100$

a main effect of Stimulus-Type [$F(3,315) = 70.93$, $p < .001$, $\eta_p^2 = 0.403$, $BF_{10} = 1.27e+31$ (i.e., $BF_{10} > 100$)]. The effect of RSI [$F(1,105) = 4.27$, $p = .041$, $\eta_p^2 = 0.039$] was only supported by anecdotal Bayesian evidence ($BF_{10} = 1.32/BF_{01} = 0.759$), although numerically faster responding ($M = 534$, $SE = 11$) was observed in short RSI as compared to long RSI (M = 566, $SE = 11$). The Stimulus Type × RSI interaction was significant [$F(3,315) = 3.23$, $p = .026$, $\eta_p^2 = 0.030$] and supported by moderate evidence ($BF_{incl} = 3.82$ for interaction alone, see Table S7 in OSM for models generated by JASP 0.14.1.0 ; JASP Team, 2017[6]).

This interaction was further decomposed by testing the simple main effect of Stimulus-Type at each level of RSI. In line with this significant effect with long [$F(3,103) = 37.21$, $p < .001$, $\eta_p^2 = 0.520$] and short RSI [$F = (3,103) = 16.41$, $p < .001$, $\eta_p^2 = 0.323$], the overall Stroop effect (M_{DR}–M_C) resulted from a significant contribution of facilitation (M_N–M_C) and interference (M_{DR}–M_N) in both RSIs (see Table 4). However, as predicted, Stroop interference observed with short-RSI was driven only by response conflict (M_{DR}–M_{SR}; $p < .001$, $BF_{10} = 531.30$), whereas it resulted from a significant contribution of both response ($p < .001$, $BF_{10} = 27744.87$) and semantic (M_{SR}–M_N; $p = .004$, $BF_{10} = 5.33$) when the long RSI was used (see OSM for the analysis of errors). Therefore, unsurprisingly, when the entire distribution of RTs was additionally analyzed in the Stimulus type × RSI × Response-speed ANOVA, Stimulus type × Response-speed ANOVA – reported here for the sake of direct comparison with the results of Experiments 1–3 – was significant (while the omnibus ANOVA was not). Its further decomposition showed that the contribution of semantic conflict to the Stroop interference effect was again only reliable in slower RTs (from Bin 4 to 5[8]), whereas that of response conflict was reliable for a wider range of RTs (from Bin 2 to 5; see Table 5 and Fig. 6 for the detailed results and see OSM for the full description of these additional analyses).

Cross-experimental (post hoc) analyses

Given that the results of Experiment 2 are at odds with those observed in the remaining experiments, a Stimulus-Type × Response-speed ANOVA was additionally conducted on RTs from all four experiments (i.e., 482 participants) collected with a typical (i.e., 2,000 ms) RSI (see OSM for detailed analyses and results). The decompositions of the significant Stimulus-Type

[8] There were at least 2,035 observations per Stimulus-condition in each of the five bins.

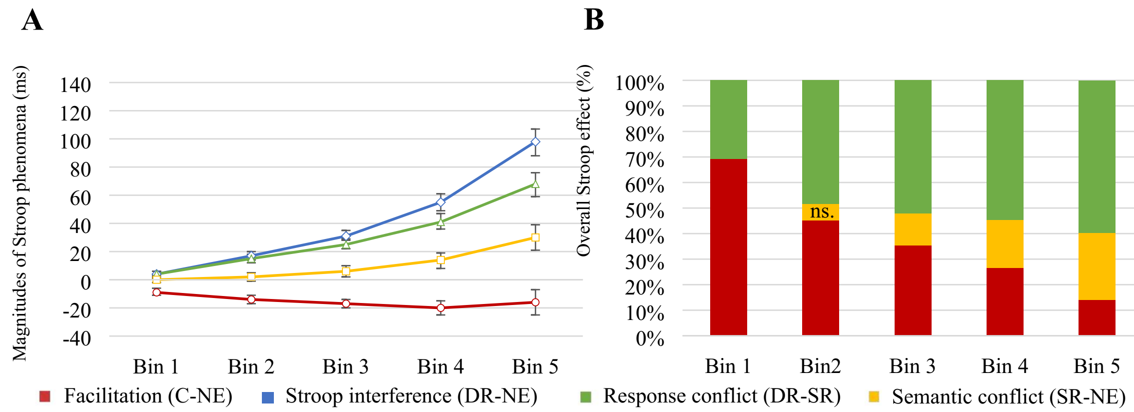

Fig. 7 Magnitudes of each Stroop effect (**panel A**) and composition of the overall Stroop (Congruency) effect in percentages (**panel B**) observed as a function of response speed in Experiments 1-4 (cross-experimental post hoc analysis). Error bars in panel A represent 95% confidence intervals; Sum of Response and Semantic conflicts represented in panel B corresponds to Stroop interference

× Response-speed interaction it showed (see Table 6) suggest that in Bin 1 and up to Bin 5, the overall Stroop effect (M_{DR}–M_C), was driven by both facilitation (M_N–M_C) and interference (M_{DR}–M_N), as in Roelofs (2010). Up to Bin 2, Stroop interference (M_{DR}–M_N) was itself entirely driven by response conflict (M_{DR}–M_{SR}). From Bin 3, semantic conflict (M_{SR}–M_N) started to contribute reliably (see Fig. 7). The additional Linear Mixed Modelling (see OSM for description and results) suggests that this Stimulus-Type × Response-speed interaction also best accounted for the hierarchical structure of the cross-experimental data (see Table S12 in OSM for all models). This idea is further reinforced by the non-significant intercept of the experiments (see Table S13 in OSM for estimate values) suggesting that the pattern of results in Experiment 2 does not deviate from the general pattern observed across the remaining experiments. At this point therefore, it is the most likely a false negative. In light of results observed in Experiment 3, the lack of semantic conflict in Experiment 2 is neither due to the fact that this latter experiment was run online, nor to the use of English words as stimuli (as compared to French ones). While more cross-language studies are needed to address this latter possibility in a more direct and systematic way than it was done in Experiment 3, it is important to note that French is characterized by greater print-to-sound transparency. As a result, one could expect semantic conflict generated by French stimuli to be of a smaller and not a greater magnitude as the one observed in English (see also Sulpizio et al., 2022, for a discussion of this issue).

General discussion

While semantic conflict was absent in Experiment 2 in both mean RTs (replicating Hasshim & Parris, 2014, see Table S3 in OSM) and across the entire distribution, the remaining three experiments and additional cross-experimental (post hoc) analyses revealed the presence of semantic conflict in slow-response trials. In a similar way to response conflict, the contribution of semantic conflict increased proportionally as a function of response speed. But while the contribution of response conflict was present across the entire distribution, that of semantic only became reliable in the tail of the RT distribution – as already suggested by the data of Hasshim and colleagues (2019). This presence of two conflicts is also in line with various event-related potential studies suggesting that there are two interference-related components (Appelbaum et al., 2009). The first – centro-medial and occurring from 350 to 500 ms after the presentation of the Stroop stimulus – is believed to arise from generators in the anterior cingulate cortex (ACC) and is thought to reflect the detection and/or resolution of response conflict. The second component – occurring from 500 to 900 ms post stimulus and maximal over the left parietal cortex – is "possibly related to the need for additional processing of word meaning" (Liotti et al., 2000, p. 701). Although the examination of response versus semantic conflict across the entire distribution (see Fig. 7, see also Table 6) strongly coincides with the aforementioned onsets, it is important to remember that the RT distribution cannot be interpreted as a timeline. Most models – including those mentioned above – assume that response conflict is determined by the activation of the underpinning semantic information (i.e., information flows from the semantic level to the response level). But the automatic processing of the word-dimension of incongruent trials is also thought to trigger semantic conflict, at least within models anticipating this conflict – occurring at the level of stimulus (in addition to Zhang and colleagues' two-conflict model, see also so-called *early selection models* that anticipate a unique (i.e., semantic) conflict, e.g., Seymour, 1977). Under this view,

Table 6 Color-identification performance (mean response times, standard errors and 95% confidence intervals) observed as a function of stimulus- or effect type and response speed in Experiments 1–4 (cross-experimental post hoc analysis)

	Bin 1		Bin 2		Bin 3		Bin 4		Bin 5	
Stimulus-type	M (SE)	CI	M (SE)	CI	M (SE)	CI	M (SE)	CI	M (SE)	CI
Different-Response (DR) Color-Incongruent Items	449 (4)	[441, 456]	550 (5)	[540, 560]	639 (6)	[627, 651]	762 (8)	[746, 778]	1046 (11)	[1024, 1068]
Same-Response (SR) Color-Incongruent Items	445 (4)	[438, 452]	535 (5)	[526, 544]	614 (6)	[602, 625]	721 (7)	[706, 735]	978 (10)	[958, 999]
Color-Neutral items	445 (3)	[438, 451]	533 (4)	[524, 542]	607 (5)	[597, 618]	707 (7)	[694, 721]	948 (10)	[929, 967]
Color-Congruent items	436 (3)	[429, 442]	519 (4)	[511, 527]	591 (5)	[580, 601]	687 (7)	[673, 701]	933 (10)	[913, 952]
Magnitudes of effects	Diff. (SE)	CI	Diff. (SE)	CI	Diff. (SE)	CI	Diff. (SE)	CI	Diff. (SE)	CI
Overall Stroop (Congruency) Effect (DR – Congruent)	$13^{\dagger / \dagger\dagger}$ (1)	[10, 15]	$31^{\dagger / \dagger\dagger}$ (2)	[28, 34]	$48^{\dagger / \dagger\dagger}$ (2)	[44, 53]	$75^{\dagger / \dagger\dagger}$ (3)	[69, 82]	$113^{\dagger / \dagger\dagger}$ (5)	[104, 123]
Facilitation Effect (Congruent – Neutral)	$-9^{\dagger / \dagger\dagger}$ (1)	[-11, -6]	$-14^{\dagger / \dagger\dagger}$ (1)	[-17, -11]	$-17^{\dagger / \dagger\dagger}$ (2)	[-20, -14]	$-20^{\dagger / \dagger\dagger}$ (3)	[-25, -15]	$-16^{\dagger / \dagger}$ (5)	[-25, -7]
Stroop Interference Effect (DR – Neutral)	$4^{\dagger / **}$ (1)	[2, 6]	$17^{\dagger / \dagger\dagger}$ (2)	[14, 20]	$31^{\dagger / \dagger\dagger}$ (2)	[28, 35]	$55^{\dagger / \dagger\dagger}$ (3)	[49, 61]	$98^{\dagger / \dagger\dagger}$ (5)	[88, 107]
Response Conflict (DR – SR)	$4^{** / *}$ (1)	[1, 6]	$15^{\dagger / \dagger\dagger}$ (2)	[12, 18]	$25^{\dagger / \dagger\dagger}$ (2)	[22, 29]	$41^{\dagger / \dagger\dagger}$ (3)	[36, 47]	$68^{\dagger / \dagger\dagger}$ (4)	[59, 76]
Semantic Conflict (SR – Neutral)	$0^{ns / ns}$ (1)	[-2, 3]	$2^{ns / ns}$ (2)	[-1, 5]	$6^{** / ***}$ (2)	[2, 10]	$14^{\dagger / \dagger\dagger}$ (3)	[8, 19]	$30^{\dagger / \dagger\dagger}$ (5)	[21, 39]

Standard frequentist inference (presented before the slash used as a separator): ^{ns}non-significant; *significant at $p < .05$; **significant at $p < .01$; †significant at $p < .001$; Bayesian inference (presented after the slash): ^{ns}no evidence of an effect, with BF_{10} value between 0–1; ^{ns*}anecdotal evidence in favor of H1, with BF_{10} value between 1–3; *moderate evidence in favor of H1, with BF_{10} value between 3–10; **strong evidence in favor of H1, with BF_{10} value between 10–30; †very strong evidence in favor of H1, with BF_{10} between 30–100; ††extreme evidence in favor of H1/of an effect, with $BF_{10} < 100$

semantic conflict reflects a slowdown that occurs whenever two distinct yet closely related semantic representations are simultaneously activated in an amodal semantic network (see, e.g., Seymour, 1977, for discussion, but see, e.g., Hock & Egeth, 1970, for the idea of perceptual rather than conceptual/semantic interference at the stimulus level). This means that while semantic and response conflict might be detected and/or resolved with different time courses, they are both likely to have an early functional processing locus.

Finally, the fact that semantic conflict is only apparent in slower responses could be interpreted as indicating that this type of conflict only emerges in trials involving larger lapses of selective attention (as initially reasoned by Scaltritti et al., 2022). This is precisely what the results of Experiment 4 suggest. In agreement with past studies, a larger overall Stroop effect was observed with long RSI compared to short RSI (e.g., Augustinova et al., 2018; De Jong et al., 1999; Jackson & Balota, 2013; Parris, 2014). Experiment 4 specifically revealed that this larger magnitude was due to the contribution of an additional (i.e., semantic) conflict.[9] These results are therefore consistent with De Jong et al.'s (1999) idea that faster responding (induced by RSI of 200 ms) significantly reduces lapses in the maintenance of the task goal and therefore optimizes participants' focus on the relevant color-dimension. Still, it should be noted that this latter causal chain (i.e., response speed influencing attention) is not warranted, as lapses of attention occurring at long RSI might as well lead to longer RTs.

Therefore, it seems important to emphasize at this point that the changes in goal maintenance anticipated by De Jong et al. (1999) might not be necessary, and that any factor that – exactly like RSI – changes response speed can consequently modify the nature of the Stroop effect. Indeed, factors such as response repetition from trial N-1 are also known to influence

9 The present findings are therefore incompatible with the idea that RSI affects response conflict but leaves semantic conflict unaffected – as previously argued by Augustinova et al. (2018). Our results are also incompatible with the idea that a short RSI reduces task conflict (i.e., a more general conflict that derives from the simultaneous activation of two task sets: word-reading vs. color-naming, e.g., Bench et al., 1993; Goldfarb & Henik, 2007; Hershman & Henik, 2020, for PET, behavioral, and pupillometric evidence; but see also Parris et al., 2023, for another perspective) as opposed to other types of conflict. In the present study, and unlike in Parris (2014), shortening the RSI failed to boost the magnitude of Stroop facilitation (while simultaneously reducing interference).

the speed of responses. To illustrate, Stroop studies – including the present one – do not usually control for associative-priming confounds (Henson et al., 2014; Hommel, 2004; Schmidt & Weissman, 2016), which can substantially impact response speed and therefore – as shown here – impact the type of conflict that is subsequently experienced on a given trial (response conflict alone or response conflict followed by semantic conflict). In sum, any factor that modifies response speed should determine how much semantic conflict is observed.

Despite the fact that the factors capable of influencing response speed in the Stroop task are not all known at present (as these are likely to range from the aforementioned sequence effects up to higher-order variables such as motivation) and other unresolved issues discussed above, this study has at least two important implications. First, it allows us to account for the existing contradictory findings that suggest that the contribution of semantic conflict in the Stroop task may be fragile. Indeed, in fast responses, semantic conflict failed to contribute significantly, as in mean RTs reported by Hasshim and Parris (2014; see also RTs from pupillometric studies of Hasshim & Parris, 2014, and of Hershman & Henik, 2020). In slower responses, on the other hand, the significant contribution of semantic conflict was indeed independent of both response conflict and of facilitation – as in mean RTs reported by Burca et al. (2022). Second, the present study adds to those showing that the overall Stroop effect increases proportionally as response speed gets slower (e.g., Jackson & Balota, 2010; Roelofs, 2010; Spieler et al., 1996). While this proportional increase is due, in part, to the fact that both facilitation and interference increase at similar rates (as showed by Roelofs, 2010), the present study allowed us to identify that larger magnitudes of Stroop interference in slower responses are specifically due to the contribution of an additional (i.e., semantic) conflict. Said differently, the distributional analyses conducted in the present study were unique in revealing the polymorphic nature of Stroop interference.

Conclusion

While it still needs to be extended to vocal response modality,[10] it is important to understand that neither unitary (e.g., Cohen et al., 1990; Logan, 1980; Roelofs, 2003) nor composite models (Zhang et al., 1999; Zhang & Kornblum, 1998) can account for this polymorphic nature of Stroop interference on their own. Therefore, these latter models might need to be amended to effectively incorporate semantic conflict (see, e.g., Kalanthroff et al. (2018) for amending Cohen et al.'s (1990) initial model to incorporate task conflict[9] that is also likely to occur in the Stroop task). Additionally, as neither class of models is very precise in predicting time-related aspects of the Stroop effects, these models also need to explain why exactly semantic conflict is restricted to longer responses. Finally, since all currently available models hold that any conflicts that are present are resolved at the level of response selection (i.e., before the actual motor action is initiated), future modelling efforts might also want to explain findings showing that Stroop effects can occur in part (Kello et al., 2000; Exp. 2) or even entirely (Bundt et al., 2018; Quétard et al., 2023) after the response has been initiated (i.e., during the response execution phase). In sum, while the present study constitutes a considerable step toward understanding the loci of the Stroop effect, this issue is still far from being solved. Further empirical and theoretical work is therefore still needed.

[10] Although it should be noted that phonological processing of the irrelevant word in the Stroop task clearly occurs with manual response modality (Parris et al., 2019a, 2019b), suggesting that the Stroop task administered with manual responses is not necessary qualitatively different from the one administered with vocal responses. If anything, semantic conflict – induced via color-associated items at least (e.g., SKY_{blue}) – tends to be smaller (and not bigger) with manual as opposed to vocal responses (see, e.g., Sharma & MacKenna, 1998; and Augustinova & Ferrand, 2014, for a broader discussion).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13421-024-01538-3.

Funding The work reported was supported by ANR-19-CE28-0013 Grant from Agence Nationale de la recherche and RIN Tremplin Grant 19E00851 from Normandie Region, France awarded to all senior authors.

Data availability All data and materials have been made publicly available on the Open Science Framework (OSF) and can be accessed at https://osf.io/8sazr/?view_only=bc6d4ddf2f0a4e629a1c14ea43806d9c (file "Data").

Code availability Programs used for running the experiments have been made publicly available on the Open Science Framework (OSF) and can be accessed at https://osf.io/8sazr/?view_only=bc6d4ddf2f0a4e629a1c14ea43806d9c (file "Stroop tasks").

Declarations

Conflicts of interest None.

Ethics approval The collection of the data reported in Experiment 1 was approved by the Clermont-Ferrand Sud-Est 6 Statutory Ethics Committee (IRB00008526), in Experiment 2 by the Bournemouth University Research Ethics (ID 35032), in Experiment 3 by the Research Ethics Committee IRB-UCA (IRB00011540-2020-26) and the Bournemouth Ethics Committee (ID 35032), and in Experiment 4 by Clermont-Ferrand Sud-Est 6 Statutory Ethics Committee (IRB00008526) and the Research Ethics Committee IRB-UCA of Université Clermont Auvergne approved this study (IRB00011540-2020-53).

Consent to participate After receiving information about the study, each participant agreed to participate by either signing a written consent (for in-person experiments) or by ticking the appropriate box/proceeding to the study (for online experiments).

Consent for publication Not applicable.

Open practices statements None of the experiments was preregistered but the data and materials for all experiments are available at https://osf.io/8sazr/?view_only=bc6d4ddf2f0a4e629a1c14ea43806d9c

References

Appelbaum, L. G., Meyerhoff, K. L., & Woldorff, M. G. (2009). Priming and backward influences in the human brain: processing interactions during the Stroop interference Effect. *Cerebral Cortex, 19*(11), 2508–2521. https://doi.org/10.1093/cercor/bhp036

Augustinova, M., & Ferrand, L. (2014). Automaticity of word reading: Evidence from the semantic Stroop paradigm. *Current Directions in Psychological Science, 23*(5), 343–348. https://doi.org/10.1177/0963721414540169

Augustinova, M., Silvert, L., Ferrand, L., Llorca, P. M., & Flaudias, V. (2015). Behavioral and electrophysiological investigation of semantic and response conflict in the Stroop task. *Psychonomic Bulletin & Review, 22*(2), 543–549. https://doi.org/10.3758/s13423-014-0697-z

Augustinova, M., Silvert, L., Spatola, N., & Ferrand, L. (2018). Further investigation of distinct components of Stroop interference and of their reduction by short response-stimulus intervals. *Acta Psychologica, 189*, 54–62. https://doi.org/10.1016/J.ACTPSY.2017.03.009

Augustinova, M., Parris, B. A., & Ferrand, L. (2019). The loci of Stroop interference and facilitation effects with manual and vocal responses. *Frontiers in Psychology, 10*, 1786. https://doi.org/10.3389/fpsyg.2019.01786

Balota, D. A., & Yap, M. J. (2011). Moving beyond the mean in studies of mental chronometry: The power of response time distributional analyses. *Current Directions in Psychological Science, 20*(3), 160–166. https://doi.org/10.1177/096372141140888

Bench, C. J., Frith, C. D., Grasby, P. M., Friston, K. J., Paulesu, E., Frackowiak, R. S. J., & Dolan, R. J. (1993). Investigations of the functional anatomy of attention using the stroop test. *Neuropsychologia, 31*(9), 907–922. https://doi.org/10.1016/0028-3932(93)90147-R

Brown, T. L. (2011). The relationship between stroop interference and facilitation effects: Statistical artifacts, baselines, and a reassessment. *Journal of Experimental Psychology: Human Perception and Performance, 37*(1), 85–99. https://doi.org/10.1037/a0019252

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of cognition, 1*(1), 9. https://doi.org/10.5334/joc.10

Bundt, C., Ruitenberg, M. F. L., Abrahamse, E. L., & Notebaert, W. (2018). Early and late indications of item-specific control in a Stroop mouse tracking study. *PLOS ONE, 13*(5), e0197278. https://doi.org/10.1371/journal.pone.0197278

Burca, M., Beaucousin, V., Chausse, P., Ferrand, L., Parris, B. A., & Augustinova, M. (2021). Is there semantic conflict in the Stroop task? *Experimental Psychology, 68*(5), 274–283. https://doi.org/10.1027/1618-3169/a000530

Burca, M., Chausse, P., Ferrand, L., Parris, B. A., & Augustinova, M. (2022). Some further clarifications on age-related differences in the Stroop task: New evidence from the two-to-one Stroop paradigm. *Psychonomic Bulletin & Review, 29*, 492–500. https://doi.org/10.3758/s13423-021-02011-x

Chen, A., Bailey, K., Tiernan, B. N., & West, R. (2011). Neural correlates of stimulus and response interference in a 2–1 mapping stroop task. *International Journal of Psychophysiology, 80*(2), 129–138. https://doi.org/10.1016/j.ijpsycho.2011.02.012

Chen, Z., Lei, X., Ding, C., Li, H., & Chen, A. (2013). The neural mechanisms of semantic and response conflicts: An fMRI study of practice-related effects in the Stroop task. *NeuroImage, 66*, 577–584. https://doi.org/10.1016/j.neuroimage.2012.10.028

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

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

De Houwer, J. (2003). On the role of stimulus-response and stimulus-stimulus compatibility in the Stroop effect. *Memory & Cognition, 31*(3), 353–359. https://doi.org/10.3758/BF03194393

De Jong, R., Berendsen, E., & Cools, R. (1999). Goal neglect and inhibitory limitations: dissociable causes of interference effects in conflict situations. *Acta Psychologica, 101*(2–3), 379–394. https://doi.org/10.1016/S0001-6918(99)00012-8

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Glaser, W. R., & Glaser, M. O. (1989). Context effects in Stroop-like word and picture processing. *Journal of Experimental Psychology: General, 118*(1), 13–42. https://doi.org/10.1037//0096-3445.118.1.13

Goldfarb, L., & Henik, A. (2007). Evidence for task conflict in the Stroop effect. *Journal of Experimental Psychology: Human Perception and Performance, 33*(5), 1170–1176. https://doi.org/10.1037/0096-1523.33.5.1170

Hasshim, N., Downes, M., Bate, S., & Parris, B. A. (2019). Response time distribution analysis of semantic and response interference in a manual response Stroop task. *Experimental Psychology, 66*(3), 231–238. https://doi.org/10.1027/1618-3169/a000445

Hasshim, N., & Parris, B. A. (2014). Two-to-one color-response mapping and the presence of semantic conflict in the Stroop task. *Frontiers in Psychology, 5*, 1157. https://doi.org/10.3389/fpsyg.2014.01157

Hasshim, N., & Parris, B. A. (2015). Assessing stimulus–stimulus (semantic) conflict in the Stroop task using saccadic two-to-one color response mapping and preresponse pupillary measures. *Attention, Perception, & Psychophysics, 77*(8), 2601–2610. https://doi.org/10.3758/s13414-015-0971-9

Henik, A., Bugg, J. M., & Goldfarb, L. (2018). Inspired by the past and looking to the future of the Stroop effect. *Acta Psychologica, 189*, 1–3. https://doi.org/10.1016/j.actpsy.2018.06.007

Henson, R. N., Eckstein, D., Waszak, F., Frings, C., & Horner, A. J. (2014). Stimulus–response bindings in priming. *Trends in Cognitive Sciences, 18*(7), 376–384. https://doi.org/10.1016/j.tics.2014.03.004

Hershman, R., & Henik, A. (2019). Dissociation between reaction time and pupil dilation in the Stroop task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(10), 1899–1909. https://doi.org/10.1037/xlm0000690

Hershman, R., & Henik, A. (2020). Pupillometric contributions to deciphering Stroop conflicts. *Memory & Cognition, 48*(2), 325–333. https://doi.org/10.3758/s13421-019-00971-z

Hock, H. S., & Egeth, H. (1970). Verbal interference with encoding in a perceptual classification task. *Journal of Experimental Psychology, 83*(2, Pt.1), 299–303. https://doi.org/10.1037/h0028512

Hommel, B. (2004). Coloring an action: Intending to produce color events eliminates the Stroop effect. *Psychological Research, 68*(2–3), 74–90. https://doi.org/10.1007/s00426-003-0146-5

Jackson, J. D., & Balota, D. A. (2013). Age-related changes in attentional selection: Quality of task set or degradation of task set across time? *Psychology and Aging, 28*(3), 744–753. https://doi.org/10.1037/A0033159

Jeffreys, H. (1961). *Theory of Probability* (3rd ed). Oxford University Press.

Jiang, J., Zhang, Q., & Van Gaal, S. (2015). EEG neural oscillatory dynamics reveal semantic and response conflict at difference levels of conflict awareness. *Scientific Reports, 5*(1), 12008. https://doi.org/10.1038/srep12008

Kalanthroff, E., Davelaar, E. J., Henik, A., Goldfarb, L., & Usher, M. (2018). Task conflict and proactive control: A computational theory of the Stroop task. *Psychological Review, 125*(1), 59–82. https://doi.org/10.1037/rev0000083

Kello, C. T., Plaut, D. C., & MacWhinney, B. (2000). The task dependence of staged versus cascaded processing: An empirical and computational study of Stroop interference in speech perception. *Journal of Experimental Psychology: General, 129*(3), 340–360. https://doi.org/10.1037/0096-3445.129.3.340

Labuschagne, E. M., & Besner, D. (2015). Automaticity revisited: When print doesn't activate semantics. *Frontiers in Psychology, 6*, Article 117. https://doi.org/10.3389/fpsyg.2015.00117

Lee, M. D., & Wagenmakers, E.-J. (2014). Bayesian cognitive modeling: A practical course. *Cambridge University Press*. https://doi.org/10.1017/CBO9781139087759

Liotti, M., Woldorff, M. G., Perez, R., & Mayberg, H. S. (2000). An ERP study of the temporal course of the Stroop color-word interference effect. *Neuropsychologia, 38*(5), 701–711. https://doi.org/10.1016/S0028-3932(99)00106-2

Logan, G. D. (1980). Attention and automaticity in Stroop and priming tasks: Theory and data. *Cognitive Psychology, 12*(4), 523–553. https://doi.org/10.1016/0010-0285(80)90019-5

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

Melara, R. D., & Algom, D. (2003). Driven by information: A tectonic theory of Stroop effects. *Psychological Review, 110*(3), 422–471. https://doi.org/10.1037/0033-295X.110.3.422

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers, 36*(3), 516–524. https://doi.org/10.3758/BF03195598

Parris, B. A. (2014). Task conflict in the Stroop task: When Stroop interference decreases as Stroop facilitation increases in a low task conflict context. *Frontiers in Psychology, 5*, 1182. https://doi.org/10.3389/fpsyg.2014.01182

Parris, B. A., Augustinova, M., & Ferrand, L. (2019). Editorial: The locus of the Stroop effect. *Frontiers in Psychology, 10*, 2860. https://doi.org/10.3389/fpsyg.2019.02860

Parris, B. A., Bate, S., Brown, S. D., & Hodgson, T. L. (2012). Facilitating goal-oriented behaviour in the Stroop task: When executive control is influenced by automatic processing. *PLOS ONE, 7*(10), e46994. https://doi.org/10.1371/JOURNAL.PONE.0046994

Parris, B. A., Sharma, D., Weekes, B. S., Momenian, M., Augustinova, M., & Ferrand, L. (2019). Phonological processing of the irrelevant word in the Stroop task with manual and vocal responses. *Experimental Psychology, 66*(5), 361–366. https://doi.org/10.1027/1618-3169/a000459

Parris, B. A., Hasshim, N., Wadsley, M., Augustinova, M., & Ferrand, L. (2022). The loci of Stroop effects: a critical review of methods and evidence for levels of processing contributing to color-word Stroop effects and the implications for the loci of attentional selection. *Psychological Research, 86*, 1029–1053. https://doi.org/10.1007/s00426-021-01554-x

Parris, B. A., Hasshim, N., Ferrand, L., & Augustinova, M. (2023). Do Task Sets Compete in the Stroop Task and Other Selective Attention Paradigms? *Journal of Cognition, 6*(1), 23. https://doi.org/10.5334/joc.272

Phaf, R. H., Van der Heijden, A. H. C., & Hudson, P. T. W. (1990). SLAM: A connectionist model for attention in visual selection tasks. *Cognitive Psychology, 22*(3), 273–341. https://doi.org/10.1016/0010-0285(90)90006-P

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of Neuroscience Methods, 162*(1–2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Quétard, B., Spatola, N., Parris, B. A., Ferrand, L., & Augustinova, M. (2023). A mouse-tracking study of the composite nature of the Stroop effect at the level of response execution. *PLoS ONE, 18*(1). https://doi.org/10.1371/journal.pone.0279036

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

Roelofs, A. (2010). Attention and facilitation: Converging information versus inadvertent reading in Stroop task performance. *Journal of Experimental Psychology: Learning Memory and Cognition, 36*(2), 411–422. https://doi.org/10.1037/A0018523

Šaban, I., & Schmidt, J. R. (2021). Stimulus and response conflict from a second language: Stroop interference in weakly-bilingual and recently-trained languages. *Acta Psychologica, 218*, 103360. https://doi.org/10.1016/j.actpsy.2021.103360

Scaltritti, M., Job, R., & Sulpizio, S. (2022). Different types of semantic interference, same lapses of attention: Evidence from Stroop tasks. *Memory & Cognition, 50*, 898–910. https://doi.org/10.3758/S13421-021-01256-0

Schmidt, J. R., & Cheesman, J. (2005). Dissociating stimulus-stimulus and response-response effects in the stroop task. *Canadian Journal of Experimental Psychology, 59*(2), 132–138. https://doi.org/10.1037/H0087468

Schmidt, J. R., Hartsuiker, R. J., & De Houwer, J. (2018). Interference in Dutch-French bilinguals. *Experimental Psychology, 65*(1), 13–22. https://doi.org/10.1027/1618-3169/a000384

Schmidt, J. R., & Weissman, D. H. (2016). Congruency sequence effects and previous response times: conflict adaptation or temporal learning? *Psychological Research, 80*(4), 590–607. https://doi.org/10.1007/s00426-015-0681-x

Seymour, P. H. K. (1977). Conceptual encoding and locus of the Stroop effect. *Quarterly Journal of Experimental Psychology, 29*(2), 245–265.

Sharma, D., & McKenna, F. P. (1998). Diferential components of the manual and vocal Stroop tasks. *Memory & Cognition, 26*(5), 1033–1040. https://doi.org/10.3758/BF03201181

Shichel, I., & Tzelgov, J. (2018). Modulation of conflicts in the Stroop effect. *Acta Psychologica, 189*, 93–102. https://doi.org/10.1016/j.actpsy.2017.10.007

Spieler, D. H., Balota, D. A., & Faust, M. E. (1996). Stroop performance in healthy younger and older adults and in individuals with dementia of the Alzheimer's type. *Journal of Experimental Psychology: Human Perception and Performance, 22*(2), 461–479. https://doi.org/10.1037/0096-1523.22.2.461

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*(6), 643–662. https://doi.org/10.1037/h0054651

Sulpizio, S., Job, R., Leoni, P., & Scaltritti, M. (2022). Prepotent task-irrelevant semantic information is dampened by domain-specific control mechanisms during visual word recognition. *Quarterly Journal of Experimental Psychology, 75*(3), 390–405. https://doi.org/10.1177/17470218211030863

Van Veen, V., & Carter, C. S. (2005). Separating semantic conflict and response conflict in the Stroop task: A functional MRI study. *NeuroImage, 27*(3), 497–504. https://doi.org/10.1016/J.NEUROIMAGE.2005.04.042

Zhang, H., & Kornblum, S. (1998). The effects of stimulus-response mapping and irrelevant stimulus-response and stimulus-stimulus overlap in four-choice Stroop tasks with single-carrier stimuli. *Journal of Experimental Psychology: Human Perception and Performance, 24*(1), 3–19. https://doi.org/10.1037/0096-1523.24.1.3

Zhang, H., Zhang, J., & Kornblum, S. (1999). A parallel distributed processing model of stimulus–stimulus and stimulus–response compatibility. *Cognitive Psychology, 38*(3), 386–432. https://doi.org/10.1006/COGP.1998.0703

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.