This manuscript has been accepted for publication in:
Moretti, L., Koch, I., Hornjak, R., & von Bastian, C. C. (in press). Quality over quantity: Focusing on high-conflict trials to improve the reliability and validity of attentional control measures. *Journal of Experimental Psychology: Learning, Memory & Cognition.* doi: 10.1037/xlm0001466

Quality Over Quantity: Focusing on High-Conflict Trials to Improve the Reliability and Validity of Attentional Control Measures

Luca Moretti[1], Iring Koch[1], Raphael Hornjak[1] & Claudia C. von Bastian[2]

[1]RWTH Aachen University, Institute for Psychology, Aachen, Germany

[2]University of Sheffield, Department of Psychology, Sheffield, UK

Author Note

Correspondence concerning this article should be addressed to Luca Moretti, Institute of Psychology, RWTH Aachen University, Jaegerstrasse 17/19, 52066 Aachen, Germany. Email: Luca.Moretti@psych.rwth-aachen.de.

Raw data and analyses scripts can be publicly accessed at: https://doi.org/10.23668/psycharchives.12584

Abstract

In conflict tasks, congruency effects are thought to reflect attentional control mechanisms needed to counteract response conflict elicited by incongruent stimuli. Although congruency effects are well-replicable experimentally, recent studies have evidenced low correlations between congruency effects measured across different paradigms, leading to a heated debate over whether these low correlations indicate a lack of construct validity or are rather attributable to high measurement error, as indicated by the poor reliability typically displayed by congruency effects. In the present study, we investigated whether the poor reliabilities of congruency effects are due to their poor theoretical specification. Specifically, we tested whether the psychometric properties of congruency effects can be improved by focusing exclusively on those trials in which response conflict is theoretically expected to be highest. We considered two factors modulating the degree of response conflict: previous trial congruency, with higher conflict following congruent trials, and the time elapsed since stimulus onset, with higher conflict in fast responses. Data from 195 participants completing a Simon and a spatial Stroop paradigm showed that generally poor split-half reliabilities for the full set of trials improved greatly when excluding post-incongruent and slow trials. Importantly, between-task correlations also increased substantially when controlling for these factors, suggesting that, with increased reliability, these tasks capture a common attentional control ability. Our results suggest that individual differences in conflict tasks can provide valid and reliable measures of inhibition as a major component of attentional control when focusing on the trials with the theoretically highest response conflict.

Keywords: Attentional control, interindividual differences, reliability, Stroop, Simon.

Quality Over Quantity: Focusing on High-Conflict Trials to Improve the Reliability and Validity of Attentional Control Measures

The term attentional control (also referred to as cognitive control or executive functions) is widely used to refer to a set of high-order abilities which allow goal-directed behaviour in the face of distraction (von Bastian et al., 2020). Attentional control has often been conceptualized as an umbrella term for multiple components (Miyake et al., 2000; for alternative taxonomies of attentional control, see: Friedman & Miyake, 2017; Stuss & Alexander, 2007), including *shifting*, the flexible switching between mental task sets, *updating*, the monitoring and removing of working memory representations, and *inhibition*, the overcoming of prepotent responses (sometimes further partitioned in response inhibition and resistance to proactive interference; Miyake & Friedman, 2004).

Over the years, an impressive number of studies in both experimental psychology and research on interindividual differences have focused on the last component using so-called conflict tasks (e.g., the Stroop and the Simon task) where an irrelevant feature of the presented stimuli, if attended, automatically triggers a task-irrelevant response, thus requiring inhibition. On the one hand, experimental psychologists have mostly been interested in providing mechanistic accounts of inhibition (e.g., Botvinick et al., 2001; Braver, 2012), positing that inhibition is triggered whenever response conflict is detected by a dedicated conflict-monitoring system (Botvinick et al., 2001; Kerns et al., 2004). On the other hand, interindividual differences researchers have focused on the relationship between this inhibitory ability to other cognitive domains such as working memory (Kane & Engle, 2003; Meier & Kane, 2013; Rey-Mermet et al., 2019; Unsworth et al., 2021; Unsworth et al., 2024), language proficiency (Borella et al., 2010) or fluid intelligence (Draheim et al. 2019, Kane & Engle, 2002; Rey-Mermet et al., 2019), in addition to a variety of every-day life aspects such as mental health (Barch & Ceaser, 2012; Kane et al., 2016; Nigg, 2000; Pievsky & McGrath, 2018), academic achievement (Blair 2002; Diamond, 2016) or substance abuse (Baler & Volkow, 2006). Although the construct of inhibition is a cornerstone of cognitive psychology

(Declerck & Koch, 2023; Friedman & Miyake, 2004; Nigg, 2000), recent studies have highlighted poor convergent validity between the paradigms putatively measuring it, calling into question the very existence of this construct (Rey-Mermet et al., 2019). In the present study, we propose a solution to this issue by exploiting the theoretical developments in experimental psychology to measure peoples' inhibitory abilities. To foreshadow, we will argue that to establish whether inhibition is a task-general ability, it is important to isolate those conditions that maximally require inhibition. Our results suggest that focusing on those conditions where response conflict is present, individual differences in attentional control can be validly and reliably captured in the Stroop and in the Simon task.

The congruency effect as a marker for inhibition

Studies investigating inhibition have employed experimental paradigms designed to trigger both task-irrelevant and task-relevant responses, such as the Stroop task (Stroop, 1935) and the Simon task (Simon, 1990). In the Stroop task, participants are asked to name the color of the letters of a color word. If the color and the word meaning are incongruent (e.g., the word BLUE presented in red), then color naming is slowed down relative to when color and word meaning are congruent (e.g., the word BLUE presented in blue). In this task, it is commonly assumed that responses are both activated by the automatic reading task and the goal-relevant color naming task (MacLeod, 1991). Similarly, in the Simon task, participants are asked to respond to a non-spatial stimulus attribute, such as the letters X or O, with a left vs. right key press while ignoring the spatial location of the stimulus, which is presented randomly on the left or on the right side of the screen. If the task-irrelevant spatial location of the stimulus is incongruent with the required response (e.g., a left response required by a stimulus presented on the right side), responding is slowed down relative to congruent stimuli. It is commonly assumed that, while the stimulus location is task-irrelevant, the spatially congruent response is nonetheless activated along with the required stimulus-categorization response (Hommel, 2011).

As these examples show, the presence of response conflict can be manipulated in so-called conflict paradigms by creating conditions in which the task-irrelevant response is either consistent (in congruent trials) or inconsistent (in incongruent trials) with the task-relevant response. The difference in performance between incongruent and congruent trials – the *congruency effect* – is thus commonly interpreted as a measure of response conflict and the inhibitory mechanisms employed to deal with it. According to current theories, once conflict between simultaneously activated responses is detected, attentional control is exerted to keep behavior goal-directed, possibly by adjusting attentional settings to the task-relevant feature (Botvinick et al., 2001; Kerns et al., 2004). Consequently, attentional control is thought to be necessary for dealing with response conflict in incongruent trials across conflict tasks.

Conflict tasks have a long tradition in experimental research on attentional control. Like other attentional control paradigms (e.g., task switching), these tasks provide two main advantages to the experimenter. First, as outlined above, they provide an intuitive way to isolate a construct of interest using condition-difference scores, which should yield high construct validity. Second, the experimental congruency effect that arises from this contrast is well-replicable. Due to their decades of success in experimental research, it is not surprising that researchers have also relied on these tasks to measure interindividual differences in attentional control abilities and assess their relation to other constructs such as working memory (e.g., Kane & Engle, 2003; Meier & Kane, 2013; Rey-Mermet et al., 2019; Unsworth et al., 2021) or fluid intelligence (e.g., Kane & Engle, 2002; Redick et al., 2016; Rey-Mermet et al., 2019).

However, recent evidence from interindividual differences research has cast doubts over whether conflict tasks truly capture a common attentional control ability. Several studies have found little shared variance between congruency effects in these paradigms (De Simoni & von Bastian, 2018; Draheim et al., 2021; Kalamala et al., 2020; Paap et al., 2020; Rey-Mermet et al., 2018; Rey-Mermet et al., 2019; Whitehead et al., 2020). Some researchers interpreted these findings as suggesting that conflict tasks do not tap similar attentional control abilities, but rather require task-

specific mechanisms (Rey-Mermet et al., 2018; for a related discussion in experimental research, see Egner, 2008; 2017), thus calling into question the construct validity of attentional control, and in particular inhibition, in these paradigms. In other words, conflict tasks may not really be measuring what we assume they are measuring. At the same time, others argued that low between-tasks correlations arise from methodological, rather than theoretical issues and, thus, should not be interpreted to inform cognitive theories (e.g., Draheim et al., 2021; Hedge et al., 2022).

In the present study, we propose that theoretically informed adjustments in measuring the congruency effect can address the methodological issues identified and improve the assessment of individual differences in attentional control substantively. Specifically, we tested whether the reliability of the congruency effect can be improved by focusing on those trials in which attentional control demands can be expected to be highest. We predicted that such more reliable measures should better capture the construct of interest (i.e., increase validity), and thus also lead to stronger correlations across tasks, which would provide convergent validity for those measures as representing measures of the same construct: inhibition. To foreshadow our results, we indeed found that our novel procedure substantially improved reliability and increased the shared variance between the two attentional control tasks investigated.

What Are the Causes of Weak Correlations Between Attentional Control Measures?

Before concluding that attentional control is either non-existent or at best task-specific (Rey-Mermet et al., 2018), a number of methodological issues have to be considered. In particular, the correlations between performance in conflict tasks may be attenuated by their low reliabilities, which are often only around $r = .50$ to $r = .70$ (Hedge, Powell & Sumner, 2018; Stahl et al., 2014; Paap & Sawi, 2016; Pettigrew & Martin, 2014; Schuch et al., 2022; Unsworth et al., 2021; von Bastian et al., 2016; for related findings with the affordance task, see: Littman et al., 2023). Conflict tasks may exhibit such low reliability for at least the following three reasons.

First, because experimental researchers aim at maximizing condition differences, experimental paradigms are designed to minimize between-person variance, which is typically

considered undesired noise. However, in correlational research, precisely that between-person variance is critical to reliably rank-order individuals within and across dimensions (Cooper et al., 2017; Hedge, Powell & Sumner, 2018). Therefore, experimental tasks designed to minimize between-person variance may be unsuitable for interindividual differences research.

Second, difference scores are problematic (Draheim et al., 2019; 2021) because their reliability is inherently limited: the higher the correlation between the two components of a difference score (e.g., RTs in congruent and incongruent trials), the lower their reliability. Recently, some attempts have been made to remediate the problems related to the use of difference scores. For example, Draheim et al. (2021; 2023) have devised new attentional control tasks and used accuracy-based modified versions of the existing conflict tasks. Although the new task versions showed a better test-retest reliability and higher between-task correlations than the traditional versions, these correlations were still relatively low ($r \approx .25$). Similarly, Rey-Mermet et al. (2019) introduced a deadline procedure and used error rate as the dependent variable. Reliabilities did improve but, again, between-task correlations were low. Finally, Burgoyne et al. (2023) have recently introduced modified versions of the most common conflict tasks. In these tasks, participants first have to focus on the relevant dimension of the stimulus, and then choose which of two response-stimuli's irrelevant dimensions matches the correct response. For example, in the modified Stroop task, if a red word is presented as stimulus, the participant should select the response-stimulus displaying the word RED. Reliability was found to be excellent. Critically, and different to the studies reviewed above, performance on these tasks loaded highly on a common latent factor, suggesting that they tapped a similar construct.

A third factor that may contribute to the low reliability of conflict tasks is trial-by-trial variability or noise (Rouder & Haaf, 2019; Rouder et al., 2019). Trial noise is particularly problematic in these paradigms given that the congruency effect is often relatively small. For example, mean Stroop effects are typically smaller than 200 ms with standard deviations that amount to at least about 35% to 50% of these means (e.g., Friedman et al., 2004; Miyake et al.,

2000). Rouder et al. (2019) estimated that observed correlations are typically reduced by half compared to true correlations, and that remediating this issue by increasing the number of trials would not be practically feasible. Instead, the problem of high trial noise has been tackled using hierarchical linear models that allow to explicitly model variability across trials (Rouder & Haaf, 2019). Nonetheless, when testing the correlation between congruency effects in Stroop and flanker tasks using hierarchical linear models, the authors found Bayesian evidence for the null hypothesis, leading them to conclude that attentional control may be task-specific.

Finally, another issue that does not relate to reliability but is directly associated with the lack of correlation between congruency effects, is the use of tasks emphasizing both speed and accuracy as equally important (Draheim et al., 2021). A recent meta-analysis showed that speed and accuracy are correlated only weakly within conflict tasks, suggesting that RTs and ERs may reflect different underlying processes (Hedge, Powell, Bompas et al., 2018). As a consequence, between-task correlations may be attenuated by interindividual differences in speed-accuracy trade-off (SAT) (Draheim et al., 2019; 2021; Hedge, Powell, Bompas et al., 2018). To address this problem, Hedge et al. (2022) used drift-diffusion models for conflict tasks (DMC, Ulrich et al., 2015), which allow to model both speed and accuracy simultaneously and distinguish conflict processing from other components of task performance, including response caution reflecting the extent to which people favor accuracy over speed. When fitting the DMC to seven data sets, Hedge et al. (2022) found no correlations between parameters reflecting conflict processing. Instead, the authors found significant correlations between the parameters reflecting general processing speed and response caution, suggesting that these conflict-unrelated processes alone could account for the (low) correlations observed in conflict tasks (cf. Rey-Mermet et al., 2021). Similarly, Löffler et al. (2024) used the standard drift-diffusion model to assess performance in conflict tasks. Their aim was to isolate inhibitory abilities, as measured with the drift rate in incongruent trials, from variability in processing speed, as measured with classical processing-speed tasks. When regressing the common inhibitory factor on the latent processing-speed factor, they found no residual variance attributable

to inhibitory abilities, thus suggesting that differences in conflict tasks reflect nothing more than differences in processing speed.

Taken together, these findings suggest possible problems related to both the reliability and the validity of the attentional control construct. Although none of the factors considered above is unique to conflict tasks per se (e.g., other difference measures, such as the task switch costs, still display good reliability, see von Bastian & Druey, 2017), their combination is problematic for obtaining reliable measures in these paradigms. Furthermore, these findings highlight that even when improving the tasks' design to better suit interindividual differences research, performance in conflict tasks still display little common variance (but see Burgoyne et al., 2023, for an exception).

Attentional Control and Response Conflict

Should we therefore indeed stop using the congruency effect to measure attentional control (Rey-Mermet et al., 2018)? We argue that this may be premature. Instead of posing the question dichotomously (i.e., is the congruency effect valid or not valid?), we propose that the validity and reliability of the congruency effect as a measure of individual differences in attentional control can be improved based on theoretical insights from experimental research. Specifically, computing response conflict using only a subset of the trials where attentional control demands are greatest will decrease measurement noise and increase reliability. Two factors known to modulate the strength of response conflict are the previous-trial congruency (herein N-1 congruency) and the time elapsed from stimulus onset.

First, the congruency effect has been found to be reduced, or even eliminated, following incongruent trials, a phenomenon known as the congruency sequence effect (CSE) or Gratton effect (Gratton et al., 1992; for a recent review, see Egner, 2017). The conflict monitoring theory of cognitive control (Botvinick et al., 2001) explains the CSE as an after effect of conflict occurrence, hypothesizing that the cognitive system adapts to conflict beyond the trial in which control was recruited (but see Hommel, 2004; Mayr et al., 2003, for alternative accounts of the CSE; see Frings et al., 2020, for a recent review). Although divergent theories exist on how such conflict adaptation

is achieved (e.g., whether it reflects active preparation, or passive carryover of control settings), for the purpose of the present paper it is important to note that conflict adaptation is conceptually distinct from monitoring and control recruitment per se. More specifically, performance in post-incongruent trials is driven by two components: The ability to recruit control in trial N-1, and the extent to which such control settings carry over to the next trial (Duthoo et al., 2014; Egner et al., 2010; but see: Schiltenwolf et al., 2023). In contrast, performance in post-congruent trials will largely depend on the on-line ability to direct attention away from the task-irrelevant feature and selectively suppress the prepotent task-irrelevant response. Therefore, congruency effects following congruent and incongruent trials likely reflect distinct cognitive mechanisms and, thus, measure different abilities rather than a coherent construct. As carryover of control settings is not intrinsically related to attentional control abilities, we argue that post-incongruent trials should be excluded to gain a more valid and reliable measure of control abilities.

Indeed, a recent study by Schuch et al. (2022) provided first evidence for the hypothesis that reliability of the congruency effect improves when considering N-1 congruency. The authors presented pictures of male or female faces that were superimposed by a female or male name. Participants had to categorize the gender of the name while ignoring the gender of the face. In line with the reasoning outlined above, split-half and test-retest reliabilities improved for both RTs and error rates when considering only N-1 congruent trials compared to the full set of trials. Even though these results provide a first anecdotal hint on how the congruency effect may more reliably measure attentional control on high-conflict trials only, a few limitations have to be noted. First, in their study, Schuch et al. focused exclusively on face-name interference. Therefore, it was not possible to determine whether a more reliable measure of the congruency effect would correlate with performance in other conflict tasks. Second, the sample size employed was small for interindividual differences research (N = 48), so that the true reliability may be rather different from the reported estimates (Schönbrodt & Perugini, 2013).

The second factor influencing response conflict is the time elapsed from stimulus onset. A vast literature has characterized the time course of the congruency effect as a function of response speed using delta plots (De Jong et al., 1994; Pratte et al., 2010; Ridderinkhof 2002; 2004; Hübner & Töbel, 2019; van den Wildenberg et al., 2010), which plot the effect on the y-axis and RT quantiles on the x-axis (Balota & Yap, 2011; Ridderinkhof, 2004). Most of this literature assumes that response activation from the task-relevant and irrelevant features follows two distinct routes (controlled and automatic, respectively) that compete for selection. The typical pattern of results is that the RT Simon congruency effect is strongest for fast responses but tends to decrease over time (Pratte et al., 2010; Proctor et al., 2011; Schwarz & Miller, 2012), whereas flanker and Stroop effects increase over time (Pratte et al., 2010). Intuitively, these results appear to indicate that the processes underlying the congruency effects in these tasks would be different. However, instead, these differences may be quantitative rather than qualitative, with stimulus-based interference decreasing as a function of time in all conflict tasks. First, although the RT pattern may differ between tasks, congruency effects in error rates have been consistently found to be strongest, if not only present, in fast responses (Gratton et al., 1992; Hübner & Töbel, 2019; Ridderinkhof, 2002; Stins et al., 2007). Second, under some circumstances, flanker and Stroop tasks can also produce negative-going delta slopes (Hübner & Töbel, 2019). Finally, the above-mentioned DMC (Ulrich et al., 2015) explicitly implements the idea that attentional control is required predominantly early in the trial irrespective of the specific task. In the DMC, evidence accumulation proceeds separately in a direct route, which is governed by the irrelevant task feature, and in a controlled route, which is under attentional control. Whereas the rate of evidence accumulation is constant for the controlled route, evidence accumulation in the direct route rapidly peaks and then returns to zero following a pulse-like function. Using DMC, Ulrich et al. (2015) were able to show that differences in the shapes of the delta plots can be reproduced by differences in the time-to-peak of the automatic process, thereby suggesting that differences between conflict tasks may be quantitative (i.e., how fast interference peaks and dissipates) rather than qualitative.

Taken together, these studies show that congruency effects may not effectively measure response conflict across the whole range of RT distributions. Instead, only responding quickly would require suppression of activation along the automatic route and/or amplification along the controlled route, whereas in slow responses automatic activation has likely decayed already. Notice that this assumption is specific to conflict tasks where automatic and controlled responses compete for selection, and should not be confused with common practices in sustained attention tasks where attention lapses are typically measured in the slow portion of RTs (e.g., the psychomotor vigilance task, Dinges & Powell, 1985). Following this rationale, slow responses originate primarily from an SAT shift where the threshold for response selection is increased. Importantly, we do not claim that response conflict is absent in slow responses, or that participants do not make use of attentional control at all in those trials. Instead, we argue that attentional control operations are already completed at the time when slow responses are given. Therefore, performance in these trials will be primarily driven by the chosen response threshold rather than reflect response conflict.

The Present Study

Although conflict tasks are highly similar on the surface level, performance in these paradigms correlates only weakly. At present, it is unclear whether this finding depends on the poor reliability of the congruency effect or whether these paradigms measure only task-specific attentional control. Should we indeed abandon conflict tasks due to their poor psychometric properties (Paap et al., 2020; Rey-Mermet et al., 2018)? A recently emerging line of research is urgently posing this question, testing different ways to improve the methodologies of currently existing paradigms (Burgoyne et al., 2023; Draheim et al., 2021; Rey-Mermet et al., 2019). The present study goes beyond the methodological aspects that may limit reliability in conflict tasks, and instead proposes an approach based on theoretical insights from experimental research. Specifically, we control for factors that may decrease the construct validity of the congruency effect as a measure of attentional control, which, in turn, will also increase reliability. We predict that reliabilities in a Simon and a spatial Stroop task will improve when focusing exclusively on N-1

congruent trials and on fast responses. This prediction is derived from findings indicating that response conflict should be stronger, if not exclusively present, in these trials. Critically, if we succeed in providing a more reliable measure of attentional control, we will also be able to test whether congruency effects in the Stroop and Simon tasks substantially correlate. Again, we predict larger correlations when focusing on fast post-congruent trials.

Method

The experiment was programmed in Tatool Web (www.tatool-web.com; von Bastian et al. 2013) and conducted online via Prolific (www.prolific.co). The study was conducted in accordance with the declaration of Helsinki. Participants were provided with information about the study, eligibility criteria for participation, their right to withdraw at any time without negative consequences, and data protection and confidentiality. All participants gave informed consent.

Participants

Participants were eligible if they were fluent English speakers (self-reported), living in the UK, and between 18-36 years of age. A total of 195 participants took part in the study (M_{age} = 23.2 ± 4.2; 69.5% female; 30.5% male). With this sample size, the estimated correlations were found to converge to the population value for true correlations of ρ = .60 or larger with a deviation of ± .1 for a 95% CI (Schönbrodt & Perugini, 2013). We found this criterion to fit well the aims of our study as reliabilities are usually considered to be satisfactory over r = .7, good over r = .8, and excellent above r = .9 (Drost, 2011; Nunnally, 1978). The study was conducted in accordance with the Declaration of Helsinki. All participants gave informed consent by clicking on the participation button. The experiment took overall about 30 min and participants were compensated £3 after study completion.

Materials

Simon Task

A circle or a square appeared on either the left or right portion of the screen, and participants had to indicate the shape identity while ignoring the location. To do so, they were asked to press the

X key for square, and M key for circle. Each trial began with a central fixation cross presented for 250 ms. After this time elapsed, the stimulus appeared either on the left or on the right of the fixation cross, and participants were required to provide a response. When the response was executed, the screen turned blank for 250 ms before the beginning of the next trial. No response deadline was implemented.

Following instructions, participants underwent 10 practice trials, in which performance feedback was provided. The test phase consisted of two blocks of 144 trials each (288 trials in total). Within each block, the geometrical shapes had an equal probability of being displayed on the left or on the right, thus creating a balanced number of congruent and incongruent trials. Furthermore, we balanced the number of N-1 congruent trials for each level of N congruency, resulting in 72 trials for each condition (+/- 1 trial in each block as the first trial cannot be classified with respect to N-1 congruency).

Spatial Stroop Task

We used the words “LEFT”, “CENTRE” and “RIGHT”, which were presented on a horizontal plane to the left, center, or right part of the screen. The participants’ task was to indicate the word identity while ignoring the position in which it was presented. Responses were given with the keys V for “LEFT”, B for “CENTRE” and N for “RIGHT”, which are also horizontally aligned on the keyboard. The timing procedure was identical as that used in the Simon task. Following a fixation cross presented for 250 ms the stimulus appeared and remained on screen until response. After response the screen turned blank for 250 ms and a new trial began.

After a brief practice phase of 10 trials, in which performance feedback was provided, three blocks followed consisting of 180 trials each (540 trials in total). Within each block, each word was presented equally often in each position. This resulted in 120 incongruent and 60 congruent trials in each block. Furthermore, we balanced the number of N-1 congruent trials for each level of N congruency, resulting in 20 cC (congruent previous trial followed by congruent current trial), 40 cI (congruent-incongruent), 40 iC (incongruent-congruent) and 80 iI (incongruent-incongruent) trials

in each block (again, +/- 1 trial in each block as the first trial cannot be classified with respect to N-1 congruency). Finally, response repetitions were as likely to occur as response switches. This feature and the 3-choice variant of the spatial Stroop were chosen to allow an assessment of episodic memory contributions to the observed effects (Braem et al., 2019; Hommel et al., 2004), as better detailed in S3 of the Supplementary Materials (see also the statistical analyses section).

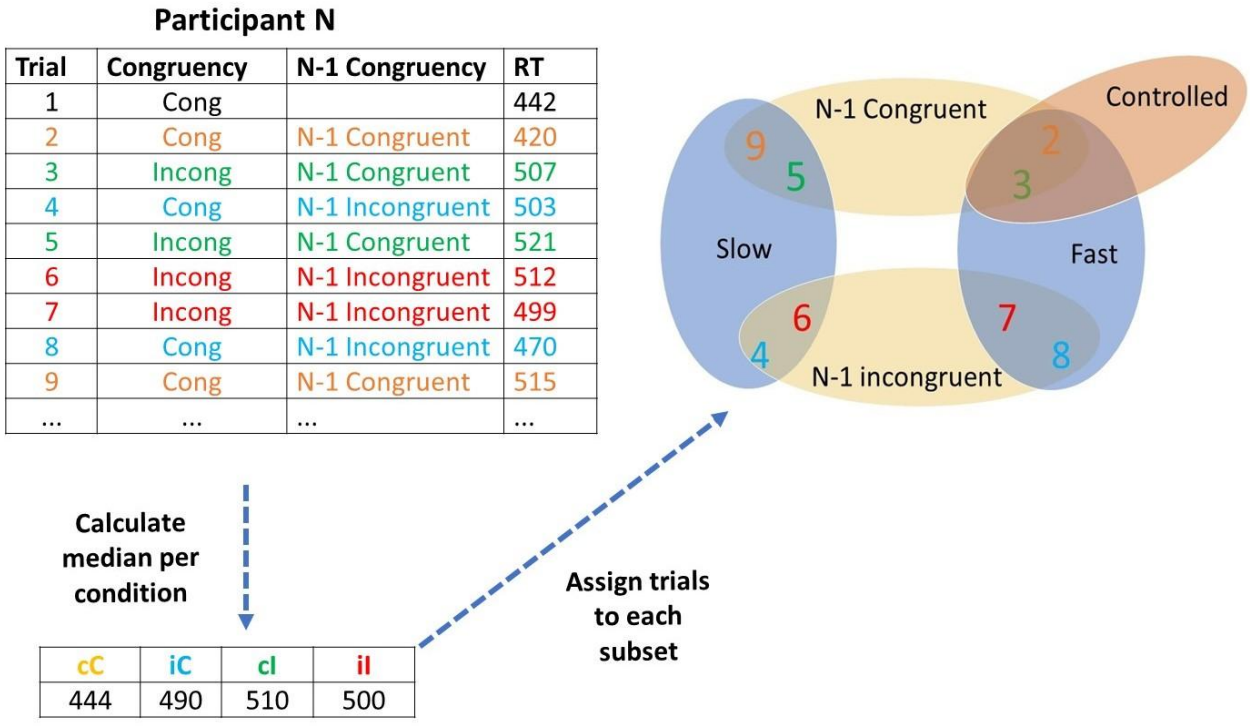
Statistical Analyses

All analyses were conducted in R version 4.1.1 (https://www.R-project.org/). Raw data and data analyses scripts can be found at: https://doi.org/10.23668/psycharchives.12584. This study was not pre-registered. First, experimental differences between conditions were assessed using analyses of variance (ANOVAs). For each task, individual 2 x 2 x 2 within-subjects ANOVAs were run on RTs and squared-root arcsine-transformed[1] error rates (ERs), with Congruency (congruent, incongruent), N-1 Congruency (N-1 congruent, N-1 incongruent) and Speed (fast, slow) as independent variables. Although conflict likely varies continuously as a function of RTs, we still opted for defining Speed as a categorical variable because, for the correlational analyses, we needed to define a threshold over which responses would be considered slow (and thus be removed). Furthermore, although splitting the RT distribution in many time-bins would allow to assess the relation between speed and conflict in a fine-grained fashion, we decided instead to maximize the number of trials in each condition. For these reasons, the Speed factor was defined by splitting the

[1] In binomial data, the variance of a distribution is a function of its mean, with means closer to the extremes (0 or 1) producing the smallest variance. Thus, if different error rates are observed across two conditions, the variance of the distributions will be unequal unless error rates in both conditions are equally distant from 0.5. Therefore, as in previous studies (Moretti et al., 2021; 2023a), we used squared-root arcsine transformation to remediate this issue by stabilizing the variance (Winer et al., 1971). Generally, squared-root arcsine transformation yields more conservative results than when using raw proportions.

RT distribution of each participant in each condition into halves defined by the median in each condition. Responses above the median were considered slow, whereas the ones below the median were considered fast. A visual depiction of how the trials were categorized is available in Figure 1. The main purpose of the experimental analyses was to ensure we replicated the expected effects, that is, the congruency effect and its modulations by N-1 Congruency and Speed. Significant interactions were further explored by performing appropriate follow-up tests (i.e., ANOVAs or t-tests). Following the suggestions of Lakens (2013), we report η_p^2 and η_G^2 for ANOVAs. For paired t-tests, we report Cohen's d_z computed by dividing the mean difference scores by their standard deviation (Brysbaert, 2019; Lakens, 2013).

Figure 1. *Graphical illustration of trials subsetting.*



Note: The figure illustrates how trials were divided into different subsets. First, the congruency in trial N-1 was determined (N-1 Congruency). Then, for each combination of Congruency and N-1 Congruency the median RT was calculated separately for each participant. If the response time for a given trial fell below the median for its condition, the trial was labelled as fast. If not, as slow. For instance, the figure shows that trial 2 was considered fast as response time in that trial (420ms) was below the median of the N-1 congruent-N congruent (cC) condition (444 ms).

Following the experimental analyses, split-half reliabilities of the congruency effect were assessed using a bootstrapping approach recently proposed by Parsons et al. (2019; see Pronk et al., 2022, for a review of methodologies). A visual description of this procedure can be found in Figure 2. First, two separate samples are created by randomly splitting the dataset into halves, with the constraint that the number of congruent and incongruent trials must be equal for each participant in each sample. Next, the congruency effect is calculated separately for each half, and Pearson's *r* correlations are computed between the samples and corrected using the Spearman-Brown prophecy (SBP) formula (see below). The same process was repeated for 20,000 times, returning a distribution of corrected Pearson's *r*s. The mean of the distribution was taken as a point estimate of the reliability coefficient. The lowest and highest 2.5% quantiles reflected the limits of 95% confidence intervals.

For each of the four dependent variables (Simon RT, Simon ER, Stroop RT, Stroop ER), reliabilities of the congruency effect were calculated for all trials and for three subsets of trials: 1) using only N-1 congruent trials, 2) using only fast responses, and 3) using only fast N-1 congruent responses (herein, *controlled subset*). To correct for the smaller number of trials in the subsets, the SBP formula was applied to allow for comparing their reliabilities to that of the full set of trials:

$$r_c = \frac{N * r_{xx}}{1 + (N-1) * r_{xx}}, \quad (1)$$

where r_{xx} is the observed reliability coefficient and r_c is the predicted value of the reliability coefficient if *N* times more trials were used. The value of *N* was set to 2 for the analysis of the full set of trials to account for the data loss intrinsic in calculating split-half reliabilities. *N* was set to 4 for the analyses only including fast responses in both tasks, and for the N-1 congruent subsets in the Simon task. In the Stroop task, 66% of the trials were incongruent, leaving only a third of trials for analyses restricted to N-1 congruent trials. Therefore, *N* was set to 6 for the analyses of this subset of trials. Finally, in the controlled sample analyses, *N* was set to 8 for the Simon task and to 12 for the Stroop task.

Although SBP correction is commonly used when assessing reliability (Draheim et al., 2021; Pettigrew & Martin, 2014; Redick et al., 2016; Rey-Mermet et al., 2018; Stahl et al., 2014; Unsworth et al., 2021) due to the random half-split of the data, it is worth noting that, when excluding a large number of trials (and, thus, using large *N*s), the effects of SBP correction can be quite large. Therefore, we report the corrected and uncorrected correlation coefficients for all analyses. Critically, however, SBP corrections do not artificially increase reliability, independent of *N*. This can be demonstrated by randomly removing $1/N^{th}$ part of the full set of trials and applying the corresponding SBP correction. For any *N* used in our procedure (i.e., 2, 4, 6, 8, 12), reliabilities were comparable to the uncorrected reliability in the whole sample, thus demonstrating that SBP corrections reflect the reliability in the full set of trials.

In addition to providing a robust measure of reliability, the bootstrapped approach adopted here offers two ways to compare reliabilities across subsets. First, it generally allows for assessing whether significant differences exist between reliabilities using the bootstrapped 95% confidence interval (CI) of the $r_{adjusted}$ distributions (e.g., Cooper et al., 2017). However, in the present context this approach would be problematic due to the dependencies between the sets of trials (Meng et al., 1992). To address this issue, at each iteration, we compared the reliabilities of the congruency effect in each subset to the reliability of the congruency effect in the full set of trials. To this aim we used Zou's (2007) approach for testing differences in correlation strength. Importantly, to account for dependencies in the data, we calculated the 95% CI for the comparison of dependent non-overlapping correlations (Meng et al., 1992). A bootstrapped *p*-value was then computed as the proportion of 95% CIs not including 0, (e.g., $p = .050$ would mean that in 95% of the iterations the 95% CI indicated a significant difference between reliability in the whole trial-set and the subset under examination).

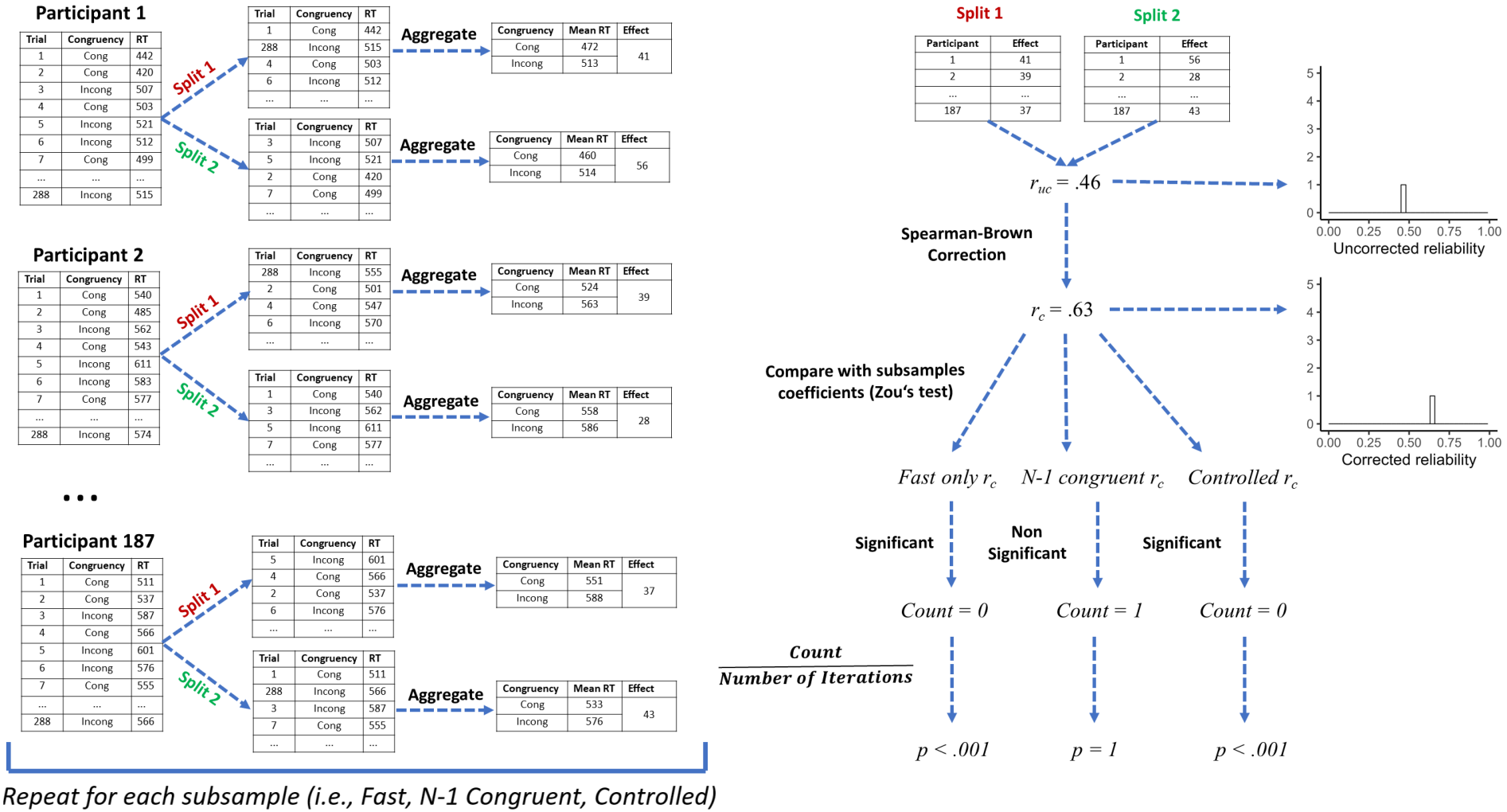
Finally, we computed correlations of the congruency effects between the two tasks for RTs and ERs separately. Importantly, this was done both for the full set of trials and for each subset. Again, to enable comparisons between these correlation coefficients, SBP corrections were used for

adjusting the Pearson's *r* depending on how many trials were excluded. Once appropriate SBP corrections were applied to the correlation matrix, we tested whether significant differences existed using the test for non-overlapping dependent correlations proposed by Zou (2007).

Additional Analyses Excluding Partial Repetitions

An important issue when dealing with CSEs is that they can as well be explained by low-level phenomena, in addition to attentional control upregulation (Braem et al., 2019; Frings et al., 2020). In particular, the episodic binding perspective proposes that, on each trial, the irrelevant stimulus feature is bound together with the produced response (Hommel, 2004). When only one of these features repeats, it automatically retrieves the other feature, thus generating interference. In two-choice conflict tasks it is not possible to assess the contribution of such partial repetitions to the CSE as they are always present when the congruency levels switch (i.e., in cI and iC trials). For this reason, we chose to use a three-choice version of the spatial Stroop task to be able to replicate our analyses excluding partial repetitions. We report these additional analyses in the S3 of the Supplementary Material. Critically, our results were essentially replicated when excluding partial repetitions.

Figure 2. *Graphical illustration of the bootstrapping procedure for calculating reliabilities.*



Note: The figure illustrates one cycle of our bootstrapping procedure in the Simon task. The left part of the graph illustrates that for each participant the unaggregated data are randomly split in half, with the only constraint that an equal number of congruent and incongruent trials would be present in each half. For each half, the mean congruency effect is calculated, and the correlation between halves is calculated.

The same procedure is then repeated for each of the subsamples (i.e., fast only, N-1 congruent and controlled). The uncorrected correlation coefficient is stored as one data point of the bootstrapped distribution of uncorrected coefficients. Then the Spearman-Brown correction is applied, and the corrected correlation coefficient is stored as one data point of the corrected correlation coefficients. Finally, the corrected correlation coefficient of the whole sample is compared with those of the subsamples using Zou's approach (2007). If the confidence interval of the test does not contain 0 (i.e., the difference is significant), we count the test result as a 0. If it does, we mark the count the result as a 1. The p-value represents the proportion of non-significant comparisons across iteration cycles.

Results

Data Trimming

Data trimming was performed separately for the Simon and the Stroop task. Responses above 2,000 ms, post-error trials, fast guesses and the responses following fast guesses were excluded. Fast guesses were defined as RT < 250 ms in the Simon task (1.0%) and RT < 300 ms in the Stroop (1.5%). Next, RT outliers were removed using a 2.5 SD deviation criterion for each participant and each congruency condition. Errors were removed from the RT analyses, amounting to 6.3% in the Simon and 7.4% in the Stroop task. After data trimming, for the included participants, 90.1% of trials were retained for the Simon task analyses. More specifically, the percentage of trials included in the analyses was: 92.9% for cC trials, 89.9% for cI trials, 92.9% for iC trials and 87.5% for iI trials. In the Stroop task 89.8% of the trials were retained. More specifically, the percentage of trials included in the analyses was: 93.5% for cC trials, 89.1% for cI trials, 93.3% for iC trials and 87.4% for iI trials. For each task, participants were excluded if they committed more than 30% errors and/or 10% fast guesses, and/or 10% slow responses, resulting in the exclusion of 8 participants in the Simon task and 12 participants in the Stroop task. For computing correlations between performance in the two tasks, all participants that were excluded in either task were removed (15 in total).

Experimental Analyses

Figures 3 and 4 illustrate the descriptive statistics of the Simon and Stroop tasks, respectively. For the sake of conciseness, we will report only higher-order interactions here. The interested reader can find further details about the lower-order effects, together with the complete ANOVA results in S2 of the Supplementary Materials.

Simon Task

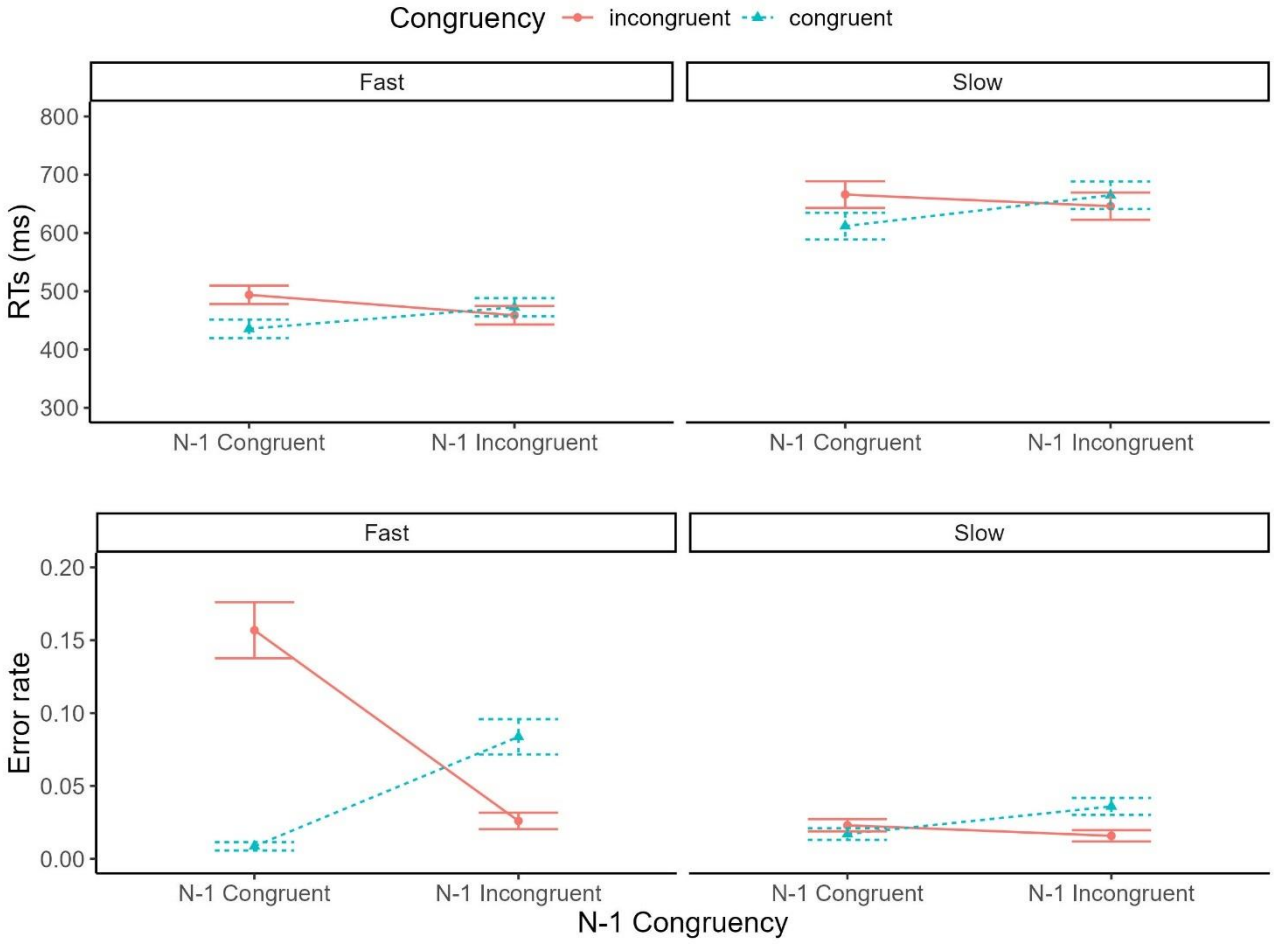
The three-way interaction between Congruency, N-1 Congruency and Speed did not reach significance for RTs, $F(1, 186) < 1$, $p = .821$, $\eta_p^2 < .01$, $\eta_G^2 < .01$. We did observe a strong

interaction effect between Congruency and N-1 Congruency, $F(1, 186) = 548.90$, $p < .001$, $\eta_p^2 = .75$, $\eta_G^2 = .02$, indicating the presence of a CSE. We found a weaker but still significant interaction effect between Congruency and Speed, $F(1, 186) = 3.92$, $p = .049$, $\eta_p^2 = .02$, $\eta_G^2 < .001$, indicating larger congruency effects in fast responses (M = 22 ms, SD = 17 ms) compared to slow responses (M = 18 ms, SD = 37 ms). Finally, we observed an interaction effect between Speed and N-1 Congruency $F(1, 186) = 76.86$, $p < .001$, $\eta_p^2 = .29$, $\eta_G^2 < .01$, indicating that in fast responses there was no difference in performance between post-congruent and post-incongruent trials (M = 0 ms, SD = 12 ms), whereas in slow responses RTs were slower in post-incongruent trials than in post-congruent trials (M = 17 ms, SD = 26 ms).

When analyzing ERs, we found a significant three-way interaction effect involving all factors, $F(1, 186) = 253.79$, $p < .001$, $\eta_p^2 = .58$, $\eta_G^2 = .11$. Decomposing the interaction for fast and slow responses showed a strong CSE in fast responses, $F(1, 186) = 476.50$, $p < .001$, $\eta_p^2 = .72$, $\eta_G^2 = .40$, and a somewhat smaller CSE in slow responses, $F(1, 186) = 52.49$, $p < .001$, $\eta_p^2 = .22$, $\eta_G^2 = .05$.

Figure 3

Experimental Effects in the Simon Task



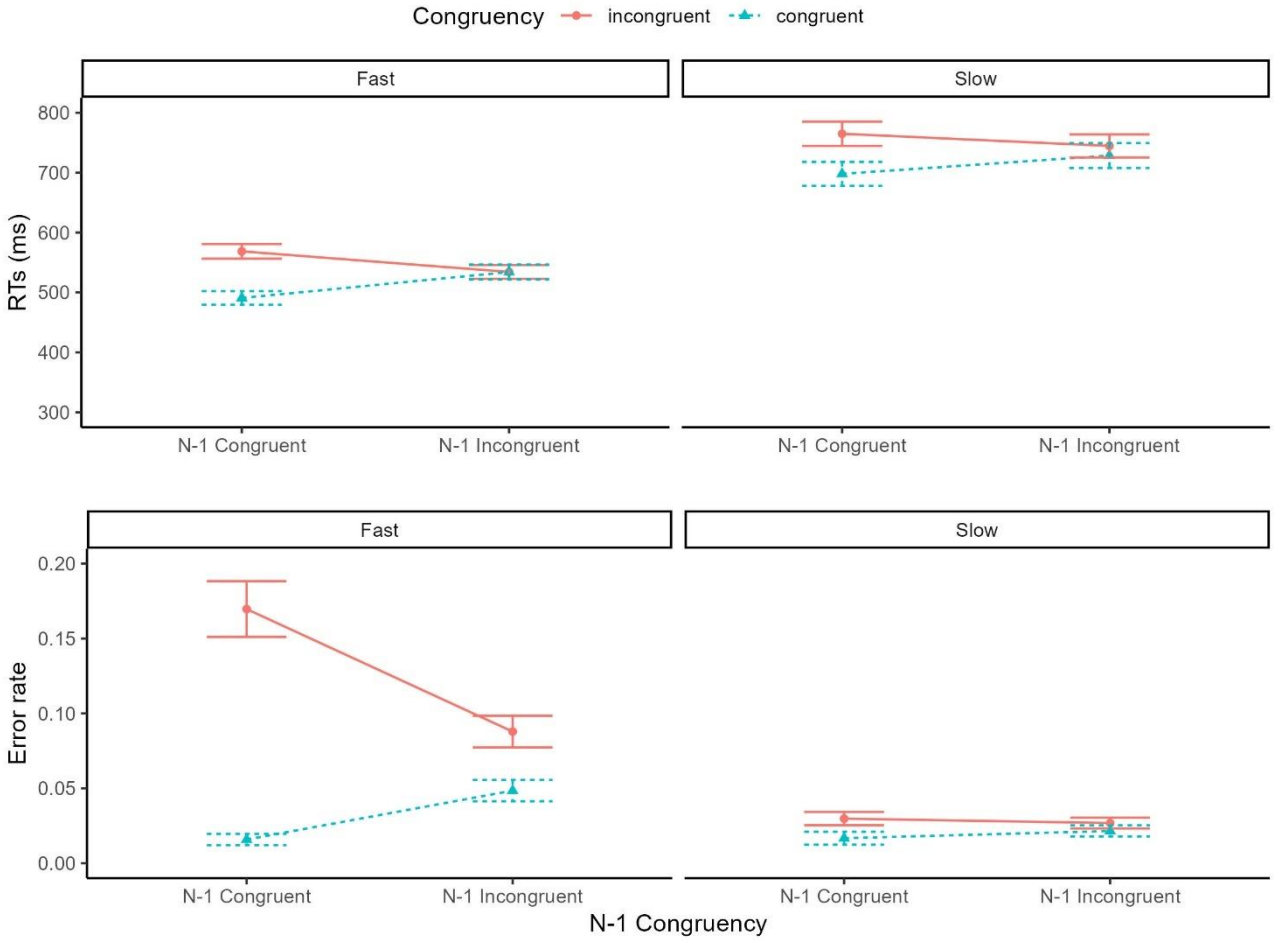
Note. Reaction times (upper panel) and error rates (lower panel) are plotted as a function of Congruency, N-1 Congruency and Speed. Error bars represent 95% confidence intervals.

Stroop Task

In RTs the three-way interaction involving all factors was significant, $F(1, 182) = 63.27$, $p < .001$, $\eta_p^2 = .26$, $\eta_G^2 < .01$. The CSE was larger in fast responses, $F(1, 182) = 821.50$, $p < .001$, $\eta_p^2 = .82$, $\eta_G^2 = .05$, compared to slow responses, $F(1, 182) = 196.80$, $p < .001$, $\eta_p^2 = .52$, $\eta_G^2 = .01$. We found the same pattern of results for ERs. Again the three-way interaction was significant, $F(1, 182) = 108.90$, $p < .001$, $\eta_p^2 = .37$, $\eta_G^2 = .04$, indicating a larger CSE in fast responses $F(1, 182) = 240.20$, $p < .001$, $\eta_p^2 = .57$, $\eta_G^2 = .14$, than in slow responses $F(1, 182) = 8.12$, $p = .005$, $\eta_p^2 = .04$, $\eta_G^2 = .01$.

Figure 4

Experimental Effects in the Stroop Task



Note. Reaction times (upper panel) and error rates (lower panel) are plotted as a function of Congruency, N-1 Congruency and Speed. Error bars represent 95% Confidence Intervals.

Reliability Analysis

Figures 5 and 6 show the distribution of SBP-corrected Pearson's *r* for the Simon and the Stroop task for both RTs and ERs. For each reliability coefficient, we report both the median SBP-corrected Pearson's coefficient r_c with their 95% CI in square brackets, followed by the uncorrected median r_{uc} with their 95% CI. When presenting reliabilities in the subsets we also report the bootstrapped *p*-value derived from the iterative comparison of r_c in that subset with r_c in the full set of trials (see *Statistical Analyses*).

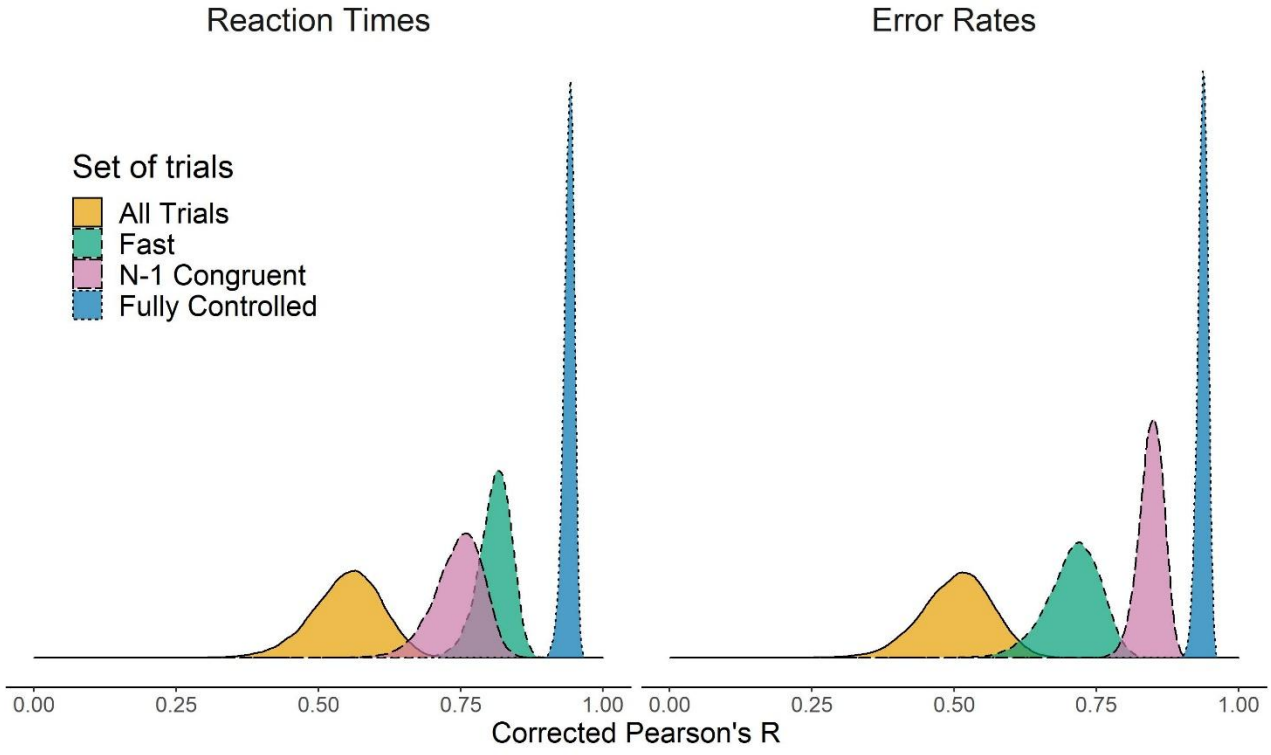
Simon Task

For the full set of trials, split-half reliability of the Simon RT congruency effect was low, r_c = .55, 95% CI [.42, .66], r_{uc} = .38 [.26, .49]. However, reliability was significantly improved when restricting the analysis to fast responses only, r_c = .81 [.74, .86], p = .002, r_{uc} = .52 [.42, .60]. A similar trend was observed when considering only N-1 congruent trials, even though the bootstrapped test did not reach significance r_c = .75 [.64, .82], p = .070, r_{uc} = .43 [.31, .53]. When controlling for both factors, split-half reliability was excellent, r_c = .94 [.92, .96], p < .001, r_{uc} = .67 [.59, .73].

Similar improvements of split-half reliability were observed for the Simon ER congruency effect. Again, split-half reliability was low when including all trials, r_c = .51 [.37, .61], r_{uc} = .34 [.23, .44]. When considering only fast responses, reliability was improved, albeit not significantly, r_c = .71 [.61, .79], p = .070, r_{uc} = .38 [.28, .49]. On the other hand, significant improvements were observed both for the post-congruent subset, r_c = .85 [.80, .88], p < .001, r_{uc} = .58 [.49, .65], and in the controlled subset, r_c = .94 [.92, .95], p < .001, r_{uc} = .65 [.58, .71].

Figure 5

Density Distributions of Corrected Reliability Coefficients in the Simon Task



Note. Bootstrapped density distributions of congruency-effect reliabilities as a function of the subset of trials used. See text for details.

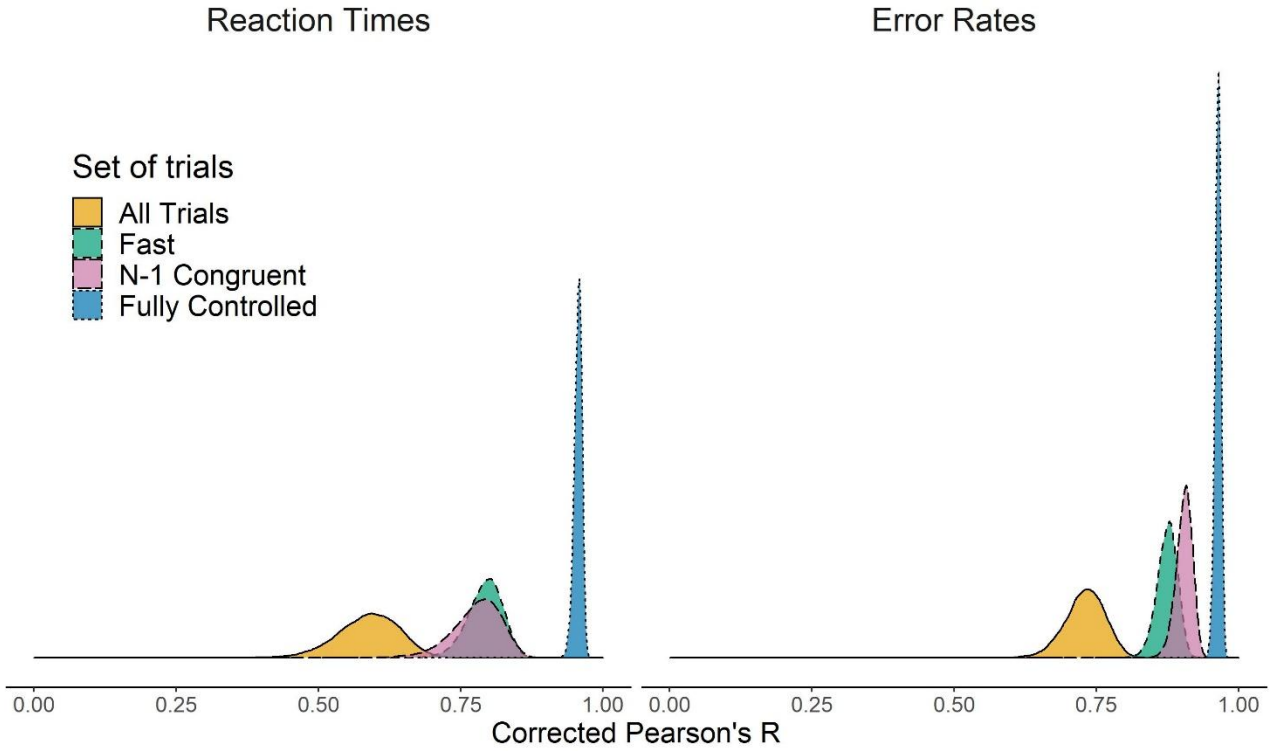
Stroop Task

In the Stroop task, we also found a low split-half reliability of the RT congruency effect when using the full set of trials, r_c = .59 [.47, .68], r_{uc} = .42 [.31, .52]. Reliability was significantly improved when considering fast responses only, r_c = .79 [.72, .85], p = .019, r_{uc} = .49 [.40, .58], and showed a trend for significant improvement with N-1 congruent trials only, r_c = .78 [.68, .85], p = .064, r_{uc} = .37 [.26, .48]. Reliabilities were excellent when controlling for both factors r_c = .96 [.94, .97], p < .001, r_{uc} = .65 [.57, .72].

In the ER, we found acceptable reliability for the whole sample dataset, r_c = .73 [.66, .79], r_{uc} = .58 [.49, .66]. Nonetheless, reliability was still significantly improved both when considering fast responses only, r_c = .88 [.84, .91], p = .005, r_{uc} = .64 [.56, .71], or N-1 congruent trials only, r_c = .91 [.87, .93], p < .001, r_{uc} = .61 [.53, .69]. Finally, when controlling for both factors, split-half reliability was again excellent, r_c = .96 [.95, .97], p < .001, r_{uc} = .69 [.63, .75].

Figure 6

Density Distributions of Corrected Reliability Coefficients in the Stroop Task



Note. Bootstrapped density distributions of congruency-effect reliabilities as a function of the subset of trials used. See text for details.

Between-Task Correlational Analysis

Having established that reliabilities were substantially improved when considering only specific subsets of the data, we moved on to test whether the low correlations between the Simon and the Stroop congruency effects commonly found in the literature are due to a lack of reliability or to a lack of true shared variance. For each correlation, we report both the corrected and uncorrected Pearson's r^2. We report no associated *p*-value because we do not aim to test the null-

[2] Because the proportion of removed trials differs in the Stroop and Simon tasks when controlling for N-1 Congruency (i.e., 1/2 and 2/3 respectively), the SBP formula takes the following form:

$$r_c = r_{xx} * \sqrt{\frac{N_1 * N_2}{(1+(N_1-1)*r_{xx}) * (1+(N_2-1)*r_{xx})}}, \quad (2)$$

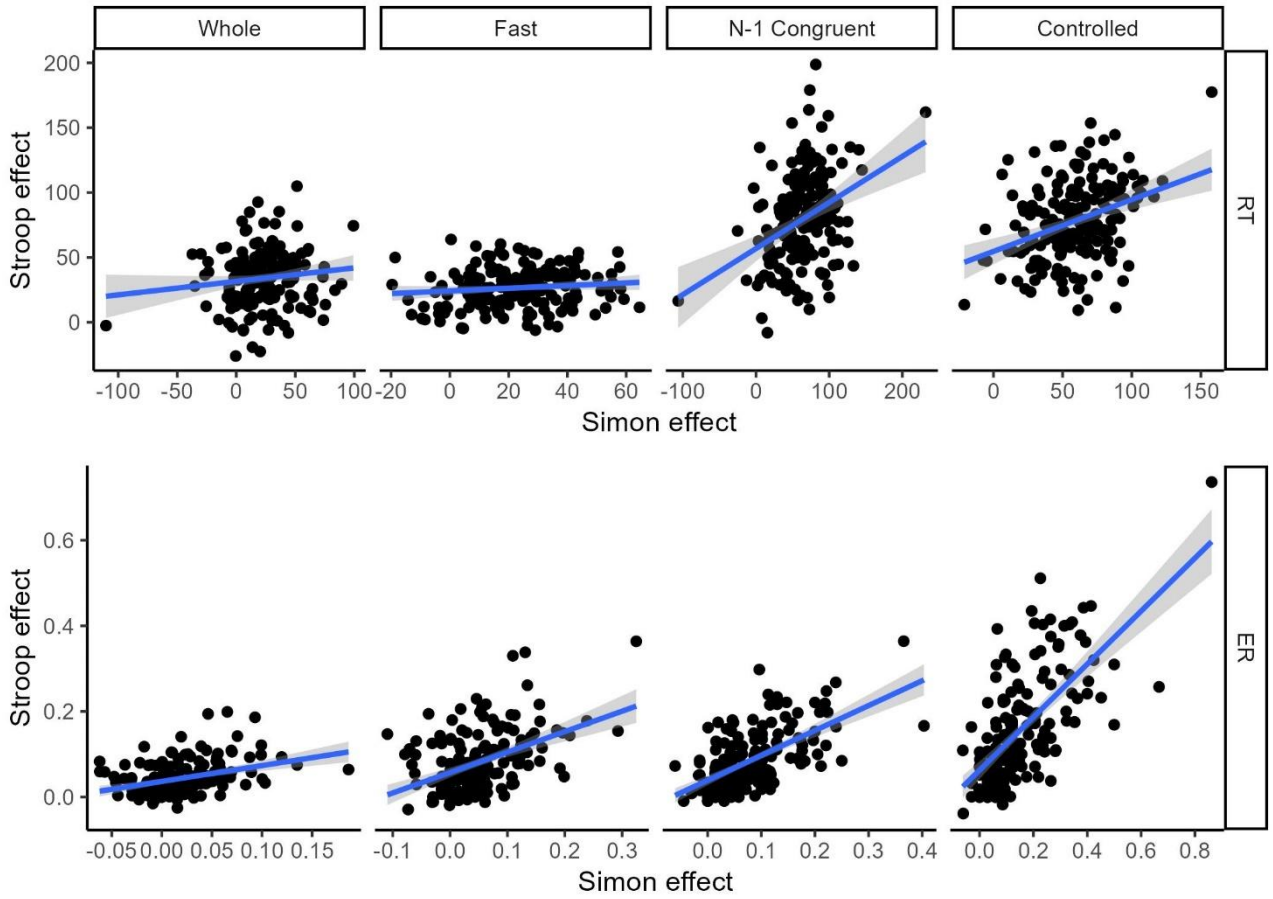
hypothesis of complete independence but instead to assess the size of the correlation coefficients. In this respect, correlations were considered to be low if r < .3, moderate for .3 > r > .5 and strong for r > .5, in line with the popular guidelines provided by Cohen (1988). For each subset, we report the difference between the observed correlation in that given subset (SBP-corrected), and the correlation observed in the full set of trials, with relative 95% CI. If the Δr 95% CI did not include 0 the difference was regarded as significant.

Figure 7 shows scatterplots of the uncorrected between-task correlations. Our results were robust to outliers exclusion, although it should be noted that correlations were slightly weaker, especially in the full set of trials (see S1 of the Supplementary Materials). We first computed Pearson's *r* for the full set of trials. In line with previous studies, we found a low correlation between congruency effects in RTs, $r(178) = .12$. In ERs, however, we did find a moderate correlation between the Stroop and Simon congruency effects, $r(178) = .36$. When considering fast responses only, the RT-based correlation did not change significantly, $r_c(178) = .20$, $\Delta r = .08$ 95% CI [-.11, .27], $r_{uc}(178) = .11$, but the ER-based correlation was significantly higher, $r_c(178) = .63$, $\Delta r = .27$ [.14, .40] $r_{uc}(178) = .46$. Including only N-1 congruent trials led to larger correlations being observed compared to the full set of trials, both in RTs, $r_c(178) = .58$, $\Delta r = .45$ [.29, .61], $r_{uc} = .36$, and in ERs, $r_c(178) = .80$, $\Delta r = .44$ [.33, .56], $r_{uc}(178) = .62$. Finally, when controlling for both factors, we observed very strong between-task correlations in both RTs, $r_c(178) = .72$, $\Delta r = .60$ [.45, .75], $r_{uc}(178) = .35$, and ERs, $r_c(178) = .90$, $\Delta r = .54$ [.44, .66], $r_{uc}(178) = .66$.

where N_1 and N_2 are the inverse of the proportion of removed trials for each subset of trials.

Figure 7

Uncorrected Correlations Between the Simon and the Stroop Congruency Effects as a Function of Trial Subset and Dependent Variable



Note. RT = reaction time; ER = error rate.

Between-Subset Variance Analysis

True between-subjects variance is fundamental for rank-ordering participants based on their individual differences. Any manipulation or analytical choice decreasing the true between-subject variance is, therefore, detrimental for observing acceptable split-half reliabilities and high between-tasks correlations (Hedge, Powell & Sumner, 2018). Although the use of difference scores has been criticized exactly for this reason (Draheim et al., 2019), in the present study we still found excellent reliabilities and between-task correlations for the congruency effect when analyzing only high-conflict trials. We thus tested whether the degree of response conflict not only impacts the size of the congruency effect, as shown in the ANOVA, but also its dispersion.

To do so, we fitted marginal models as implemented in the *gls* function of the nlme package in R (Pinheiro et al., 2017). Marginal models contain a fixed part only but no random effects which could model heteroscedasticity and dependencies between the repeated measures (Snijders & Bosker, 2011*)*. Therefore, both of these will emerge in the residuals. The *gls* function allows for modelling the residual covariance matrix and, therefore, for an assessment of heteroscedasticity and dependencies between the repeated measures. This is important as this approach allows to model different covariance structures of the residuals. In particular, the marginal models were fit to the data from the whole set of trials and another subset of interest. The resulting structure of the covariance matrix is as follows:

$$\begin{bmatrix} \sigma^2(Whole) & COV(Whole, Subset) \\ COV(Whole, Subset) & \sigma^2(Subset) \end{bmatrix} \quad (3)$$

For each pair of data sets, we fit a model where all elements in the covariance matrix were free to vary (i.e., the covariance matrix was unstructured), and one in which the variance parameters were kept constant across conditions (i.e., a compound symmetry covariance matrix). Next, we compared the models based on their difference in Akaike's information criterion (AIC) and the Bayesian information criterion (BIC). Models with a lower AIC and/or BIC were considered to fit substantially better if the difference with the other model was $\Delta > 2$ (Raftery, 1995). Because variance measures are largely influenced by outliers, we removed participants who displayed a clearly abnormal congruency effect in at least one subset of data before data analysis using the same approach as for calculating outlier-free correlations (see S1 of the Supplementary Material), resulting in the exclusion of four participants.

The results are summarized in Table 1. In general, variability was significantly larger in high-conflict subsets compared to the full set of trials in almost all conditions, except for the RTs of the subset only including fast responses. In this subset, as can be expected, variability was reduced

compared to the full set of trials. Furthermore, there was some evidence in RTs for the equality of variance between the controlled subset and the full set of trials, most likely also due to the limited variability in fast responses. This finding was limited to the Simon task, and evidence in favor of equality was only substantial when considering the BIC, but not the AIC. Taken together, this set of analyses showed that restricting the congruency effect only to trials with high conflict and, in particular, N-1 congruency trials, does not only increase its reliability and correlation with conflict effects from other tasks, but also its variability.

Table 1

Variance and Variance Comparisons as a Function of Subset

Set of trials	Simon			Stroop		
	S^2	ΔAIC	ΔBIC	S^2	ΔAIC	ΔBIC
RTs (ms)						
Full set	529			452		
Fast	271	39.7	35.8	218	44.2	40.4
N-1 Congruent	998	36.5	32.7	1227	75.4	71.6
Controlled	622	-0.4	-4.3	872	22.2	18.3
ERs (%)						
Full set	12			13		
Fast	36	207.5	203.7	43	282.2	278.3
N-1 Congruent	39	108.7	104.8	39	156.0	152.2
Controlled	136	302.2	298.4	137	370.6	366.8

Note. For each subset in each task and for each dependent variable the sample variance of the congruency effect (S^2) is reported. ΔAIC and ΔBIC refer to the model comparison between a marginal model assuming an unstructured covariance matrix and an identical model assuming compound symmetry in the covariance matrix. Positive values represent differences in AIC/BIC in favor of the unstructured model. The model pairs were fit to the data from all trials (full set) and from a subset of interest. If the unstructured model had the smallest AIC/BIC by a difference of $\Delta > 2$, we concluded that the variance in the subset of interest was significantly different from the whole sample.

Focusing on N-1 incongruent and slow trials

Above we have shown that focusing on N-1 congruent and fast trials improves congruency effect reliability in both the Simon and the Stroop tasks, irrespective of whether it is measured with RTs or error rates. Moreover, contrary to previous findings, such reliable measures correlate strongly between tasks, with the only exception of the RT congruency effect in the fast subset. To further strengthen our findings, we here replicate the same analyses focusing on the other portions of the dataset: Slow, and N-1 incongruent trials. Our reasoning is that neither slow responses, nor N-1 incongruent trials should adequately capture attentional control abilities. Therefore, we predict that neither reliabilities, nor between-task correlations should be improved in these subsets.

Reliability Analysis

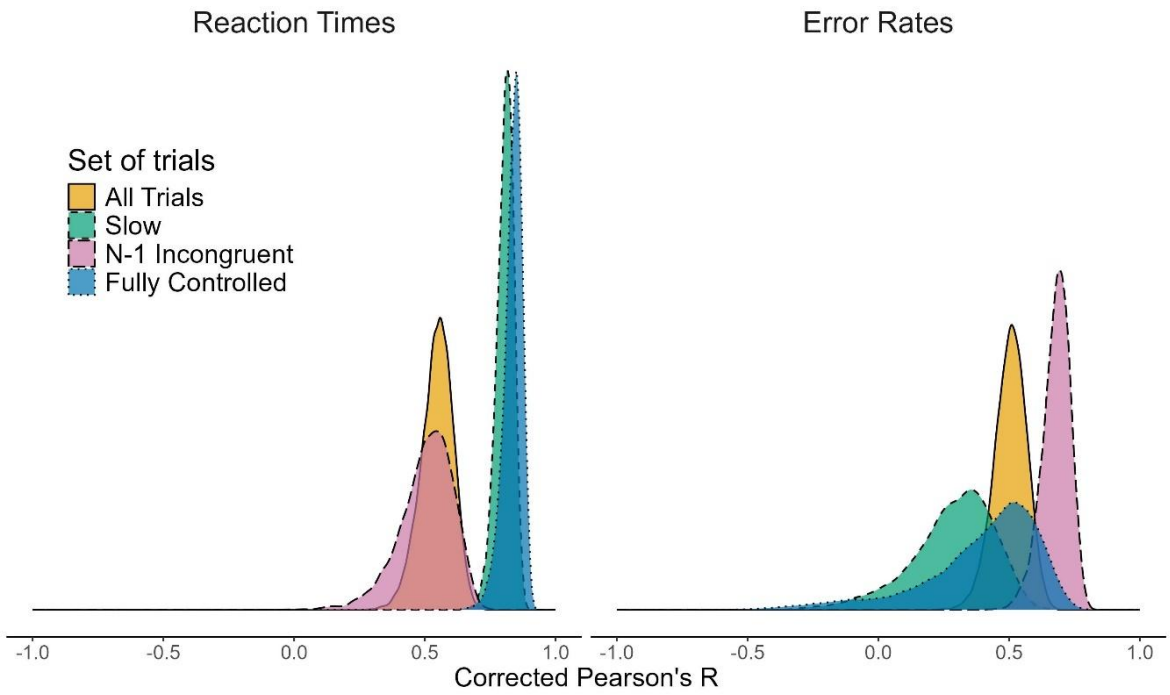
Simon Task

Figures 8 and 9 show the distribution of SBP-corrected Pearson's *r* for the Simon and the Stroop task for both RTs and ERs. As mentioned above, reliability in the full set of trials was r_c = .55, 95% CI [.42, .66], r_{uc} = .38 [.26, .49]. Reliability was very similar when restricting the analyses to post-incongruent trials, r_c = .52 [.27, .77], p = .913, r_{uc} = .21 [.08, .33]. Contrary to our expectations however, reliability significantly improved in the slow subsample, r_c = .81 [.73, .86], p = .003, r_{uc} = .52 [.41, .61], and when controlling for both factors, r_c = .84 [.75, .89], p = .002, r_{uc} = .40 [.27, .51].

In the ERs data, split-half reliability was low when including all trials, r_c = .51 [.37, .61], r_{uc} = .34 [.23, .44]. No significant difference was observed in the post-incongruent trials subset, r_c = .68 [.56, .77], p = .226, r_{uc} = .35 [.24, .45], or when focusing on slow responses, r_c = .31 [-.03, .53], p = .995, r_{uc} = .09 [-.03, .22]. Similarly, reliability was not improved when controlling for both factors, r_c = .45 [-.20, .68], p = .890, r_{uc} = .09 [-.03, .21].

Figure 8

Density Distributions of Corrected Reliability Coefficients in the Simon Task



Note. Bootstrapped density distributions of congruency-effect reliabilities as a function of the subset of trials used. See text for details.

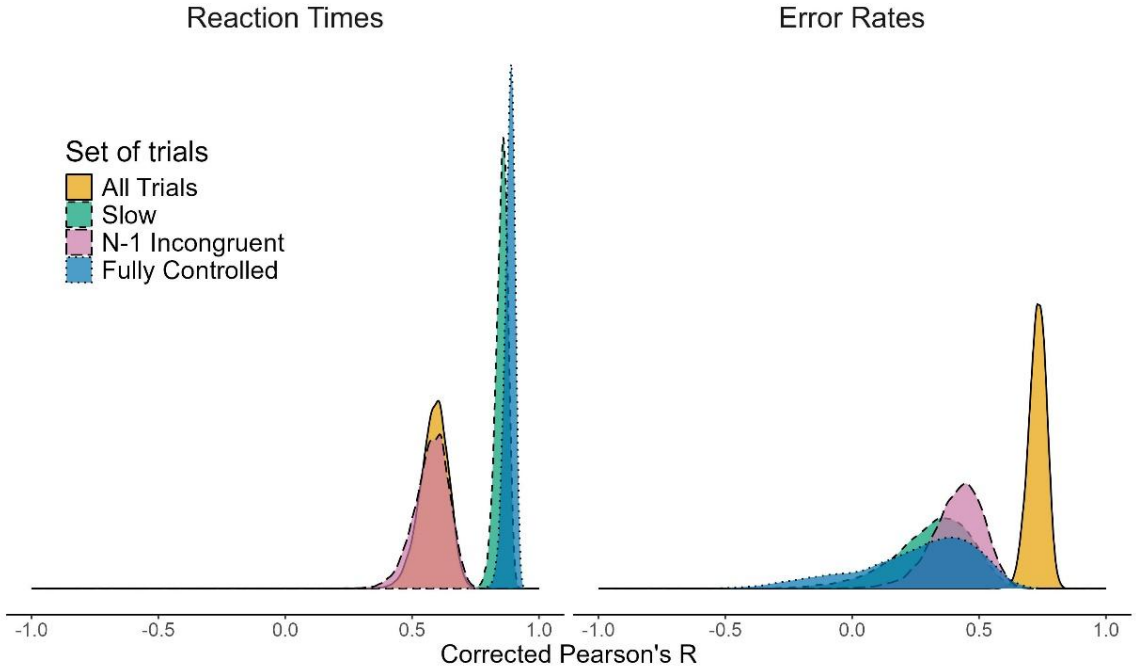
Stroop Task

As mentioned above, reliability of the RT congruency effect was low in the Stroop task, r_c = .59 [.47, .68], r_{uc} = .42 [.31, .52]. This value was virtually identical when restricting analyses to post-incongruent trials, r_c = .58 [.42, .69], p = .904, r_{uc} = .32 [.20, .43]. However, similarly to what observed in the Simon RTs, reliability significantly improved in the slow subset, r_c = .86 [.80, .89], p < .001, r_{uc} = .60 [.50, .68], and in the controlled dataset, r_c = .89 [.84, .91], p < .001, r_{uc} = .57 [.47, .65].

In the whole sample dataset reliability was acceptable for the ER congruency effect, r_c = .73 [.66, .79], r_{uc} = .58 [.49, .66]. Crucially, it did not improve when only including N-1 incongruent trials, r_c = .43 [.19, .59], p = 1, r_{uc} = .20 [.07, .32], or in the slow subset, r_c = .33 [-.05, .56], p = 1, r_{uc} = .11 [-.01, .24]. Finally, when controlling for both factors, split-half reliability was again poor, r_c = .30 [-.26, .59], p = 1, r_{uc} = .07 [-.05, .20].

Figure 9

Density Distributions of Corrected Reliability Coefficients in the Stroop Task



Note. Bootstrapped density distributions of congruency-effect reliabilities as a function of the subset of trials used. See text for details.

Between-task correlational analyses

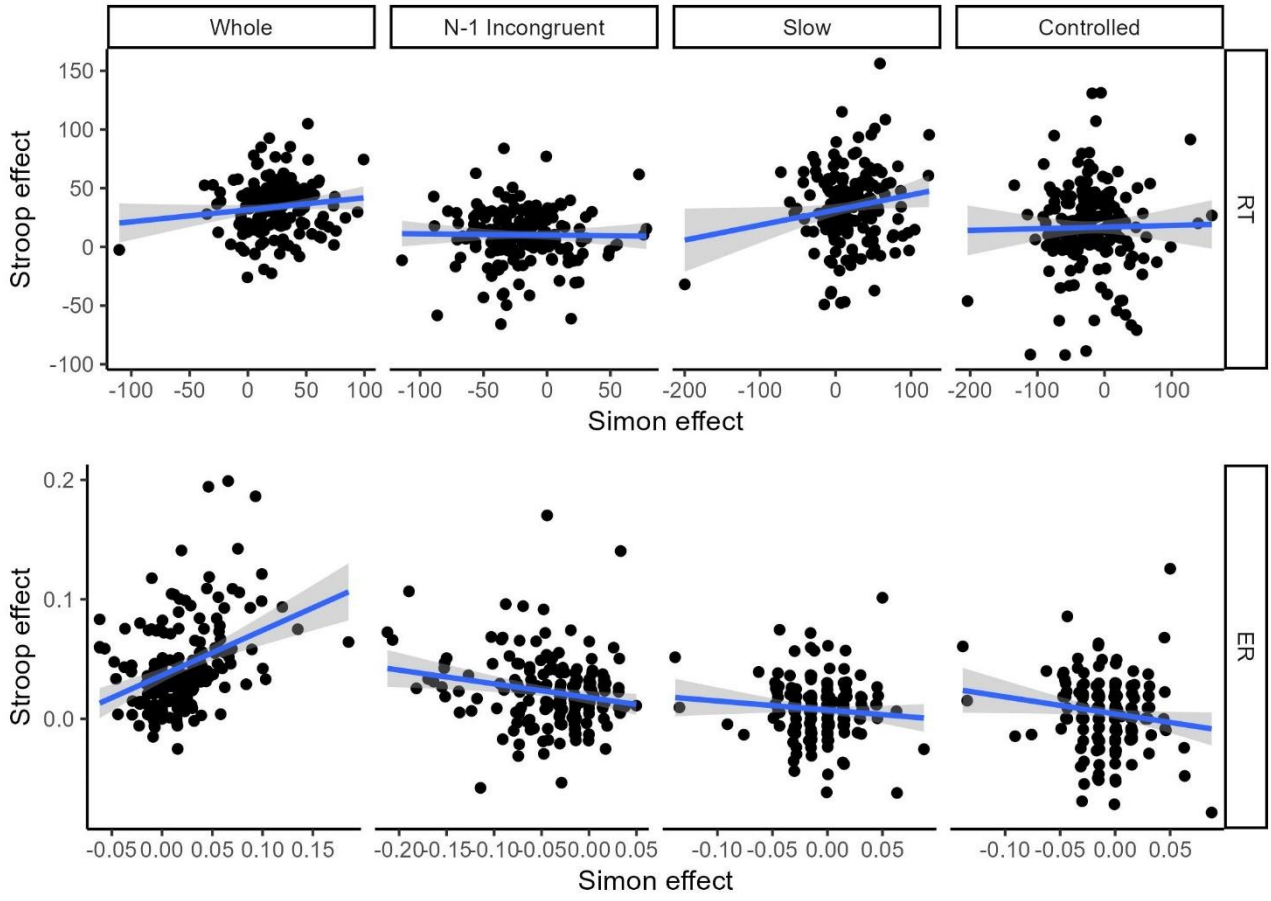
The reliability analyses showed that reliabilities generally did not improve when focusing on slow and N-1 incongruent trials, except for the RT congruency effect in the slow subset. Restricting the analyses to slow responses was also the only responsible for improved reliabilities in the RT congruency effect in the controlled sample. We thus wish to test whether focusing on these subsets would, contrary to our expectations, increase between-task correlations.

Figure 10 shows scatterplots of the uncorrected between-task correlations. As mentioned above, the RT congruency effect in the Simon and the Stroop tasks correlated poorly, $r(178) = .12$, whereas the correlation between ER congruency effects was moderate, $r(178) = .36$. When considering post-incongruent trials only, the RT-based correlation was next to null, $r_c(178) = .06$, $\Delta r = -.06$ 95% CI [-.28, .16], $r_{uc}(178) = -.02$. In the ERs, the between-task correlation turned to be negative, being significantly lower than in the whole sample, $r_c(178) = -.29$, $\Delta r = -.66$ 95% CI [-.87, -.43], $r_{uc}(178) = -.19$. When considering slow responses only, the RT-based correlation did not

change significantly, $r_c(178) = .27$, $\Delta r = .14$ 95% CI [-.04, .33], $r_{uc}(178) = .15$, whereas the ER-based correlation was significantly lower, $r_c(178) = -.17$, $\Delta r = -.54$ [-.72, -.35], $r_{uc}(178) = -.09$. Finally, when controlling for both factors, we observed correlations next to zero in the RTs, $r_c(178) = .06$, $\Delta r = -.06$ [-.28, .16], $r_{uc}(178) = .02$, and negative in the ERs, $r_c(178) = -.37$, $\Delta r = -.74$ [-.90, -.56], $r_{uc}(178) = -.15$.

Figure 10

Uncorrected Correlations Between the Simon and the Stroop Congruency Effects as a Function of Trial Subset and Dependent Variable



Note. RT = reaction time; ER = error rate.

Discussion

Recent years have seen an extensive debate over whether the commonly found low correlations between congruency effects in conflict tasks indicate poor measurement (Draheim et

al., 2019) or poor construct validity of attentional control ability (Rey-Mermet et al., 2018). In the present study, building on findings from experimental research (Gratton et al., 1992; Ridderinkhoff, 2002), we proposed that the congruency effect may indeed measure individual differences in attentional control but only on the subsets of trials where response conflict is largest, or present at all. Specifically, response conflict has been shown to be markedly reduced (1) following incongruent trials due to conflict adaptation (Botvinick et al., 2001), and (2) in responses long after stimulus onset due to selective suppression and/or decay of activation along the automatic route (Ridderinkhof, 2002). Therefore, we predicted that the split-half reliability of the congruency effect can be improved by using only the data from N-1 congruent trials and fast responses. Our results confirmed these predictions, although they also cast some doubts on the utility of focusing on fast trials when measuring the RT congruency effect.

Experimentally, we generally found stronger CSEs in fast responses relative to slow responses. This pattern held true for both tasks in the ERs, and for the Stroop task in the RTs. However, this three-way interaction was not present for RTs in the Simon task, possibly due to one or both of the following two reasons. First, whereas the CSE is somewhat weaker in slow trials in the Stroop task, the congruency effect is numerically reversed in post-incongruent trials both for fast and slow responses in the Simon task. Hence, control settings may be maintained, and thus carry over, more strongly in the Simon task compared to the Stroop task. Second, the difference in results could be due to different methodological choices in the Stroop and the Simon task. Specifically, the Stroop task comprised a higher proportion of incongruent trials. However, typically, a larger proportion of incongruent trials leads to more sustained control effects (Braver et al., 2012) and, therefore, the Stroop task, and not the Simon task, should display strong CSEs in both fast and slow trials. Importantly, the lack of a three-way interaction in Simon RTs is not critical for the interpretation of the present results, as we still observed that the congruency effect was reduced both in slow responses and following incongruent trials. Taken together, the results

from experimental analyses consistently indicate that both Speed and N-1 Congruency modulated response conflict in the expected direction.

In line with previous reports, split-half reliability of the congruency effect was poor when considering the full set of trials in both tasks for both RTs and ERs ($r_c \approx .50$), with the only exception of the Stroop ER effect ($r_c = .73$). Critically, however, such reliabilities were substantially improved when including only post-congruent trials (see also Schuch et al., 2022), or when focusing on fast responses. Finally, reliabilities were excellent ($r_c \approx .95$) when controlling for both factors. It is important to note that, for the Stroop task, both the experimental and the reliability findings were replicated when excluding partial repetitions of distractor and response (see S3 in the Supplementary Materials). This is a critical test to ensure that we are measuring attentional control abilities rather than episodic binding effects. When restricting the analyses to post-congruent trials, congruency effects may reflect the episodic mismatch of the previous response-distractor binding (i.e., "event file", Hommel, 2004) with the current trial (Frings et al., 2020; Hommel, 2004). This account was not supported in the three-choice Stroop task when excluding partial repetitions of distractor and response. Admittedly, as these additional analyses were limited to the three-choice spatial Stroop task, we cannot exclude that memory-based interference did play a role in the two-choice Simon task. However, to the best of our knowledge, there is no reason to believe that episodic binding and retrieval would occur any differently in these paradigms. Most importantly, the results of the correlational analyses strongly suggest that the congruency effect in these two tasks tap similar constructs. Therefore, we conclude that our approach is successful in measuring attentional control abilities rather than episodic binding. Furthermore, the inclusion of both two-choice and three-choice tasks demonstrates the general feasibility of using our approach irrespective of the number of response alternatives. This is an important feature of our design, as the number of response alternatives likely influences both attentional control processes per se as well as conflict adaptation phenomena (Bugg, 2014).

As we were successful in extracting reliable measures of attentional control in both the Simon and the spatial Stroop task, we next assessed the correlations between the congruency effects in these tasks for both RTs and ERs and the different subsets of trials. When using the full set of trials, RT congruency effects were poorly correlated (r = .12), and ER congruency effects were moderately correlated (r = .37). Critically, correlations for both RT and ER-based measures were substantially higher when excluding N-1 incongruent trials. ER-based congruency effects also correlated more strongly when focusing on fast responses. However, despite reliability increased in both tasks, their RT-based correlation remained low.

In order to strengthen our results, we conducted further analyses on the subsets of trials that do not require strong attentional control abilities (i.e., slow and N-1 incongruent trials). In line with our expectations, reliabilities did not generally improve in these subsets, with the exception of the RT congruency effect in slow responses (which also drove reliabilities in the controlled subsets to be higher). Importantly, between-task correlations remained low for all subsets, including the RT-based congruency effects computed on slow responses.

In sum, our findings unambiguously demonstrate (1) that congruency effects can exhibit excellent reliability when focusing on the subset of trials where response conflict is highest, and (2) that these more reliable – and arguably theoretically more valid – measures of attentional control yield strong correlations between conflict tasks, suggesting that the Simon and the Stroop tasks indeed tap a common ability. This conclusion comes with the caveat that focusing on either fast or slow responses held equivocal results for RT-based measures. In both subsets reliabilities significantly improved, whereas between-task correlations remained low. This seems to suggest that restricting the range of RTs allows to capture some construct reliably, which is however different from attentional control per se. It is possible for example that restricting the RT range would help in isolating processing speed, which however may not lead to high correlations between conflict effects across tasks. Given the ambiguity of these results, we suggest that focusing on fast or slow responses should be avoided to improve the reliability of RT-based congruency effects.

Overall, our data lend support to the position that the attentional control abilities measured in the conflict paradigms are domain-general, rather than being limited to the resolution of conflict in one or the other paradigm. At the same time, given the poor reliabilities and low correlation for the full set of trials, our results also substantiate the concerns that have been expressed about the validity of the congruency effect as it has usually been measured in cognitive individual differences research (e.g., Rey-Mermet et al., 2018; Rouder & Haaf, 2019). To increase both validity and reliability of attentional control measures, we suggest that future research should consider using the subset of data where response conflict is the strongest for computing attentional control scores. In what follows, we consider the feasibility of this approach.

Implications for Future Research

The main aim of the present study was to contribute to the existing debate on the validity of attentional control. However, we also illustrate an approach that future researchers may want to use to unravel the structure of attentional control, or its relation with other abilities. In what follows, we consider some implications that our approach holds for future research.

The first implication derived from our results is that quality matters even more than quantity. Reliabilities and between-task correlations were indeed increased in most subsets also when considering the uncorrected coefficients. This finding suggests that, although reducing the number of trials inevitably increases measurement error, this disadvantage is well compensated by the use of a more valid measure of attentional control. Nonetheless, researchers may still wish to maximize the number of trials that can be used for analysis. One possibility could be to design future studies including a high proportion of congruent trials, so to minimize data loss when controlling for N-1 congruency. For example, including 75% congruent trials more than doubles the number of cC trials while reducing by 25% the number of cI trials. Using 66% congruent trials increases the number of cC trials by 50% while losing very few cI trials (~ 10%). Our results showed that excluding N-1 incongruent trials increased both reliabilities and between-task correlations more so than using only fast responses, suggesting that controlling only for N-1 congruency may suffice (see also Schuch et

al., 2022). An additional benefit of increasing the proportion of congruent trials is that participants do not predict, and thus do not prepare for conflict occurrence (Braver, 2007). The relaxation of such proactive mode of control results in larger conflict when incongruent trials are actually encountered (Braver, 2007; Braver et al., 2012; Bugg & Crump, 2012), which, in line with the evidence reported in the present study, may also be beneficial for the reliability and validity of the congruency effect.

A second implication that deserves attention concerns the use of SB correction. The SB correction provides a useful tool for comparing correlations between sets of trials of differing length and should thus be used whenever researchers are interested in such a comparison (as in the present manuscript). This approach is equivalent to the standard practice of using SB correction with N = 2 when reporting split-half reliability. In both cases, the correction is applied to correct for the intrinsic data loss related to subsetting. However, researchers interested in correlating attentional control measures with other constructs should consider reporting uncorrected between-task correlations. Under these circumstances, corrected coefficients would only indicate the correlation that is expected by administering N times more trials and would therefore be of little practical use. The between-task correlations were moderate to strong when excluding N-1 incongruent trials, even for uncorrected coefficients (r = .36 and r = .62 in RTs and ERs respectively). Hence, excluding N-1 incongruent trials may be enough to observe robust between-task correlations even when running experiments of limited length. In the present study, the Simon task and the three-choice Stroop task consisted of 288 and 540 trials, respectively. Therefore, excluding N-1 incongruent trials left 144 and 180 trials. Experiments designed to increase the proportion of congruent trials may even be shorter than the ones reported here, being in the order of 200-300 trials.

To summarize, our approach does two things. First, it encourages to define more precisely what construct one wishes to measure. Second, it excludes the conditions in which the construct of interest is poorly measured. These guiding principles can be extended beyond response-conflict tasks to improve the psychometric properties of other attentional control measures. In this regard,

our theoretically-driven approach differs from other solutions proposed in the literature, which focused on improving reliability by modifying existing conflict paradigms (e.g., Burgoyne et al., 2023; Draheim et al., 2021; 2023). For example, in the task switching paradigm, participants are asked to alternate between simple classification tasks, such as indicating the stimulus' color or its shape, as indicated on each trial by a task cue (for recent reviews, see: Koch et al., 2018; Koch & Kiesel, 2022). On the surface level, task switching is similar to other conflict paradigms in that the employed stimuli simultaneously trigger two cognitive representations, in this case two conflicting tasks (Moretti et al., 2023b; Waszak et al., 2003). Therefore, on the one hand, the task switching paradigm can be used for measuring the participants' ability to deal with such stimulus-based *task conflict*, and possibly correlate it with the participants' ability to deal with stimulus-based *response conflict* in the Stroop task. However, the switch cost, despite being the most frequently used measure in task-switching research, may be unapt to this purpose. The reason is that, in addition to the processes needed to counteract stimulus-based task conflict, the switch cost also measures pre-stimulus processes such as cue-based task-set reconfiguration (Rogers & Monsell, 1995). Our approach suggests minimizing the contribution of task-set reconfiguration by using a long interval between the cue and the stimulus. It is conceivable indeed that, with enough time to prepare, each participant will have completed task-set reconfiguration by the time the stimulus is presented and the switch cost will be a better measure of post-stimulus processes. On the other hand, if one is interested in assessing interindividual-differences in reconfiguration processes, these will be more easily detectable in performance if a short cue-stimulus interval is provided. To conclude, our approach can be generalized to other attentional control tasks by focusing on those conditions where the construct of interest contributes maximally to performance costs.

Methodological Considerations

Different aspects of our results speak to the methodological considerations outlined in the introduction. First, several authors have stressed the importance of between-participants variability for obtaining reliable cognitive measures (Hedge, Powell & Sumner, 2018; Kucina et al., 2022). In

our study, we found that controlling for N-1 Congruency and/or Speed substantially increased variability in difference scores. Notably, this effect is not attributable to an increase in error variance due to trials exclusion, as uncorrected reliabilities were higher in these subsets compared to the full set of trials. Therefore, true between-participants variability seems to be increased in high-conflict trials. Presumably, this increase in variability is due to using a more valid measure of attentional control, which is able to capture true variability in this ability. However, we cannot rule out that the larger mean congruency effect underpins, or at least contributes to, the increase in variability. Critically, though, the only subset that showed reduced variance compared to the full set of trials was, as can be expected, the RT subset including fast responses only. Interestingly, this was also the only subset in which between-tasks correlations did not improve. These results highlight, once again, the importance of taking between-participant variability into account when conducting interindividual differences research (Hedge, Powell & Sumner, 2018).

Another methodological aspect previously discussed concerns the type of dependent variable that should be used in interindividual differences research. It has been argued that RT measures are unsuitable to this purpose (Draheim et al., 2019), leading to the suggestion to develop tasks in which the speed component is experimentally controlled rather than measured (e.g., using staircase procedures, Draheim et al., 2021). Indeed, in our data, inter-task correlations were generally higher for ERs, particularly when analyzing the full set of trials, the condition which compares best to previous research. Similarly, whereas RTs reliabilities were consistently poor in the full set of trials ($r < .6$), ER reliabilities were acceptable for the Simon congruency effect. However, when focusing on high-conflict trials, RT measures were also found to be reliable, and inter-task correlations were generally high. Therefore, although we agree that relying on both measures can complicate the interpretation of results due to SATs (Hedge et al., 2021), our data support that RT measures are overall as suitable as accuracy measures for interindividual differences research.

Another source of low reliability in conflict tasks is trial-to-trial variability. Because congruency effects are usually small in absolute size (especially in RTs), even slight variability may have serious consequences for their psychometric properties. By focusing on high-conflict trials, we may have reduced the impact of variability simply by increasing the absolute size of our effects. Yet, even larger effects would still be unreliable if error variance would proportionally increase. Here, we found that reliabilities and between-task correlations in the subsets were higher compared to the full set of trials even without correcting for data loss (i.e., when comparing the uncorrected coefficients). Hence, with the present approach, above and beyond an increase in observed variance in the subsets, we found that the true variance increased more than the error variance.

Conclusions

In recent years, the debate on why congruency effects correlate poorly across conflict tasks has received considerable attention (e.g., Rey-Mermet et al., 2018). Whereas some authors have argued that these poor correlations suggest that there is simply no domain-general attentional control ability (Rey-Mermet et al., 2018), or that, at best, it cannot be measured with congruency effects in conflict tasks (Rouder & Haaf, 2019), others have argued that the problems are rooted in the methodological shortcomings of using experimental-research paradigms in interindividual differences research (Draheim et al., 2019; Hedge, Powell & Sumner, 2018). The truth seems to fall in-between, with the findings of the present study showing that the low correlations are due to both theoretical and methodological problems: the congruency effect reliably measures attentional control only when response conflict is highest. Therefore, we conclude that a common attentional control mechanism is indeed employed in the different conflict tasks, and that it can be reliably measured when using the appropriate subset of trials. We are hopeful that such an approach to validity will be fruitful for future interindividual differences research, and that it will foster discussion on the boundary conditions under which cognitive control tasks truly elicit conflict.

Constraint on Generality

Our results provide evidence that the congruency effects in a Simon task and in a spatial Stroop task are reliable measures tapping a similar construct, presumably attentional control. This only holds to be true when focusing on trials where response conflict is highest (i.e. on post-congruent and on fast trials). We expect our findings to generalize to other conflict tasks (e.g. the flanker task), as well as to other stimulus materials (e.g. the color-word Stroop task). Furthermore, as previous research has shown that congruency effects in conflict tasks do not differ between in-lab and online settings (Crump et al., 2013), we expect our results to be replicable in the laboratory. However, given that our sample was made of young (age range: 18-36 years) healthy participants, we do not exclude that different results may be produced using older samples, or populations known to differ in cognitive control abilities (e.g. Schizophrenia patients; Lesh et al., 2011). We have no reason to believe that the results depend on other characteristics of the participants, materials, or context.

References

Baler, R. D., & Volkow, N. D. (2006). Drug addiction: the neurobiology of disrupted self-control. *Trends in Molecular Medicine*, *12*(12), 559-566. https://doi.org/10.1016/j.molmed.2006.10.005

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. http://dx.doi.org/10.1177/0963721411408885

Blair, C. (2002). School readiness: Integrating cognition and emotion in a neurobiological conceptualization of children's functioning at school entry. *American Psychologist*, *57*(2), 111. https://doi.org/10.1037/0003-066X.57.2.111

Borella, E., Carretti, B., & Pelegrina, S. (2010). The specific role of inhibition in reading comprehension in good and poor comprehenders. *Journal of Learning Disabilities*, *43*(6), 541-552. https://doi.org/10.1177/0022219410371676

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652. http://dx.doi.org/10.1037/0033-295X.108.3.624

Braem, S., Bugg, J. M., Schmidt, J. R., Crump, M. J., Weissman, D. H., Notebaert, W., & Egner, T. (2019). Measuring adaptive control in conflict tasks. *Trends in Cognitive Sciences*, *23*(9), 769-783. http://dx.doi.org/10.1016/j.tics.2019.07.002

Braver, T. S., Gray, J. R., & Burgess, G. C. (2007). Explaining the many varieties of working memory variation: Dual mechanisms of cognitive control. In A. R. A. Conway, C. Jarrold, M. J. Kane (Eds.) & A. Miyake & J. N. Towse (Ed.), *Variation in working memory* (pp. 76–106). Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195168648.003.0004

Braver, T. S. (2012). The variable nature of cognitive control: A dual mechanisms framework. *Trends in Cognitive Sciences*, *16*(2), 106–113. http://dx.doi.org/10.1016/j.tics.2011.12.010

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, *2*(1). http://dx.doi.org/10.5334/joc.72

Bugg, J. M. (2014). Conflict-triggered top-down control: Default mode, last resort, or no such thing? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 567–587. https://doi.org/10.1037/a0035032

Bugg, J. M., & Crump, M. J. (2012). In support of a distinction between voluntary and stimulus-driven control: A review of the literature on proportion congruent effects. *Frontiers in Psychology*, *3*, 367. http://dx.doi.org/10.3389/fpsyg.2012.00367

Burgoyne, A. P., Tsukahara, J. S., Mashburn, C. A., Pak, R., & Engle, R. W. (2023). Nature and measurement of attention control. *Journal of Experimental Psychology: General*, *152*(8), 2369–2402. https://doi.org/10.1037/xge0001408

Clayson, P. E., & Larson, M. J. (2013). Psychometric properties of conflict monitoring and conflict adaptation indices: Response time and conflict N 2 event-related potentials. *Psychophysiology*, *50*(12), 1209-1219. http://dx.doi.org/10.1111/psyp.12138

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum. http://dx.doi.org/10.4324/9780203771587

Cooper, S. R., Gonthier, C., Barch, D. M., & Braver, T. S. (2017). The role of psychometrics in individual differences research in cognition: A case study of the AX-CPT. *Frontiers in Psychology*, *8*, 1482. http://dx.doi.org/10.3389/fpsyg.2017.01482

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS ONE*, *8*(3), e57410. https://doi.org/10.1371/journal.pone.0057410

Declerck, M., & Koch, I. (2023). The concept of inhibition in bilingual control. *Psychological Review*, *130*(4), 953–976. https://doi.org/10.1037/rev0000367

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*(4), 731–750. http://dx.doi.org/10.1037/0096-1523.20.4.731

De Simoni, C., & von Bastian, C. C. (2018). Working memory updating and binding training: Bayesian evidence supporting the absence of transfer. *Journal of Experimental Psychology: General*, *147*(6), 829-858. http://dx.doi.org/10.1037/xge0000453

Diamond, A. (2016). Why improving and assessing executive functions early in life is critical. In J. A. Griffin, P. McCardle, & L. S. Freund (Eds.), *Executive function in preschool-age children: Integrating measurement, neurodevelopment, and translational research* (pp. 11–43). American Psychological Association. https://doi.org/10.1037/14797-002

Dinges, D. F., & Powell, J. W. (1985). Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. *Behavior Research Methods, Instruments, & Computers*, 17(6), 652–655. https://doi.org/10.3758/BF03200977

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

Draheim, C., Tsukahara, J. S., Martin, J. D., Mashburn, C. A., & Engle, R. W. (2021). A toolbox approach to improving the measurement of attention control. *Journal of Experimental Psychology: General*, *150*(2), 242-275. http://dx.doi.org/10.1037/xge0000783

Drost, E. A. (2011). Validity and reliability in social science research. *Education Research and Perspectives*, *38*(1), 105-123. https://search.informit.org/doi/10.3316/informit.491551710186460

Duthoo, W., Abrahamse, E. L., Braem, S., & Notebaert, W. (2014). Going, going, gone? Proactive control prevents the congruency sequence effect from rapid decay. *Psychological Research*, *78*(4), 483-493. http://dx.doi.org/10.1007/s00426-013-0498-4

Egner, T., Ely, S., & Grinband, J. (2010). Going, going, gone: characterizing the time-course of congruency sequence effects. *Frontiers in Psychology*, *1*, 154. http://dx.doi.org/10.3389/fpsyg.2010.00154

Egner, T. (2008). Multiple conflict-driven control mechanisms in the human brain. *Trends in Cognitive Sciences*, *12*(10), 374–380. http://dx.doi.org/10.1016/j.tics.2008.07.001

Egner, T. (2017). Past, present, and future of the congruency sequence effect as an index of cognitive control. In T. Egner (Ed.), *The Wiley handbook of cognitive control* (pp. 64-78). Wiley-Blackwell. https://doi.org/10.1002/9781118920497.ch4

Feldman, J. L., & Freitas, A. L. (2016). An investigation of the reliability and self-regulatory correlates of conflict adaptation. *Experimental Psychology*, *63*(4), 237–247. http://dx.doi.org/10.1027/1618-3169/a000328

Friedman, N. P., & Miyake, A. (2004). The relations among inhibition and interference control functions: a latent-variable analysis. *Journal of Experimental Psychology: General*, *133*(1), 101-135. http://dx.doi.org/10.1037/0096-3445.133.1.101

Friedman, N. P., & Miyake, A. (2017). Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex*, *86*, 186-204. https://doi.org/10.1016/j.cortex.2016.04.023

Friedman, N. P., Miyake, A., Altamirano, L. J., Corley, R. P., Young, S. E., Rhea, S. A., & Hewitt, J. K. (2016). Stability and change in executive function abilities from late adolescence to early adulthood: A longitudinal twin study. *Developmental Psychology*, *52*(2), 326–340. http://dx.doi.org/10.1037/dev0000075

Frischkorn, G. T., Schubert, A.-L., & Hagemann, D. (2019). Processing speed, working memory, and executive functions: Independent or inter-related predictors of general intelligence. *Intelligence*, *75*, 95–110. http://dx.doi.org/10.1016/j.intell.2019.05.003

Frings, C., Hommel, B., Koch, I., Rothermund, K., Dignath, D., Giesen, C., Kiesel, A., Kunde, W., Mayr, S., Moeller, B., Möller, M., Pfister, R., & Philipp, A. M. (2020). Binding and

Retrieval in Action Control (BRAC). *Trends in Cognitive Sciences, 24(5),* 375-387. http://dx.doi.org/10.1016/j.tics.2020.02.004

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*(4), 480–506. http://dx.doi.org/10.1037/0096-3445.121.4.480

Gustavson, D. E., Panizzon, M. S., Franz, C. E., Friedman, N. P., Reynolds, C. A., Jacobson, K. C., Xian, H., Lyons, M. J., & Kremen, W. S. (2018). Genetic and environmental architecture of executive functions in midlife. *Neuropsychology*, *32*(1), 18-30. http://dx.doi.org/10.1037/neu0000389

Hedge, C., Powell, G., Bompas, A., & Sumner, P. (2022). Strategy and processing speed eclipse individual differences in control ability in conflict tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(10), 1448–1469. http://dx.doi.org/10.1037/xlm0001028

Hedge, C., Powell, G., Bompas, A., Vivian-Griffiths, S., & Sumner, P. (2018). Low and variable correlation between reaction time costs and accuracy costs explained by accumulation models: Meta-analysis and simulations. *Psychological Bulletin*, *144*(11), 1200–1227. http://dx.doi.org/10.1037/bul0000164

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

Hedge, C., Vivian-Griffiths, S., Powell, G., Bompas, A., & Sumner, P. (2019). Slow and steady? Strategic adjustments in response caution are moderately reliable and correlate across tasks. *Consciousness and Cognition*, *75*, 102797. http://dx.doi.org/10.1016/j.concog.2019.102797

Hommel, B. (2004). Event files: Feature binding in and across perception and action. *Trends in Cognitive Sciences*, *8*(11), 494-500. http://dx.doi.org/10.1016/j.tics.2004.08.007

Hommel, B. (2011). The Simon effect as tool and heuristic. *Acta Psychologica*, *136*(2), 189-202. http://dx.doi.org/10.1016/j.actpsy.2010.04.011

Hübner, R., & Töbel, L. (2019). Conflict resolution in the Eriksen flanker task: Similarities and differences to the Simon task. *PloS ONE*, *14*(3), e0214203. http://dx.doi.org/10.1371/journal.pone.0214203

Kałamała, P., Szewczyk, J., Chuderski, A., Senderecka, M., & Wodniecka, Z. (2020). Patterns of bilingual language use and response inhibition: A test of the adaptive control hypothesis. *Cognition*, *204*, 104373. https://doi.org/10.1016/j.cognition.2020.104373

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review*, *9*(4), 637-671. http://dx.doi.org/10.3758/BF03196323

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*(1), 47–70. http://dx.doi.org/10.1037/0096-3445.132.1.47

Kane, M. J., Meier, M. E., Smeekens, B. A., Gross, G. M., Chun, C. A., Silvia, P. J., & Kwapil, T. R. (2016). Individual differences in the executive control of attention, memory, and thought, and their associations with schizotypy. *Journal of Experimental Psychology: General*, *145*(8), 1017–1048. https://doi.org/10.1037/xge0000184

Kerns, J. G., Cohen, J. D., MacDonald III, A. W., Cho, R. Y., Stenger, V. A., & Carter, C. S. (2004). Anterior cingulate conflict monitoring and adjustments in control. *Science*, *303*(5660), 1023-1026. http://dx.doi.org/10.1126/science.1089910

Keye, D., Wilhelm, O., Oberauer, K., & Van Ravenzwaaij, D. (2009). Individual differences in conflict-monitoring: testing means and covariance hypothesis about the Simon and the

Eriksen Flanker task. *Psychological Research*, *73*(6), 762-776. http://dx.doi.org/10.1007/s00426-009-0257-8

Koch, I., & Kiesel, A. (2022). Task switching: Cognitive control in sequential multitasking. In A. Kiesel, L. Johannsen, I. Koch, & H. Müller (Eds.), *Handbook of human multitasking* (pp. 85–143). Springer. https://doi.org/10.1007/978-3-031-04760-2_3

Koch, I., Poljac, E., Müller, H., & Kiesel, A. (2018). Cognitive structure, flexibility, and plasticity in human multitasking—An integrative review of dual-task and task-switching research. *Psychological Bulletin*, *144*(6), 557. https://doi.org/10.1037/bul0000144

Kucina, T., Wells, L., Lewis, I., de Salas, K., Kohl, A., Palmer, M., ... & Heathcote, A. (2022). A solution to the reliability paradox for decision-conflict tasks. PsyArXiv. Doi: 10.31234/osf.io/bc6nk

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863. http://dx.doi.org/10.3389/fpsyg.2013.00863

Lesh, T. A., Niendam, T. A., Minzenberg, M. J., & Carter, C. S. (2011). Cognitive control deficits in schizophrenia: mechanisms and meaning. *Neuropsychopharmacology*, *36*(1), 316-338. https://doi.org/10.1038/npp.2010.156

Littman, R., Hochman, S. & Kalanthroff, E. (2023). Reliable affordances: A generative modeling approach for test-retest reliability of the affordances task. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-023-02131-3

Löffler, C., Frischkorn, G.T., Hagemann, D., Sadus, K. & Schubert, A. -L. (2024) The common factor of executive functions measures nothing but speed of information uptake. *Psychological Research*, 88, 1092–1114. https://doi.org/10.1007/s00426-023-01924-7

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

Mayr, U., Awh, E., & Laurey, P. (2003). Conflict adaptation effects in the absence of executive control. *Nature Neuroscience*, *6*(5), 450-452. http://dx.doi.org/10.1038/nn1051

Meier, M. E., & Kane, M. J. (2013). Working memory capacity and Stroop interference: Global versus local indices of executive control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 748–759. http://dx.doi.org/10.1037/a0029200

Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, *111*(1), 172-175. http://dx.doi.org/10.1037/0033-2909.111.1.172

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49-100. https://doi.org/10.1006/cogp.1999.0734

Moretti, L. (2023, March 15). Raw data and analysis scripts of "Quality Over Quantity: Focusing on High-Conflict Trials to Improve the Reliability and Validity of Attentional Control Measures". Retrieved from https://doi.org/10.23668/psycharchives.12584

Moretti, L., Koch, I., Steinhauser, M., & Schuch, S. (2021). Errors in task switching: Investigating error aftereffects in a N-2 repetition cost paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(10), 1720–1737. https://doi.org/10.1037/xlm0001034

Moretti, L., Koch, I., Steinhauser, M., & Schuch, S. (2023a). Disentangling task-selection failures from task-execution failures in task switching: an assessment of different paradigms. *Psychological Research*, *87*(3), 929-950. https://doi.org/10.1007/s00426-022-01708-5

Moretti, L., Koch, I., Steinhauser, M. & Schuch, S. (2023b). Stimulus-triggered task conflict affects task-selection errors in task switching: A Bayesian multinomial processing tree modeling approach. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advanced online publication. https://doi.org/10.1037/xlm0001245

Nigg, J. T. (2000). On inhibition/disinhibition in developmental psychopathology: Views from cognitive and personality psychology and a working inhibition taxonomy. *Psychological Bulletin*, *126*(2), 220–246. https://doi.org/10.1037/0033-2909.126.2.220

Nunnally, J. C. (1978). An overview of psychological measurement. In B. Wolfman (Ed.), *Clinical Diagnosis of Mental Disorders: A Handbook* (97-146). Springer. http://dx.doi.org/10.1007/978-1-4684-2490-4_4

Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Advances in Methods and Practices in Psychological Science*, *2*(4), 378–395. http://dx.doi.org/10.1177/2515245919879695

Paap, K. R., Anders-Jefferson, R., Zimiga, B., Mason, L., & Mikulinsky, R. (2020). Interference scores have inadequate concurrent and convergent validity: Should we stop using the flanker, Simon, and spatial Stroop tasks? *Cognitive Research: Principles and Implications*, *5*(1), 1-27. http://dx.doi.org/10.1186/s41235-020-0207-y

Paap, K. R., & Greenberg, Z. I. (2013). There is no coherent evidence for a bilingual advantage in executive processing. *Cognitive Psychology*, *66*, 232–258. http://dx.doi.org/10.1016/j.cogpsych.2012.12.002

Paap, K. R., & Sawi, O. (2016). The role of test-retest reliability in measuring individual and group differences in executive functioning. *Journal of Neuroscience Methods*, *274*, 81-93. http://dx.doi.org/10.1016/j.jneumeth.2016.10.002

Pettigrew, C., & Martin, R. C. (2014). Cognitive declines in healthy aging: Evidence from multiple aspects of interference resolution. *Psychology and Aging*, *29*(2), 187–204. http://dx.doi.org/10.1037/a0036085

Pievsky, M. A., & McGrath, R. E. (2018). The neurocognitive profile of attention-deficit/hyperactivity disorder: A review of meta-analyses. *Archives of Clinical Neuropsychology*, *33*(2), 143-157. https://doi.org/10.1093/arclin/acx055

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., Heisterkamp, S., Van Willigen, B., & Maintainer, R. (2017). Package 'nlme'. *Linear and nonlinear mixed effects models, version*, *3*(1), 274.

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*(7), 2013–2025. http://dx.doi.org/10.3758/APP.72.7.2013

Proctor, R. W., Miles, J. D., & Baroni, G. (2011). Reaction time distribution analysis of spatial correspondence effects. *Psychonomic Bulletin & Review*, *18*(2), 242-266. http://dx.doi.org/10.3758/s13423-011-0053-5

Pronk, T., Molenaar, D., Wiers, R. W., & Murre, J. (2022). Methods to split cognitive task data for estimating split-half reliability: A comprehensive review and systematic assessment. *Psychonomic Bulletin & Review*, *29*(1), 44-54. https://doi.org/10.3758/s13423-021-01948-3

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 111-163.

Redick, T. S., Shipstead, Z., Meier, M. E., Montroy, J. J., Hicks, K. L., Unsworth, N., Kane, M. J., Hambrick, D. Z., & Engle, R. W. (2016). Cognitive predictors of a common multitasking ability: Contributions from working memory, attention control, and fluid intelligence. *Journal of Experimental Psychology: General*, *145*(11), 1473–1492. http://dx.doi.org/10.1037/xge0000219

Rey-Mermet, A., Gade, M., & Oberauer, K. (2018). Should we stop thinking about inhibition? Searching for individual and age differences in inhibition ability. *Journal of Experimental*

Psychology: Learning, Memory, and Cognition, *44*(4), 501–526. http://dx.doi.org/10.1037/xlm0000450

Rey-Mermet, A., Gade, M., Souza, A. S., von Bastian, C. C., & Oberauer, K. (2019). Is executive control related to working memory capacity and fluid intelligence? *Journal of Experimental Psychology: General*, *148*(8), 1335–1372. http://dx.doi.org/10.1037/xge0000593

Rey-Mermet, A., Singmann, H., & Oberauer, K. (2021). Neither measurement error nor speed–accuracy trade-offs explain the difficulty of establishing attentional control as a psychometric construct: Evidence from a latent-variable analysis using diffusion modeling. PsyArXiv. https://doi.org/10.31234/osf.io/3h26y

Ridderinkhof, K. R. (2002). Micro- and macro-adjustments of task set: activation and suppression in conflict tasks. *Psychological Research*, *66*(4), 312–323. http://dx.doi.org/10.1007/s00426-002-0104-7

Ridderinkhof, K. R., van den Wildenberg, W. P., Wijnen, J., & Burle, B. (2004). Response inhibition in conflict tasks is revealed in delta plots. In M. I. Posner (Ed.), *Cognitive neuroscience of attention* (pp. 369-377). The Guilford Press.

Rogers, R. D., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General, 124(2)*, 207-231. http://dx.doi.org/10.1037/0096-3445.124.2.207

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review*, *26*(2), 452–467. http://dx.doi.org/10.3758/s13423-018-1558-y

Rouder, J., Kumar, A., & Haaf, J. M. (2019). Why most studies of individual differences with inhibition tasks are bound to fail. Doi: 10.31234/osf.io/3cjr5

Schiltenwolf, M., Kiesel, A., & Dignath, D. (2023). No temporal decay of cognitive control in the congruency sequence effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *49*(8), 1247–1263. https://doi.org/10.1037/xlm0001159

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612. http://dx.doi.org/10.1016/j.jrp.2013.05.009

Schuch, S., Philipp, A. M., Maulitz, L., & Koch, I. (2022). On the reliability of behavioral measures of cognitive control: retest reliability of task-inhibition effect, task-preparation effect, Stroop-like interference, and conflict adaptation effect. *Psychological Research*, *86*(7), 2158-2184. http://dx.doi.org/10.1007/s00426-021-01627-x

Schwarz, W., & Miller, J. (2012). Response time models of delta plots with negative-going slopes. *Psychonomic Bulletin & Review*, *19*(4), 555-574. http://dx.doi.org/10.3758/s13423-012-0254-6

Shilling, V., Chetwynd, A., & Rabbitt, P. M. (2002). Individual inconsistency across measures of inhibition: an investigation of the construct validity of inhibition in older adults. *Neuropsychologia*, *40*(6), 605–619. http://dx.doi.org/10.1016/S0028-3932(01)00157-9

Simon, J. R. (1990). The effects of an irrelevant directional cue on human information processing. In R. W. Proctor & T. G. Reeve (Eds.), *Stimulus-response compatibility: An integrated perspective* (pp. 31–86). North-Holland. http://dx.doi.org/10.1016/S0166-4115(08)61218-2

Snijders, T. A., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage.

Stahl, C., Voss, A., Schmitz, F., Nuszbaum, M., Tüscher, O., Lieb, K., & Klauer, K. C. (2014). Behavioral components of impulsivity. *Journal of Experimental Psychology: General*, *143*(2), 850–886. http://dx.doi.org/10.1037/a0033981

Stins, J. F., Polderman, J. T., Boomsma, D. I., & de Geus, E. J. (2007). Conditional accuracy in response interference tasks: Evidence from the Eriksen flanker task and the spatial conflict task. *Advances in Cognitive Psychology*, *3*(3), 409-417. http://dx.doi.org/10.2478/v10053-008-0005-4

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

Stuss, D. T., & Alexander, M. P. (2007). Is there a dysexecutive syndrome?. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 901-915. https://doi.org/10.1098/rstb.2007.2096

Ulrich, R., Schröter, H., Leuthold, H., & Birngruber, T. (2015). Automatic and controlled stimulus processing in conflict tasks: Superimposed diffusion processes and delta functions. *Cognitive Psychology*, *78*, 148–174. http://dx.doi.org/10.1016/j.cogpsych.2015.02.005

Unsworth, N., Miller, A. L., & Robison, M. K. (2021). Are individual differences in attention control related to working memory capacity? A latent variable mega-analysis. *Journal of Experimental Psychology: General*, *150*(7), 1332-1357. http://dx.doi.org/10.1037/xge0001000

Unsworth, N., Miller, A. L., & Strayer, D. L. (2024). Individual differences in attention control: A meta-analysis and re-analysis of latent variable studies. *Psychonomic Bulletin & Review*, 1-47. https://doi.org/10.3758/s13423-024-02516-1

Van Den Wildenberg, W. P., Wylie, S. A., Forstmann, B. U., Burle, B., Hasbroucq, T., & Ridderinkhof, K. R. (2010). To head or to heed? Beyond the surface of selective action inhibition: a review. *Frontiers in Human Neuroscience*, *4*, 222. http://dx.doi.org/10.3389/fnhum.2010.00222

von Bastian, C. C., Blais, C., Brewer, G., Gyurkovics, M., Hedge, C., Kałamała, P., ... & Wiemers, E. (2020). Advancing the understanding of individual differences in attentional control: Theoretical, methodological, and analytical considerations. PsyArXiv. http://10.31234/osf.io/x3b9k

von Bastian, C. C., & Druey, M. D. (2017). Shifting between mental sets: An individual differences approach to commonalities and differences of task switching components. *Journal of Experimental Psychology: General*, *146*(9), 1266–1285. https://doi.org/10.1037/xge0000333

von Bastian, C. C., Locher, A., & Ruflin, M. (2013). Tatool: A Java-based open-source programming framework for psychological studies. *Behavior Research Methods*, *45*(1), 108–115. http://dx.doi.org/10.3758/s13428-012-0224-y

von Bastian, C. C., Souza, A. S., & Gade, M. (2016). No evidence for bilingual cognitive advantages: A test of four hypotheses. *Journal of Experimental Psychology: General*, *145*(2), 246–258. http://dx.doi.org/10.1037/xge0000120

Waszak, F., Hommel, B., & Allport, A. (2003). Task-switching and long-term priming: Role of episodic stimulus–task bindings in task-shift costs. *Cognitive Psychology*, *46*(4), 361-413. https://doi.org/10.1016/S0010-0285(02)00520-0

Whitehead, P. S., Brewer, G. A., & Blais, C. (2019). Are cognitive control processes reliable? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(5), 765–778. http://dx.doi.org/10.1037/xlm0000632

Whitehead, P. S., Brewer, G. A., & Blais, C. (2020). Reliability and convergence of conflict effects: An examination of evidence for domain-general attentional control. *Experimental Psychology*, *67*(5), 303–313. http://dx.doi.org/10.1027/1618-3169/a000497

Winer, B. J., Brown, D. R., & Michels, K. M. (1971). *Statistical principles in experimental design*. New York: McGraw-Hill.

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, *12*(4), 399–413. http://dx.doi.org/10.1037/1082-989X.12.4.399