Journal of Experimental Psychology: Learning, Memory, and Cognition

Are Cognitive Control Processes Reliable?

Peter S. Whitehead, Gene A. Brewer, and Chris Blais

Online First Publication, July 26, 2018. http://dx.doi.org/10.1037/xlm0000632

CITATION

Whitehead, P. S., Brewer, G. A., & Blais, C. (2018, July 26). Are Cognitive Control Processes Reliable?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. http://dx.doi.org/10.1037/xlm0000632

Journal of Experimental Psychology:
Learning, Memory, and Cognition

0278-7393/18/$12.00

http://dx.doi.org/10.1037/xlm0000632

Are Cognitive Control Processes Reliable?

Peter S. Whitehead
Duke University

Gene A. Brewer and Chris Blais
Arizona State University

Recent work on cognitive control focuses on the *conflict-monitoring hypothesis*, which posits that a performance monitoring mechanism recruits regions in the dorsolateral prefrontal cortex to ensure that goal-directed behavior is optimal. Critical to this theory is that a single performance monitoring mechanism explains a large number of behavioral effects including the sequential congruency effect (SCE) and the error-related slowing (ERS) effect. This leads to the prediction that the size of these effects should correlate across cognitive control tasks. To this end, we conducted three large-scale individual differences experiments to examine whether the SCE and ERS effect are correlated across Simon, Flanker, and Stroop tasks. Across all experiments, the results revealed a correlation for the error-related slowing effect, but not for the sequential congruency effect across tasks. We discuss the implications of these results in regards to the hypothesis that a domain-general performance monitoring mechanism drives both effects.

Keywords: conflict-monitoring, sequential congruency effect, error-related slowing, individual differences, cognitive control

Over the last two decades, research into the underlying mechanisms of cognitive control has focused heavily on how top-down resources are engaged to avoid the homunculus problem (Baddeley & Della Sala, 1996; Cohen, Dunbar, & McClelland, 1990; Norman & Shallice, 1986). This effort culminated in the *conflict-monitoring* hypothesis (Botvinick, Braver, Barch, Carter, & Cohen, 2001). According to this view, response conflict is detected by the anterior cingulate cortex (ACC) and used to recruit executive functions residing in the dorsolateral prefrontal cortex (DLPFC) in proportion to their need. These executive functions (i.e., Miller & Cohen, 2001) act by biasing (see Desimone & Duncan, 1995) motor and/or behavioral responses to optimize goal-directed performance.

The conflict-monitoring hypothesis has been applied to dozens of literatures, and nearly all of the primary work is experimental. Specifically, researchers use conflict tasks such as Simon, Stroop, and Flanker interchangeably to investigate the mechanisms of cognitive control under the assumption that the control effects caused by these tasks are generated by the same mechanism. Recent work, supported by neural evidence, continues to accept the premise that conflict monitoring across these tasks may be mediated by the same or similar cognitive control process (Cavanagh & Frank, 2014; Feldman & Freitas, 2016; Nigbur, Ivanova, & Stürmer, 2011; Ridderinkhof, Ullsperger, Crone, & Nieuwenhuis, 2004; but see Töllner et al., 2017).

However, a growing body of research suggests that simple conflict effects from different executive function tasks do not exhibit convergent validity and may even have low internal reliability (Bender, Filmer, Garner, Naughtin, & Dux, 2016; Feldman & Freitas, 2016; Paap & Sawi, 2016; Ward, Roberts, & Phillips, 2001). Some researchers may find this claim unsurprising given the body of work investigating the different driving perceptual mechanisms of conflict effects in these three tasks (see Kornblum, Hasbroucq, & Osman, 1990). Yet, often within cognitive control literature there is a weak theoretical justification for the utilization of a specific control task to investigate a specific cognitive process. More specifically, cognitive control tasks are used interchangeably to assess models of cognitive control.

The following studies were therefore undertaken with the aim to probe the extent to which predictions of a common control mechanism made by the conflict monitoring hypothesis are observed in the between-task shared variance of behavioral effects, and to also highlight the critical need for a more theoretically guided experimental design in regards to task and paradigm choice within the cognitive control and associated literature. We focus on two classic behavioral cognitive control effects—the sequential congruency effect (SCE) and error-related slowing (ERS)—which are commonly conceptualized as markers of control system engagement (Egner, 2007; Gratton, Coles, & Donchin, 1992; Rabbitt, 1968).

Peter S. Whitehead, Center for Cognitive Neuroscience, Duke University; Gene A. Brewer and Chris Blais, Department of Psychology, Arizona State University.

A version of this article was accepted by Barrett Honors College at Arizona State University for partial fulfillment PSW's undergraduate honors thesis. PSW and CB designed and analyzed the data. PSW collected the data. All authors wrote the article. We'd like to thank the following research assistants for their time and effort in data collection: Victoria Jacoby, Grace Kennedy, Hayley Lambertus, Nowed Patwary, Candace Rizzi-Wise, and Cameron Robins. This work was partially supported by National Science Foundation Grant 1632291 awarded to GAB. All materials and data from these experiments can be downloaded from our laboratory website asumaclab.com.

Correspondence concerning this article should be addressed to Chris Blais, Department of Psychology, Arizona State University, 950 South McAllister Avenue, Tempe AZ, 85287-1104. E-mail: chris.blais@gmail.com

Markers of Cognitive Control

The sequential congruency effect (Gratton et al., 1992) refers to the observation that the size of the conflict effect is reduced following a high conflict trial. For instance, in a Stroop task, the size of the Stroop effect is smaller following an incongruent trial. Theoretically, this reduction occurs because additional resources are recruited to resolve response conflict on an incongruent trial, and this causes the system to more efficiently process the color and ignore the word on the subsequent trial (Botvinick et al., 2001). This account of the effect is so popular that the effect is often referred to as the conflict adaptation effect. This behavioral effect has been corroborated in BOLD activity, as well. Just as response conflict signals ACC activation and subsequent DLPFC activation for the current trial, ACC activation also predicts the activation of DLPFC in the proceeding trial (Kerns, 2006; Kerns et al., 2004). This is such that ACC activation for an incongruent trial not only signals DLPFC activation for the recruitment of executive functions in that trial, but a study by Kerns (2006) presents evidence that greater DLPFC activation is seen in the following trials if the previous trial was incongruent rather than congruent. Although it is generally accepted that conflict-adaptation plays at least some role in the SCE (see the Ullsperger, Bylsma, & Botvinick, 2015 response to Mayr, Awh, & Laurey, 2003 and Egner, 2014 for a review), one must be careful to consider other factors that contribute to the SCE. This single issue actually determined the implementation of each of the studies reported below.

The error-related slowing effect refers to the observation that response times are slower following an error (Rabbitt, 1968). The theoretical mechanism describing this effect has a similar consequence as the sequential congruency effect. The error-related slowing effect is argued to result from increased response caution (Botvinick et al., 2001). Much like the explanation of the sequential congruency effect, it is thought that the top-down regulation of control processes directed by the ACC causes an adaptive change in response threshold on the next trial. Thus, greater ACC activity on error trials has been associated with greater behavioral adjustment on the subsequent posterror trial (Kerns et al., 2004). This effect is thought to be a product of the cognitive control system and has been tied to fMRI BOLD activity in the ACC (Kerns et al., 2004) and an ERP component called the error-related negativity (ERN; Hajcak, McDonald, & Simons, 2003).

General Approach

By measuring performance on several cognitive control tasks in the same subject, we can assess the extent to which a common mechanism accounts for variance across participants and the extent to which these tasks should be interchangeably used. As noted, nearly all work on cognitive control is experimental. However, there is much theoretical traction to be gained by adopting an individual-differences approach (Underwood, 1975). If the same control process drives error-related slowing and sequential congruency effects then (a) subjects who have a large error-related slowing effect in one task should also have a large error-related slowing effect in a similar task, (b) subjects who have a large sequential congruency effect in one task should also have a large sequential congruency effect in another task, and (c) subjects who have a large sequential congruency effect in one task should also have a large error-related slowing effect in another task. In addition, the reliability of both of these measures should be relatively large.

In Experiment 1[1] we sought to test this hypothesis using the "purest" measure of conflict adaptation: the component of the SCE that is independent from other components that affect performance in the same direction (e.g., feature overlap; response contingencies). To preview the results, Experiment 1 yielded no between task correlations for the SCE, and perhaps more surprising, extremely low reliability within a task. We therefore examined whether the near-zero reliability of the conflict-adaptation component of SCE was the result of diluting the magnitude of the SCE by excluding components known to influence its magnitude which also "fall out" of the conflict monitoring framework by allowing these components to covary with the "pure" measure Specifically, Experiment 2 included feature repetitions (e.g., Mayr, Awh, & Laurey, 2003) and found a similarly low reliability of the SCE and a lack of correlation of that measure across tasks and Experiment 3 included both feature repetitions and contingencies between the target and distractor dimensions—a manipulation known to yield reliable first-order conflict effects (Borgmann, Risko, Stolz, & Besner, 2007; Engle & Kane, 2003). The results were still both extremely low reliability and no between task correlations for the SCE. However, in contrast to the results of the SCE, across all three experiments the ERS effect was robustly correlated between tasks. To our knowledge, this is the first high-powered individual differences study to assess the reliability and shared variance in performance on the three most common tasks used to study cognitive control: Simon, Flanker, and Stroop.

General Method

Given the massive impact the conflict monitoring hypothesis has had in shaping the literature, we wanted to have at least .80 power to detect a cross-task correlation in the SCE as small as |.225| (i.e., 5% of the variance; Hulley, Cummings, Browner, Grady, & Newman, 2013). G-Power (Faul, Erdfelder, Lang, & Buchner, 2007) indicated a sample size greater than 154 individuals was required. Our participants were recruited from the introductory psychology research pool at Arizona State University in exchange for course credit in accordance with the IRB. All participants were English speakers. Inclusion criteria were: >70% accuracy in each of the four cells contributing to the SCE for all three tasks. Correct responses greater than 3,000 ms (outlier) or less than 200 ms (anticipatory) were also excluded.[2] Based on similar experiments in our lab with this population of students, we expected replacing approximately 15% of our subjects for having <70% accuracy in any one cell. The number of valid (i.e., >70% accuracy in each cell for all three tasks) subjects in each experiment was monitored every 7–10 days and data collection stopped after the total number of valid subjects exceeded 154.

To determine the minimum number of trials needed to observe a reliable SCE effect, we first examined previously published data (Blais, Stefanidi, & Brewer, 2014; Blais et al., 2012) in which $N =$

[1] Data collected occurred in the opposite order as is reported here. That is, Experiment 3 was the first conducted and Experiment 1 was the last.

[2] Other trimming procedures (e.g. exceeding 2.5 *SD* or 3.5 *SD* per cell; outside of 200 ms–1,500 ms) yielded a qualitatively similar pattern of findings.

15 subjects performed $N = 19{,}000$ trials across 190 blocks of 100 trials in a vocal Stroop experiment over the span of roughly 5–7 days. An odd-even split-half reliability analysis across all trials (cell counts of approximately 3,000 trials in the cC and iI conditions and 1,500 in the cI and iC conditions) yields $r_{14} = .94$, $p < .001$. To determine the minimum number of trials necessary to observe adequate reliability while keeping the length of the study reasonable, we examined how each additional block of 100 trials (i.e., 50 trials per half divided across the four cells required to compute the SCE) improved reliability. For these data, the first 100 trials yielded split-half reliability of $r_{14} = .37$. This quickly increased and stabilized to around $r_{14} = .65$–$.75$ for between 400 to 1,500 total trials. It did not exceed 0.80 until we surpassed 2,000 total trials. We therefore settled on presenting 720 experimental trials that were preceded by 240 practice trials which should be sufficient to observe odd-even split-half reliability of around .70. An alternative approach to computing power is presented in the General Discussion but yields similar findings.

We chose to report the Spearman's rho correlation in all three experiments. Pearson correlations benchmark the linear relationship between two variables, while Spearman's rho benchmarks the monotonic relationship. While the conflict monitoring hypothesis assumes the magnitude of an individual's SCE and ERS in one task will be related to the magnitude those effects in another, it does not assume this relationship is linear. Therefore, our use of Spearman's rho will allow us to better characterize the shared variance between the SCE and ERS without assuming a linear relationship. That said, we also examined the Pearson coefficients, which showed the same qualitative pattern of results as the Spearman coefficients.

Across all three experiments, the Flanker task presented subjects with a string of five letters and asked to identify the middle one (D, F, J, or K), ignoring the flanking letters using those keys on a standard QWERTY keyboard. In the Simon[3] task, subjects were presented with a directional word (RIGHT, LEFT, UP, DOWN) appearing to the right, left, top, or bottom of a fixation cross and were instructed to respond the word, not its location, using the arrow keys. In the Stroop task, subjects were presented with color words RED, BLUE, GREEN, and YELLOW in those same colors. Subjects were instructed to respond to the color of the text. In the Stroop task, stimulus responses were randomly mapped to the d, f, j, and k keys, but responses remained fixed for the Simon and Flanker task. The stimulus remained onscreen until the subject responded. All instructions were presented on the computer screen prior to the start of the experiment and were clarified by the experimenter as needed.

A random inter-trial-interval (intertribal interval [ITI]) between 400 ms and 1,000 ms (in 200-ms intervals sampled from a univariate distribution with replacement; see Blais, 2008) separated trials. In Experiments 1 and 2 there were eight blocks of 120 trials per block in each task. The first two blocks were considered practice and not used in the analysis. In Experiment 3 there were six blocks of 120 trials per block in each task. The first block was considered practice and not used in the analysis. During the practice phase each of the response labels were presented on the screen in order corresponding to the response keys and feedback was provided in the form of a + or − symbol. This served as the fixation marker for the next trial. For the final six experimental blocks the response labels were removed from the screen and there was no feedback; an asterisk (*) was used as the fixation marker. All instructions were presented on the computer screen prior to the start of the experiment and were clarified by the experimenter as needed.

Experiment 1: "Pure" Conflict Adaptation

There is a detailed literature of episodic memory based influences on conflict measures (Akçay & Hazeltine, 2007; Mayr et al., 2003; Schmidt, 2013; Schmidt, Crump, Cheesman, & Besner, 2007; Schmidt & Weissman, 2014). One of these episodic memory components of control tasks is commonly referred to as the feature-repetition or feature-integration confound (Akçay & Hazeltine, 2007; Mayr et al., 2003; Schmidt & Weissman, 2014). Feature-repetition occurs when the target feature of the stimulus repeats between trials, when the target feature becomes the distractor feature in the next trial, and vice versa, thus causing the memory of the previous trial and its either direct or indirect repetition of feature components to bias the response to the current trial.

The other of these episodic memory influences on conflict measures is a color–word contingency (Schmidt, 2013; Schmidt et al., 2007; Schmidt & Weissman, 2014). This influence is such that an implicit relationship is learned between the frequency of a stimulus category (i.e., congruent vs. incongruent) and the frequency of the items that make up each stimulus category (i.e., all the possible combinations of the congruent stimulus category in a Stroop task). In a 50:50, congruent to incongruent, proportion Stroop task with four choices and four responses, for example, there are only four possible congruent items in this design (RED_{RED}, $\text{BLUE}_{\text{BLUE}}$, $\text{GREEN}_{\text{GREEN}}$, and $\text{YELLOW}_{\text{YELLOW}}$) but 12 possible incongruent items. Thus, each congruent item will be presented more often than each incongruent item. This implicit contingency learning influence contributes to stronger stimulus-response relationships in congruent trials, biasing responses such that, in the above example, each congruent item is faster as a result of being presented more often than each incongruent item (Melara & Algom, 2003).

These episodic memory influences create a purportedly impure measure of control engagement and much effort has been spent in order to circumvent them, with some studies showing they even drive behavioral effects such as the SCE (but see Blais et al., 2014; Mayr et al., 2003; Schmidt & De Houwer, 2011). In order to measure an SCE and ERS free of these influences biasing responses, we utilized the design of Schmidt and Weissman (2014) to create Simon, Flanker, and Stroop tasks without feature-repetition or contingency learning influences. This alternating presentation design splits a four-choice, four-response task into two groups (i.e., two-choice, two-response), such that stimuli from each of these groups are presented in alternating order and thus preventing any feature repetitions in initial task presentation. This design allowed us to assess the shared variance of effects and reliability of these tasks, accurately testing the predictions of a single mechanism control system using a "pure" measure of conflict adaptation, without decreasing our measurement precision by

[3] This version of the Simon task is best classified as a spatial variant of the Stroop task in that participants are instructed to respond to the meaning of the word.

excluding a large number of trials to remove feature repetitions as would be done in other paradigms.

Method

A total of 221 participants were recruited. The study required approximately 1 hr to complete. Subjects performed Stroop, Flanker, and Simon tasks in random order. Each of these tasks was a four-choice, four-response button task, presented in alternating order with a 50:50 proportion of incongruent to congruent stimuli, thus preventing any feature repetition or color-word contingency influences. This alternating order only allows for two sets of four items, rather than the entire set of 16 possible items (see Table 1).

Results

Thirty-five subjects were excluded from the analysis for failing to correctly complete one or more of the tasks (<70% accuracy in any condition; five in Simon, four in Flanker, 19 in Stroop, seven in both Simon and Flanker, four in both Simon and Stroop, two in both Stroop and Flanker, and two in all three tasks) leaving N = 178 in the final sample. Incorrect trials were excluded from reaction time (RT) analysis (5.7%, 6.0%, and 8.2% of trials in the Simon, Flanker, and Stroop tasks, respectively), and outliers in the remaining correct response times (<200 ms or >3,000 ms) were eliminated (0.5%, 0.3%, and 1.7% for Simon, Flanker, and Stroop tasks, respectively). To assess the presence an error-related slowing effect, we conducted a one-way repeated measures ANOVA comparing postincorrect trials to postcorrect trials. To assess the presence of a sequential congruency effect, we conducted a 2 (previous trial congruency) × 2 (current trial congruency) repeated measures ANOVA. Response times for each condition are in Table 2 and trial counts for each cell are in Table 3. For our purposes, there was an SCE (postcongruent vs. postincongruent conflict effect for each task: Simon: 72 ms vs. 63 ms; Flanker: 52 ms vs. 43 ms; Stroop: 84 ms vs. 70 ms) and ERS effect (Simon: 115 ms; Flanker: 163 ms; Stroop: 304 ms). The complete results of these analyses are in Table 4.

Table 1
An Example of the Item Combinations Displayed to Each Participant for Each Experiment, Using the Simon Task as an Illustration

	Position			
Word	⇑	⇓	⇐	⇒
Experiment 1				
UP	90	90		
DOWN	90	90		
LEFT			90	90
RIGHT			90	90
Experiment 2				
UP	45	45	45	45
DOWN	45	45	45	45
LEFT	45	45	45	45
RIGHT	45	45	45	45
Experiment 3				
UP	75	25	25	25
DOWN	25	75	25	25
LEFT	25	25	75	25
RIGHT	25	25	25	75

Reliability Within and Correlations Across Tasks

The most common approach for estimating the split-half reliability of a measure would be to compare the size of an effect in the first half versus the second half of the task. This approach leads to SCE reliability estimates of r_s = .02, .05, and −.02 (ns) for Stroop, Simon, and Flanker, respectively. However, because practice (among other things) differs in the first versus second half of the experiment, and more critically, the magnitude of the Stroop effect changes as a function of practice (e.g., Davidson, Zacks, & Williams, 2003; Pratte, Rouder, Morey, & Feng, 2010), the interpretation of these results is unclear. Therefore, the approach we adopted here was to compare odd trials to even trials. To determine the upper-bound for the correlations between tasks, we computed the split-half reliability for each of the sequential congruency and error-related slowing effects by numbering each trial odd or even, computing the effect separately for the odd and even trials, and then correlating these values across subjects (see Table 5). This has the advantage of keeping practice effects relatively constant. We corrected the split half correlation using the Spearman-Brown correction for split-half data (i.e., 2r/[1 + r]; Allen & Yen, 1979). These values are reported on the diagonal of Table 5. The lower diagonal reflects the standard Spearman's rho correlation across tasks. The upper diagonal of these tables reflects the reliability-corrected (i.e., r_{xy}/([r_{xx} r_{yy}]ˆ[1/2]; (Murphy & Davidshofer, 1988) Spearman's rho correlation. Figure 1 shows scatterplots for the between task correlation for error-related slowing and sequential congruency. Similar to what is found in neural evidence (Riesel, Weinberg, Endrass, Meyer, & Hajcak, 2013), the error-related slowing effect was correlated across tasks. Pearson's correlation coefficient was also calculated for all split-half and between task correlations, showing quantitatively similar results.

Each of the Spearman's rho values were Fisher-z transformed to allow comparisons between one another. For the SCE, all correlations were statistically equal to one another and no different from zero. For the ERS, all correlations were equal to each other, and significantly greater than zero. Finally, for each pair of tasks (e.g., the Stroop/Simon/Flanker correlation), the ERS was statistically greater than the SCE.

Discussion

At the group level, we replicate the standard observation showing an SCE and ERS effect in the Stroop, Simon, and Flanker tasks. Although the SCE is statistically reliable, its effect size is much smaller than what is observed in more traditional designs (e.g., Experiments 2 and 3), but similar to other reports using this design and others that control for feature repetitions and target-distractor contingencies (but see Blais et al., 2014; Schmidt & De Houwer, 2011). In contrast to the predictions of the conflict-monitoring hypothesis, however, while ERS is uniformly reliable and correlated across tasks, the SCE shows no significant between task correlations and low reliability that is only significant for the Flanker task (r_s = .17, $p < .05$). Furthermore, there were also no significant within-task, between effects correlations (i.e., the ERS effect in the Flanker task was not correlated to the SCE in the same

Table 2
Mean Response Times (ms) and Standard Deviations for Each Conditions of Each Effect, in All Three Experiments

	Simon		Flanker		Stroop	
Conflict effect	Congruent	Incongruent	Congruent	Incongruent	Congruent	Incongruent
Experiment 1	559 (91)	626 (95)	671 (141)	718 (149)	757 (140)	832 (157)
Experiment 2	580 (70)	647 (80)	767 (116)	845 (128)	698 (108)	730 (111)
Experiment 3	564 (73)	668 (77)	719 (124)	780 (135)	774 (124)	889 (139)
ERS	Postcorrect	Posterror	Postcorrect	Posterror	Postcorrect	Posterror
Experiment 1	586 (91)	700 (166)	688 (142)	851 (270)	777 (145)	1,082 (274)
Experiment 2	620 (77)	771 (137)	714 (109)	869 (180)	804 (121)	1,090 (229)
Experiment 3	605 (73)	776 (172)	743 (127)	905 (210)	816 (127)	1,103 (233)
SCE	Previous Congruent	Previous Incongruent	Previous Congruent	Previous Incongruent	Previous Congruent	Previous Incongruent
Experiment 1						
Congruent	550 (89)	567 (92)	666 (139)	674 (139)	740 (136)	772 (144)
Incongruent	622 (94)	630 (98)	718 (152)	716 (146)	824 (159)	842 (158)
Experiment 2						
Congruent	557 (74)	587 (70)	682 (107)	701 (110)	745 (122)	770 (119)
Incongruent	646 (80)	645 (81)	729 (114)	728 (111)	846 (132)	840 (127)
Experiment 3						
Congruent	544 (74)	582 (74)	706 (123)	732 (128)	755 (125)	792 (128)
Incongruent	668 (78)	665 (78)	782 (136)	780 (134)	886 (139)	891 (143)

Note. SCE = Sequential Congruency Effect; ERS = Error Related Slowing.

Flanker task; Table 6). It is important to remember that even though the Flanker SCE showed a small but significant reliability, this is only after applying the Spearman-Brown correction which is known to be problematic with low reliability (Lumsden, 1976; Spearman, 1904).

The fact that we failed to observe a reliable SCE was quite surprising particularly given how robust the effect is at the group level. This lack of reliability also appears inconsistent with the conflict monitoring hypothesis, and thus Experiments 2 and 3 attempted to increase reliability including components (e.g., feature repetitions) known to increase the size of the effect that are consistent with, but do not provide unique support for, the conflict monitoring hypothesis.

The first of these are the episodic memory influences that are now known to contribute to the SCE—these influences were included in Experiments 2 and 3, but excluded from Experiment 1. Specifically, feature repetitions are one of the most widely known and addressed influences on the SCE within cognitive control (Akçay & Hazeltine, 2007; Mayr et al., 2003; Schmidt & Weissman, 2014). In most cases, removing such features decreases the magnitude of the SCE (but see Blais et al., 2014). It is unclear whether the reduction in the magnitude of the SCE with this design change is also accompanied by a decrease in task reliability. As these feature repetition contributions would add signal to the noise of the SCE, their exclusion may have decreased the reliability of the SCE and its between task correlations.

Table 3
Mean Cell Counts (Trials) and Standard Deviations for Each Conditions of Each Effect, in All Three Experiments

	Simon		Flanker		Stroop	
ERS	Postcorrect	Posterror	Postcorrect	Posterror	Postcorrect	Posterror
Experiment 1	681 (58)	39 (23)	683 (65)	35 (22)	646 (74)	48 (27)
Experiment 2	670 (59)	48 (24)	667 (71)	46 (24)	635 (69)	55 (26)
Experiment 3	566 (17)	29 (17)	572 (20)	28 (19)	567 (19)	33 (19)
SCE	Previous Congruent	Previous Incongruent	Previous Congruent	Previous Incongruent	Previous Congruent	Previous Incongruent
Experiment 1						
Congruent	186 (9)	188 (9)	180 (12)	183 (12)	175 (13)	177 (15)
Incongruent	170 (16)	176 (15)	177 (16)	179 (15)	170 (16)	172 (16)
Experiment 2						
Congruent	45 (6)	141 (10)	43 (5)	136 (12)	43 (6)	131 (10)
Incongruent	130 (12)	396 (27)	134 (13)	396 (33)	127 (11)	385 (27)
Experiment 3						
Congruent	147 (7)	151 (7)	147 (6)	153 (6)	147 (7)	153 (7)
Incongruent	139 (10)	158 (10)	149 (6)	151 (6)	146 (8)	154 (8)

Note. SCE = Sequential Congruency Effect; ERS = Error Related Slowing.

Table 4
Results for Each One-Way Repeated Measures ANOVA to Confirm the Error-Related Slowing Effect and the 2 × 2 Repeated Measures ANOVA to Confirm the Sequential-Congruency Effect Conducted in Each Experiment, With Response Times as the Dependent Variable

	Simon				Flanker				Stroop			
ANOVA factor	*F*	*df*	p	η_p^2	*F*	*df*	*p*	η_p^2	*F*	*df*	*p*	η_p^2
Experiment 1												
Error-related slowing	152.6	1,177	<.001	.46	132.7	1,177	<.001	.43	365.6	1,177	<.001	.67
Congruency	943.1	1,177	<.001	.84	374.7	1,177	<.001	.68	372.4	1,177	<.001	.68
Previous congruency	96.6	1,177	<.001	.35	3.0	1,177	.085	.02	92.7	1,177	<.001	.34
Interaction (i.e., SCE)	14.3	1,177	<.001	.08	9.1	1,177	.003	.05	10.1	1,177	.002	.05
Experiment 2												
Error-related slowing	397.6	1,194	<.001	.67	335.0	1,194	<.001	.63	611.6	1,194	<.001	.76
Congruency	1259.8	1,194	<.001	.87	340.7	1,194	<.001	.64	421.1	1,194	<.001	.69
Previous congruency	93.8	1,194	<.001	.33	21.4	1,194	<.001	.10	13.1	1,194	<.001	.06
Interaction (i.e., SCE)	139.4	1,194	<.001	.42	32.5	1,194	<.001	.14	29.6	1,194	<.001	.13
Experiment 3												
Error-related slowing	612.8	1,209	<.001	.75	278.9	1,209	<.001	.57	331.8	1,209	<.001	.64
Congruency	1116.9	1,209	<.001	.84	825.9	1,209	<.001	.80	2660.8	1,209	<.001	.93
Previous congruency	63.6	1,209	<.001	.23	40.0	1,209	<.001	.16	182.5	1,209	<.001	.47
Interaction (i.e., SCE)	38.3	1,209	<.001	.16	68.4	1,209	<.001	.25	312.0	1,209	<.001	.60

Note. SCE = Sequential Congruency Effect.

Experiment 2: Conflict Adaptation and Feature Repetitions

To address the possibility that reliability and between task correlations of the SCE are driven by the inclusion of feature repetitions in cognitive control tasks, this experimental paradigm specifically allowed for the inclusion of this episodic memory component in the presentation of the Simon, Flanker, and Stroop tasks. Importantly, the design of these three tasks did not include the target-distractor contingency that has been argued to influence the magnitude of the SCE effect (Schmidt & de Houwer, 2011) by using a 25:75 ratio of congruent to incongruent trials for each of the three tasks, such that each incongruent and congruent item was presented an equal amount of times. While addressing the same fundamental question as the last experiment, conceptually, this design allows us to test whether the combination of "pure" conflict adaptation and feature repetitions will (a) increase and the reliability of the SCE for each task and (b) increase the correlation of the SCE across tasks.

Table 5
Spearman's Rho Correlations of Response Time (ms) Data for the Sequential Congruency Effect and Error Related Slowing Effect for Each Experiment

	Error-related slowing			Sequential-congruency effect		
Task	Simon	Flanker	Stroop	Simon	Flanker	Stroop
Experiment 1						
Simon	.75**	.64**	.52**	.11	.34**	−1.12
Flanker	.46**	.69**	.40**	.05	.17*	.07
Stroop	.42**	.30**	.84**	−.10	.01	−.07
Experiment 2						
Simon	.67**	.49**	.68**	−.01	−.45**	−5.42
Flanker	.33**	.69**	.55**	−.01	−.06	3.60
Stroop	.50**	.41**	.81**	−.08	.11	.02
Experiment 3						
Simon	.64**	.87**	.45**	.03	−.20*	.44**
Flanker	.48**	.47**	.66**	−.01	.05	.59**
Stroop	.29**	.36**	.64**	.01	.02	−.03

Note. The middle diagonal is split-half reliability in the tasks, corrected using the Spearman-Brown correction for split-half reliability ((2*R)/(1 + R)) the lower diagonals are the correlation between the tasks, and the top diagonal is the correlation between the tasks corrected for low reliability, using the correction for attenuation formula (R_{Xy} / ([R_{Xx} R_{Yy}]^½). Significance of ($p < .01$) indicated by ** and of ($p < .05$) of * next to rho coefficient.

Method

A total of 222 individuals were recruited. The study required approximately 1 hr to complete. Subjects performed Stroop, Flanker, and Simon tasks in a pseudorandomized order. The tasks were identical to those in Experiment 1 except for the following changes. Incongruent and congruent stimuli were presented in a 25:75 proportion, where the proportion of congruent trials is 25% and incongruent 75%.

Results

Twenty-seven subjects were excluded from the analysis for failing to correctly complete one or more of the tasks (<70% accuracy in any condition; four in Simon, one in Flanker, 13 in Stroop, two in Stroop and Flanker, five in Simon and Stroop, and two in all three tasks) leaving $N = 195$ in the final sample. Incorrect trials were excluded from RT analysis (6.7%, 6.3%, and 8.6% of trials in the Simon, Flanker, and Stroop tasks, respectively), and outliers in the remaining correct response times (<200 ms or >3,000 ms) were eliminated (0.3%, 0.4%, and 1.7% for Simon, Flanker, and Stroop tasks, respectively). To assess the presence of an ERS effect, we conducted a one-way repeated measures ANOVA comparing postincorrect trials to postcorrect trials. To assess the presence of the SCE, we conducted a 2 (previous trial congruency) × 2 (current trial congruency) repeated measures ANOVA. For our purposes, there was an ERS

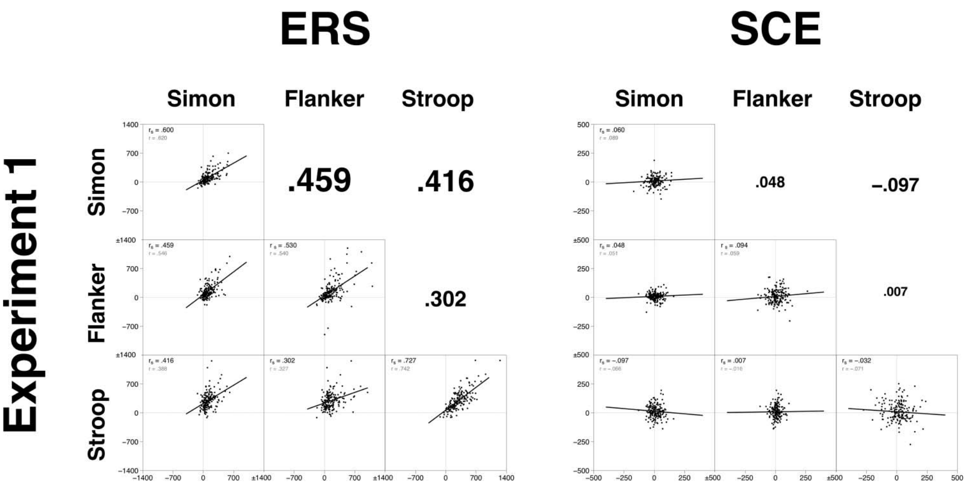

Figure 1. Correlations for error-related slowing (ERS) effect and sequential-congruency effect (SCE) in milliseconds (ms) for Experiment 1. The graphs on the diagonal are the split-half reliabilities for each effect, on the bottom diagonal are the between-task correlations, and on the upper diagonal are the corresponding between task reliabilities for the size for the correlation. The line in each graph is the line of best fit.

(Simon: 151 ms; Flanker: 155 ms; Stroop: 286 ms) and SCE (postcongruent vs. postincongruent conflict effect for each task: Simon: 89 ms vs. 58 ms; Flanker: 47 ms vs. 27 ms; Stroop: 101 ms vs. 70 ms). Full response times for each condition are in Table 2, trial counts for each cell are in Table 3, and results of these analyses are in Table 4.

Split-half reliability and between task correlations were calculated the same as in Experiment 1. Figure 2 shows scatterplots for the split-half reliability and between task correlations for the ERS and SCE effects. The ERS was the only effect that was significantly correlated between tasks, as well as internally reliable. Again, each of the Spearman's rho values was transformed into a Fisher-*z* score and showed the same pattern of results as in Experiment 1. A Pearson's correlation was also calculated for all split-half and between task correlations, and the qualitative pattern of results remained.

Discussion

The results of this experiment mirrored those of Experiment 1. We saw robust and significant group level effects ($p < .001$ for all effects, $n_p^2 > .60$ for the ERS in all tasks, and $n_p^2 = .42$, .14, and .13 for the SCE in the Simon, Flanker, and Stroop tasks, respectively). Again, however, the split-half reliabilities and between task correlations painted a different picture. Only the ERS effect was reliable, via the split-half reliability measure, across all three tasks (all $r_s > .65$). It was also the only effect that was correlated across all three tasks which is consistent with the interpretation that the mechanism driving the ERS is common across the three tasks. Conversely, the SCE showed the same pattern as demonstrated in Experiment 1; both a lack of reliability and the absence of any shared variance across tasks. Unlike in Experiment 1, we did see a small correlation between the ERS and the SCE of the Flanker task ($r = .142$; see Table 6). However, the between effects correlations for the Simon and Stroop task were nonsignificant. Thus, when both "pure" conflict adaptation and episodic memory contributions in the form of feature repetitions are allowed to contribute to the SCE, we still fail to observe any common variance in this measures across our three tasks.

Table 6
Spearman's Rho Correlations of Response Time (ms) Data Between the Sequential Congruency Effect (SCE) and Error Related Slowing (ERS) Effect for Each Experiment

	ERS		
SCE	Simon	Flanker	Stroop
Experiment 1	.07	.03	−.12
Experiment 2	.07	.14*	.08
Experiment 3	.01	−.06	−.10

Note. Significance of ($p < .05$) of * next to rho coefficient.

Perhaps one reason the SCE was not reliable within a task has to do with the fact that the relative number of trials used to estimate each cell comprising the interaction score differed drastically from Experiment 1. Assuming perfect accuracy, each of the 2 × 2 cells contained 180 trials in Experiment 1. However, because the proportion of congruent trials decreased from 50% to 25% in Experiment 2, the cell counts were approximately 38 in the congruent followed by congruent (cC) cell, 338 in the incongruent followed by incongruent (iI) cell, and 122 in each of the other two cells (cI and iC). Thus, while the SCE may actually be (slightly) more reliable when feature repetitions are included, we may have failed to observe this because noisy estimate in the cC cell. It is actually quite easy to quantify exactly how much of a decrease in the precision there was. Specifically, the observed *SD* for an individual subject's SCE in E1 was around 380 ms, making the SEM = $380/180^{0.5}$ = 28 ms. In Experiment 2, the overall *SD* was similar. This imbalance to the cell count yields an expected SEM around $((380^2/405 + 380^2/135 + 380^2/135 + 380^2/45)/4)^{0.5} = 38$ ms, or roughly a 33% increase in variance. We return to this issue more formally in the General Discussion, but it turns out to have a relatively minor impact on the minimum correlation we have the

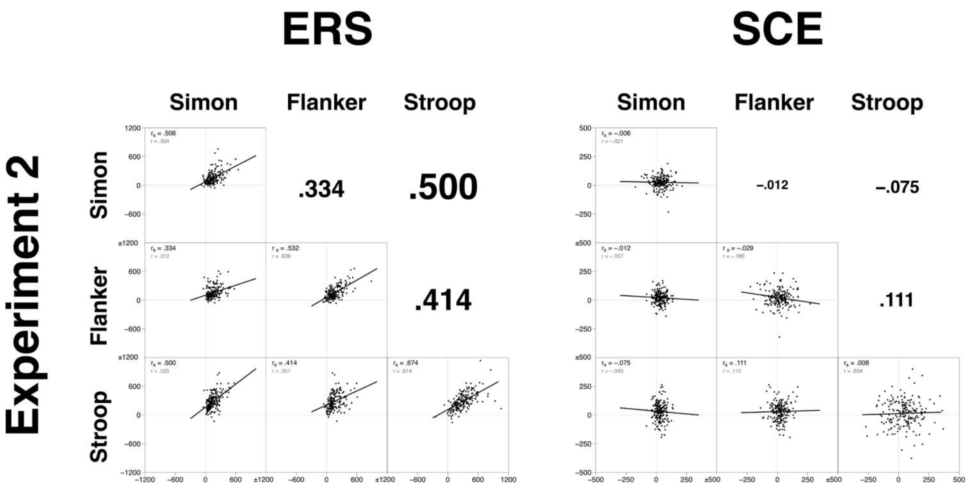

Figure 2. Correlations for error-related slowing (ERS) effect and sequential-congruency effect (SCE) in milliseconds (ms) for Experiment 2. The graphs on the diagonal are the split-half reliabilities for each effect, on the bottom diagonal are the between-task correlations, and on the upper diagonal are the corresponding between task reliabilities for the size for the correlation. The line in each graph is the line of best fit.

power to detect (i.e., instead of a minimum of around $r = .225$, the drop in precision means we need an effect closer to $r = .275$). That said, the psychometric value of a test–retest reliability below around $r = .70$—which we have $>.99$ power to detect—is quite limited.

Experiment 3: Conflict Adaptation, Feature Repetitions, and a Color–Word Contingency

The previous two experiments show that the ERS effect is both reliable for all tasks and the rank-order correlations across tasks have a high degree of shared variance. Neither of these outcomes holds for the SCE. Rather, the SCE is consistently unreliable with little shared variance across tasks, even under more favorable conditions such as when episodic memory contributions such as feature repetitions contribute to our estimate of the SCE. This pattern of results is inconsistent with the conflict monitoring hypothesis which states that a common mechanism—namely a conflict monitor—drives both the SCE and ERS effect in these (and other similar) tasks.

As noted, many earlier studies of cognitive control failed to account for episodic memory contributions to the SCE. While accounting for feature repetitions Mayr, Awh, and Laurey (2003) has become ubiquitous, other memory contributions such as target-distractor contingencies (e.g., Melara & Algom, 2003) can contribute to the magnitude of the SCE in some cases (Schmidt et al., 2011 but see Blais et al., 2014). This target-distractor contingency takes the form that any specific congruent item might be responded to faster simply because it has been seen more often than each of the incongruent items (Logan, 1988, 1990). Because there are only four possible congruent items in this design (i.e., e.g., in the Stroop task RED_{RED}, $\text{BLUE}_{\text{BLUE}}$, $\text{GREEN}_{\text{GREEN}}$, $\text{YELLOW}_{\text{YELLOW}}$), but 12 possible incongruent items, as the total proportion of incongruent to congruent items is 50:50, then each congruent item needs to be presented more often than each incongruent item to maintain a 50:50 congruent to incongruent ratio.

One benefit of increasing the proportion of congruent to incongruent is that the reliability of the conflict effect increases (Borgmann et al., 2007). Thus, increasing the proportion of congruent trials may increase the reliability of the SCE. While the working memory capacity literature tends to use between 70% and 80% congruent trials when using the Stroop task as an index of attention control (Engle & Kane, 2003) we decided to use a 50:50 congruency ratio so that the four cell counts that comprise the SCE interaction have an equal number of trials. Experiment 3 therefore examines a composite SCE that is comprised of "pure" conflict adaptation, episodic memory contributions in the form of feature repetitions, and a moderate contingency between the target-distractor features.

Method

A total of 215 individuals were recruited. The study required approximately 1 hr to complete and were the first tasks run within a battery of studies that required 2.5 hr to complete. Subjects performed Stroop, Flanker, and Simon tasks in that fixed order. Each of these was a four-alternative forced choice button-press task that allowed us to remove feature repetitions, and therefore more directly test the conflict adaptation aspect of the conflict monitoring hypothesis therefore improving on our ability to draw inferences from our results. Incongruent and congruent stimuli were presented in an equal, 50/50 proportion. The tasks were implemented to repeat all incorrect and slow (RT >3,000 ms) trials at the end of a block recursively until participants responded correctly and fast enough. The repeated trials were not included in the analyses reported, but this did not impact the pattern of data.

Results

Five subjects were excluded from the analysis for failing to correctly complete one or more of the tasks (<70% accuracy in any condition; four in Simon and one in Stroop) leaving $N = 210$ in the final sample. Incorrect trials were excluded from RT analysis (5.2%, 5.1%, and 6.3% of trials in the Simon, Flanker, and Stroop tasks, respectively), and outliers in the remaining correct response times (<200 ms or >3,000 ms) were eliminated (0.2%, 0.4%, and 1.4% for Simon, Flanker, and Stroop tasks, respectively).

To confirm the presence of an error-related slowing effect, we conducted a one-way repeated measures ANOVA comparing postincorrect trials to postcorrect trials. To confirm the presence of a sequential congruency effect, we conducted a 2 (previous trial congruency) × 2 (current trial congruency) repeated measures ANOVA. For our purposes, there was an SCE (postcongruent vs. postincongruent conflict effect for each task: Simon: 124 ms vs. 83 ms; Flanker: 76 ms vs. 48 ms; Stroop: 131 ms vs. 99 ms) and ERS (Simon: 171 ms; Flanker: 162 ms; Stroop: 287 ms). Full response times for each condition are in Table 2, trial counts for each cell are in Table 3, and results of these analyses are in Table 4.

The method for determining split-half reliability and between task correlations was identical to Experiment 1 and produced a similar qualitative pattern of results (see Table 5). This method was used on both data that had feature repetitions included and with feature repetitions removed (see Table 5). Figure 2 shows scatterplots for the split-half reliability and between task correlations for the ERS and SCE effects. The ERS was the only effect that was significantly correlated between tasks, as well as internally reliable. There was no correlation between the SCE and ERS effect in any task. As in the previous two experiments, each of the Spearman's rho values was transformed into a Fisher-*z* score and showed the same pattern of results as in Experiment 1. Again, a Pearson correlation was also calculated for all split-half and between task correlations, showing the qualitatively same pattern of results.

Discussion

As in the previous experiments, the overarching goal of Experiment 3 was to test the shared variance across tasks for the SCE and ERS effects. Experiment 3 is the most typical implementation of an experiment examining the SCE in that is uses four response alternatives (so that feature repetitions can be excluded from the analysis) and has a 50:50 congruency ratio. Even so, the results of this experiment mirror those from the previous two. Both the SCE and ERS effects are highly robust at the group level ($p < .001$). The corrected split-half reliability for the ERS effect is high (*r*s > .46) and there is substantial shared variance across tasks (*r*s > .28). Neither of these apply to the SCE. As in Experiments 1 and 2 the SCE has very low reliability (statistically zero) and therefore unsurprisingly no shared variance across tasks. Building a target-distractor contingency into our experiment failed to increase the reliability of the SCE. Again, we failed to see significant between effects correlations for each task in this experiment, replicating the general conclusions of Experiments 1 and 2. Overall, the results of Experiment 3 are consistent with those in Experiments 1 and 2 and together raise questions regarding the assumption that both the SCE and ERS are generated by a common mechanism.

General Discussion

The primary goal of this article was to investigate the extent to which the predictions made by the conflict monitoring hypothesis of a common control mechanism are observed in the between-task shared variance of behavioral effects—the SCE (Gratton et al., 1992) and ERS effect (Rabbitt, 1968)—commonly conceptualized as markers of control system engagement. Across three experiments, despite yielding significant SCE and ERS effects at the group level (see Table 4), our individual differences approach showed that only the error-related slowing effect was consistently correlated across tasks (Table 5; Figures 1–3). Additionally, across these experiments the ERS effect was uniformly reliable, which

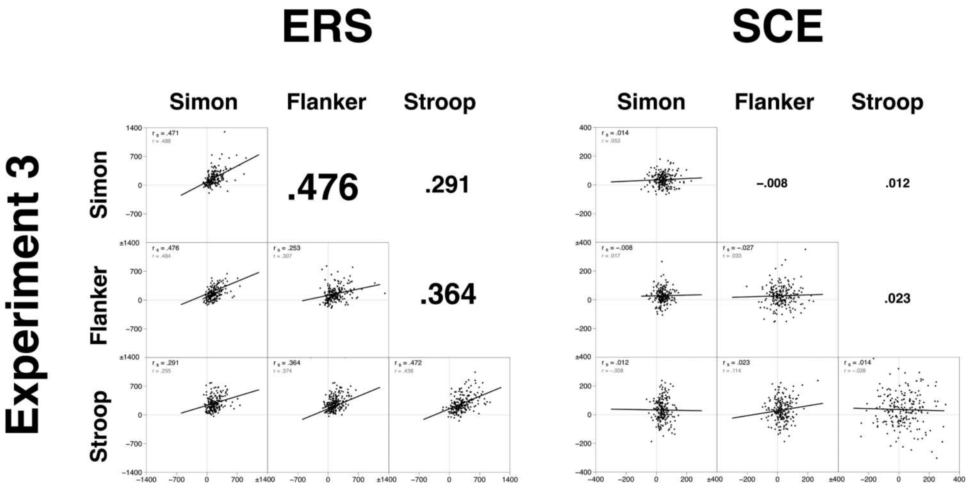

Figure 3. Correlations for error-related slowing (ERS) effect and sequential-congruency effect (SCE) in milliseconds (ms) for Experiment 3. The graphs on the diagonal are the split-half reliabilities for each effect, on the bottom diagonal are the between-task correlations, and on the upper diagonal are the corresponding between task reliabilities for the size for the correlation. The line in each graph is the line of best fit.

was contrasted by the unreliability of the SCE. Finally, across these three experiments we did not observe a pattern of shared variance between effects in the same task, either. The inclusion or exclusion of episodic memory influences did not change the qualitative pattern of results between experiments, and further demonstrated the lack of between tasks shared variance and reliability for the SCE in under the most favorable experimental conditions.

Experiment 1 used a four-response, four-feature task modeled after Schmidt and Weissman (2014) that did not allow feature repetitions, and the use of a 50:50 congruent-to-incongruent ratio prevented differential target-distractor contingencies from contributing to our measures of the SCE and ERS effects. Despite robust group level effects for the SCE and ERS, only the ERS showed shared variance across our three tasks. The SCE failed to demonstrate shared variance across the tasks, but this could have been the result of very low reliability of the SCE.

Experiments 2 and 3 sought to improve the reliability of the SCE by including feature repetitions, and feature repetitions and a contingency between the target and distractor features, respectively. While this certainly increased the size of the SCE effect in magnitude (going from an average of 11 ms in Experiment 1, to 27 ms and 34 ms in Experiments 2 and 3, respectively), it did not increase the reliability of the SCE and thereby the shared variance of this component across tasks. Nearly all updates or deviations from the conflict monitoring hypothesis (Blais, Robidoux, Risko, & Besner, 2007; Blais & Verguts, 2012; Schmidt, 2013) argue that the system learns which stimulus and response dimensions co-occur regularly in the tasks (see Melara & Algom, 2003). These newly bound stimuli then influence subsequent trials that have an overlapping feature. Because the frequency with which these features overlap changes as a function of condition (Mayr et al., 2003), they contribute to the size of the sequential congruency effect. So, conflict adaptation, feature integration, and feature/response repetition processes contribute to the sequential congruency effect, but the relative contribution may vary as a function of the task characteristics. Additionally, it should be noted that in Experiments 2 and 3, feature repetitions could be removed from the data (resulting in a loss of around 50% of trials). This is accomplished by removing four trial types: (1) the trial$_{n-1}$ target is the trial$_n$ target, (2) the trial$_{n-1}$ target is the trial$_n$ distractor, (3) the trial$_{n-1}$ distractor is the trial$_n$ target, and (4) the trial$_{n-1}$ distractor is the trial$_n$ distractor. Unsurprisingly, doing this in Experiments 2 and 3 yields the same qualitative pattern reported above. The failure to observe a correlation across Experiments 2 and 3 for the sequential congruency effect demonstrated the inability of these nonconflict adaptation processes to contribute to increased reliability and between-task correlations even though previous research demonstrated their effect of group level analysis.

The shared variance and high reliability across all three tasks for the ERS effect is consistent with prior literature. Dutilh et al. (2012) notes that virtually all theories of error-related slowing theorize a common underlying mechanism driving such an effect. Though differences between tasks may elicit an ERS effect that is influenced by that task's specific qualities, the shared variance between tasks in our data supports the conclusion that the underlying mechanism contributing to ERS is common between differing tasks (Dutilh et al., 2012). Additionally, the high reliability of the ERS in our data supports previous findings that the magnitude of the ERS is reliable across time (Danielmeier & Ullsperger, 2011), and supports finding in the neural literature that show that show similar posterror activity is highly reliable (Riesel et al., 2013). In sum, the magnitude of the ERS effect between individuals seems to be robust to a task's specific demands.

However, taken together, the results of the SCE from our current set of experiments and their individual differences results—specifically the lack of a correlation in the magnitude of the SCE across tasks—at best raises issues about the validity of interchangeably using these tasks to assess conflict adaptation (a common practice in the cognitive control literature e.g., Nieuwenhuis et al., 2006; Notebaert, Gevers, Verbruggen, & Liefooghe, 2006; Puccioni & Vallesi, 2012; Torres-Quesada, Funes, & Lupiáñez, 2013), and at worst questions the notion of a common performance monitoring mechanism (e.g., Botvinick, Cohen, & Carter, 2004). Recall that this model posits that the SCE occurs as the ACC detects response conflict and recruits executive functions in the DLPFC to alleviate the conflict. That we find no shared variance in the SCE across tasks using the most common ways of measuring it (pure, pure + feature reps, pure + feature reps + contingency) should cause concern.

Experimental approaches have also drawn similar conclusions about the generalizability of conflict adaptation as proposed by the *conflict-monitoring hypothesis* (Botvinick et al., 2001). For example, Funes, Lupiáñez, and Humphreys (2010) use a mixed-task design in which multiple types of conflict stimuli were presented in alternating fashion—presenting Stroop and Flanker stimuli within the same task, alternating between the two types of stimuli. In finding that conflict adaptation effects are not transferred from one type of conflict to the other, such that sequential effects were not seen if the stimulus at *n* was a Stroop stimulus but the stimulus at *n-1* was a Flanker stimulus, Funes et al. (2010) reinforces the specificity of conflict effects to their respective tasks. That is, each control task could produce a control signal specific to the task's bottom-up characteristics, which would be consistent with a lack of correlation for the SCE between different tasks in the current study.

Our results do not dispute these conclusions, and importantly, the evidence from these studies altogether suggests a solid theoretical ground to doubt the generalizability and reliability of conflict effects. However, the variability in findings between them stresses the importance of examining the relationship between low-level control effects, and their reliability, using individual differences analyses. We add to this literature using experimentally valid task designs with the ability to measure "pure," control effects, as well as conflict effects which are influenced by the episodic memory processes engaged by feature repetitions and color–word contingencies, thus leading us to draw stronger claims on the reliability and validity of single-monitor theories of cognitive control.

The Reliability of Cognitive Tasks

Recent work has also highlighted issues regarding reliability in cognitive control tasks. For example, Hedge, Powell, and Sumner (2017) tested the reliability of the basic conflict effect across seven tasks (which included versions of Stroop and Flanker tasks), finding the reliability of all tasks to be between .36 and .77. Note that these are below .80 which is the clinical standard for a measure to be considered excellent. They highlight the low, within sample variance as a possible cause for their findings, noting, however, that this may be inherent to these tasks. Many tasks within the attention and cognitive control literature have robust group means and effects (a trait for which they are particularly used for, in some cases), which necessitates a low variance between participants. Without sufficiently high variance between participants, it is difficult to correlate effects. Recent efforts have also been made to investigate the efficacy of error-related measures from these tasks for use in clinically applied, individual differences research, though (see Larson et al., 2016; Maurer et al., 2016). It should be noted that our specific question differed from Hedge et al. (2017), and while their conclusions may be applicable to our own work, we submit further that perhaps the number of trials in each of our experiments were not sufficient to reliably measure control system engagement via the SCE, although the length of our tasks were consistent with others in the literature, if not longer. It is also possible that implementation of control is more variable than originally accounted for, contributing to the observed unreliability in behavioral estimates of control system engagement.

Our set of experiments expands on a growing body of recent work investigating the reliability and cross-task variance for several common cognitive tasks (Bender et al., 2016; Feldman & Freitas, 2016; Keye, Wilhelm, Oberauer, & Stürmer, 2013; Keye, Wilhelm, Oberauer, & van Ravenzwaaij, 2009; Paap & Sawi, 2016; Ward et al., 2001). For example, Feldman and Freitas (2016) examined whether performance on a Wisconsin Card Sorting Task (WCST) and a Stroop-Trajectory task were correlated. Using a between task variables correlational approach they found that sequential congruency effects in the Stroop task were not significantly correlated to any WCST variables. These findings seem to be in line with our own, thus consistent with our suggestion that a task-general, conflict adaptation mechanism may not produce conflict effects across different control tasks in a similar fashion. In contrast to our approach and results, however, they used a test–retest approach to show the Stroop sequential congruency effect was significantly correlated between sessions, drawing a positive conclusion in support of the reliability of this measure. This could be attributed to the design of their Stroop task, in which the distractors of the Stroop stimuli are presented before the target, potentially contributing to a different sequence of cognitive processing in the production of the SCE. Additionally, the underlying neural mechanisms of each task posit a more varied engagement of different brain regions involved in the production of behavioral effects from the WCST (Lie, Specht, Marshall, & Fink, 2006) than in tasks such as the Stroop (Carter & van Veen, 2007; MacDonald, Cohen, Stenger, & Carter, 2000), which could contribute to their demonstrated behavioral similarity between these tasks.

Other previous studies have utilized structural equation modeling to show a lack of shared variance between tasks for the sequential congruency effect. Keye, Oberauer, and van Ravenzwaaij (2009) administered a two-choice, two-response vertical-Simon and Flanker task, and in a structural equation modeling approach found only a moderate relationship that accounted for a low proportion of shared variance between the Simon and Flanker task sequential congruency effects. In Keye, Wilhelm, Oberauer, and Stürmer (2013), using a two-choice, two-response vertical- and horizontal-Simon task they find a similar result in a confirmatory factor analysis, demonstrating a weak relationship between the tasks' sequential congruency effects. Both of these studies reach a similar view as presented here (Keye et al., 2013, 2009), challenging the existence of a single performance monitor as conceived by the *conflict monitoring hypothesis*. It should be noted that through their structural equation modeling they determine that feature repetitions inherent in their design (i.e., episodic memory influences on cognitive control) were not significant factors in conflict adaptation effects and do not impede their analysis.

The Lack of a Reliable SCE

A striking observation in our data was the fact that the split-half reliability for the SCE in each task was quite simply, nonexistent. This held even after correcting for the halving of data sets using the Spearman-Brown correction (Allen & Yen, 1979). It should be noted that using the Spearman-Brown correction to estimate the reliability of split-half data essentially predicts a doubling of the amount of data used, and thus should be interpreted accordingly. Furthermore, using the correction for attenuation to adjust the between task correlations for the low reliability in their respective tasks is not without its limitations. First, as the reader will note from the correlation tables (see Table 5), dividing near-zero numbers by other near-zero numbers can produce wildly inaccurate "corrected" between task correlations (i.e., $r > 1.00$), a fact that was acknowledged in the original work (Spearman, 1904). Thus, in using the correction for attenuation formula, the actual correlation between tasks is likely overestimated (Lumsden, 1976). Given this issue, the uncorrected between task correlations may more accurately reflect the degree of correlation across these tasks. Murphy and Davidshofer (1988) note that while the correction for attenuation is theoretically justified, there are practical grounds to criticize it as it makes an impossible assumption—that a task can exist without measurement error—an assumption all researchers know to be false.

A different method for computing power is by simulation. There are two sources of variance that will impact our ability to observe a relation: the interindividual variance and the intersubject variance. To estimate the former, consider two distributions, D_1 and D_2, with means μ_1 and μ_2 and standard deviations σ_1 and σ_2. The difference between these distributions will have a mean of μ_1-μ_2, and a standard deviation of $(\sigma_1^2 + \sigma_2^2 - 2\text{cov}(\sigma_1,\sigma_2))^{0.5}$, where $\text{cov}(\sigma_1,\sigma_2)$ represents the covariance between D_1 and D_2. If we assume that the covariance is zero,[4] the expected standard deviation is simply $(\sigma_1^2+\sigma_2^2)^{0.5}$. Extending this to the SCE which is a difference of difference scores, the resulting *SD* for a given subject is $(\sigma_{cC}^2+\sigma_{cI}^2+\sigma_{iC}^2+\sigma_{iI}^2)^{0.5}$. Assuming that the *SD* across conditions is the same for a given subject, this simplifies to $(4\sigma_{\text{overall}}^2)^{0.5}$, or 2σ. Based on our data, the average *SD* for a given subject was around $SD = 190$ ms, so $SD_{\text{SCE}} = 380$ ms.

The second issue to consider is the interindividual variance in the size of the SCE. Too little variance leads to a so-called restriction of range issue that decreases the strength of the observed correlation. To estimate how these sources of variability impact our ability to detect a reliable SCE we used MATLAB to

[4] This is probably not a valid assumption. For example, if you correlate trial_n with trial_{n-1} across all trials (i.e., the autocorrelation), we get $r\text{s} \sim=$ .15, though for a few subjects it approaches $r\text{s} = .40$ (see also Laming, 1979).

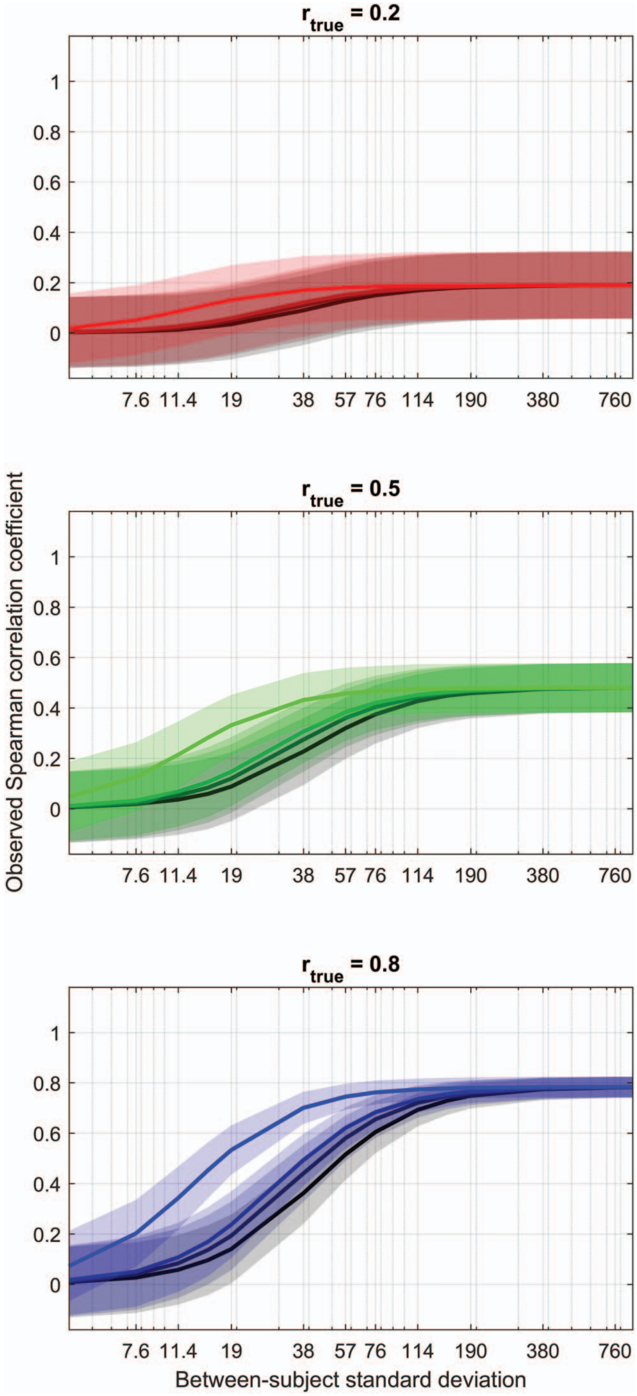

Figure 4. The observed Spearman correlation coefficient as a function of within and between subject variance when the true correlation is .80 (bottom), .50 (middle), and .20 (top), with the shaded area around each line representing the 95% CI for the observed correlation. The most saturated (i.e. most blue/green/red) line is when within-subject variance is smallest and decreases as the color gets darker (i.e., 380/sqrt(*N*)) for N = 1,800, 180, 135, and 90. Note that the width of the 95% CI decreases as both between-subject variance increases and as the true correlation increases. See the online article for the color version of this figure.

conduct N = 10,000 simulations with N = 200 subjects (i.e., close to our sample size) that are shown in Figure 4.

Each set of N = 200 subjects was created to have a between-subjects variance (*SD*s = [3.8, 38, 76, 95, 190, 380, 760, 1,520, 1,900, 3,800]) that was equivalent at Time 1 and Time 2, and the vectors representing the Time 1 and Time 2 data were zero-centered to ensure that the overall SCE was zero. We then used Cholesky factorization to set the true correlation, r_{true}, to .20 (top panel), .50 (middle panel), or .80 (bottom panel). Finally, for each of the N = 200 subjects at each of Time 1 and Time 2, we added an SCE = 10 ± 380/sqrt(N_{trials}) ms, where N_{trials} (per cell) was 90, 180, or 1,800. N_{trials} is depicted as a separate line with shading representing the 95% CI of the estimate. Note that one-tailed .95 power would be the lower limit of this shading. The *x*-axis shows the between-subjects standard deviation, and the *y*-axis shows the observed Spearman correlation coefficient. For our purposes, the critical observation is that $r_{obs} = r_{true}$ as long as the true between subject *SD* is above 190 ms. Moreover, for within-subject *SD* = 380 ms, so long as the between-subjects *SD* remains above ~70 ms, the measured correlation ought to be an accurate reflection of the underlying true correlation. Therefore, the fact that we failed to find a reliable split-half SCE is consistent with our interpretation that the SCE is not stable within in an individual.

Conclusion

Over the past 20 years, the interaction between DLPFC and ACC has formed the basis for our understanding of cognitive control. The conflict monitoring hypothesis' purported ability to explain a multitude of effects across a range of paradigms has contributed greatly to its success. However, the fact that the size of the error-related slowing effect is correlated across a set of similar tasks, but the size of the sequential congruency effect is not correlated across those same tasks add to a body of literature (e.g., Funes, Lupiáñez, & Humphreys, 2010) that questions the extent to which these effects arise from the same mechanism. Additionally, the fact that the SCE is not stable within an individual highlight the need to think more critically about the other sources of information cognitive control uses to allow us to produce goal-directed behavior.

References

Akçay, C., & Hazeltine, E. (2007). Conflict monitoring and feature overlap: Two sources of sequential modulations. *Psychonomic Bulletin & Review, 14,* 742–748. http://dx.doi.org/10.3758/BF03196831

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Belmont, CA: Wadsworth. Inc.

Baddeley, A., & Della Sala, S. (1996). Working memory and executive control. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences, 351,* 1397–1403. http://dx.doi.org/10.1098/rstb.1996.0123

Bender, A. D., Filmer, H. L., Garner, K. G., Naughtin, C. K., & Dux, P. E. (2016). On the relationship between response selection and response inhibition: An individual differences approach. *Attention, Perception & Psychophysics, 78,* 2420–2432. http://dx.doi.org/10.3758/s13414-016-1158-8

Blais, C. (2008). Random without replacement is not random: Caveat emptor. *Behavior Research Methods, 40,* 961–968. http://dx.doi.org/10.3758/BRM.40.4.961

Blais, C., Robidoux, S., Risko, E. F., & Besner, D. (2007). Item-specific adaptation and the conflict-monitoring hypothesis: A computational model. *Psychological Review, 114,* 1076–1086. http://dx.doi.org/10.1037/0033-295X.114.4.1076

Blais, C., Stefanidi, A., & Brewer, G. A. (2014). The Gratton effect remains after controlling for contingencies and stimulus repetitions. *Frontiers in Psychology, 5,* 1207. http://dx.doi.org/10.3389/fpsyg.2014.01207

Blais, C., & Verguts, T. (2012). Increasing set size breaks down sequential congruency: Evidence for an associative locus of cognitive control. *Acta Psychologica, 141,* 133–139. http://dx.doi.org/10.1016/j.actpsy.2012.07.009

Borgmann, K. W. U., Risko, E. E., Stolz, J. A., & Besner, D. (2007). Simon says: Reliability and the role of working memory and attentional control in the Simon task. *Psychonomic Bulletin & Review, 14,* 313–319. http://dx.doi.org/10.3758/BF03194070

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

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Sciences, 8,* 539–546. http://dx.doi.org/10.1016/j.tics.2004.10.003

Carter, C. S., & van Veen, V. (2007). Anterior cingulate cortex and conflict detection: An update of theory and data. *Cognitive, Affective & Behavioral Neuroscience, 7,* 367–379. http://dx.doi.org/10.3758/CABN.7.4.367

Cavanagh, J. F., & Frank, M. J. (2014). Frontal theta as a mechanism for cognitive control. *Trends in Cognitive Sciences, 18,* 414–421. http://dx.doi.org/10.1016/j.tics.2014.04.012

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

Danielmeier, C., & Ullsperger, M. (2011). Post-error adjustments. *Frontiers in Psychology, 2,* 233.

Davidson, D. J., Zacks, R. T., & Williams, C. C. (2003). Stroop interference, practice, and aging. *Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition, 10,* 85–98.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18,* 193–222. http://dx.doi.org/10.1146/annurev.ne.18.030195.001205

Dutilh, G., Vandekerckhove, J., Forstmann, B. U., Keuleers, E., Brysbaert, M., & Wagenmakers, E. J. (2012). Testing theories of post-error slowing. *Attention, Perception & Psychophysics, 74,* 454–465.

Egner, T. (2007). Congruency sequence effects and cognitive control. *Cognitive, Affective & Behavioral Neuroscience, 7,* 380–390. http://dx.doi.org/10.3758/CABN.7.4.380

Egner, T. (2014). Creatures of habit (and control): A multi-level learning perspective on the modulation of congruency effects. *Frontiers in Psychology, 5,* 1247.

Engle, R. W., & Kane, M. J. (2003). *Executive attention, working memory capacity, and a two-factor theory of cognitive control*. New York, NY: Elsevier. http://dx.doi.org/10.1016/S0079-7421(03)44005-X

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191.

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

Funes, M. J., Lupiáñez, J., & Humphreys, G. (2010). Sustained vs. transient cognitive control: Evidence of a behavioral dissociation. *Cognition, 114,* 338–347. http://dx.doi.org/10.1016/j.cognition.2009.10.007

Gratton, G., Coles, M. G., & Donchin, E. (1992). Optimizing the use of information: Strategic control of activation of responses. *Journal of Experimental Psychology: General, 121,* 480–506. http://dx.doi.org/10.1037/0096-3445.121.4.480

Hajcak, G., McDonald, N., & Simons, R. F. (2003). To err is autonomic: Error-related brain potentials, ANS activity, and post-error compensatory behavior. *Psychophysiology, 40,* 895–903. http://dx.doi.org/10.1111/1469-8986.00107

Hedge, C., Powell, G., & Sumner, P. (2017). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*. Advance online publication. http://dx.doi.org/10.3758/s13428-017-0935-1

Hulley, S. B., Cummings, S. R., Browner, W. S., Grady, D. G., & Newman, T. B. (2013). *Designing clinical research*. Philadelphia, PA: Lippincott Williams & Wilkins.

Kerns, J. G. (2006). Anterior cingulate and prefrontal cortex activity in an FMRI study of trial-to-trial adjustments on the Simon task. *NeuroImage, 33,* 399–405. http://dx.doi.org/10.1016/j.neuroimage.2006.06.012

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

Keye, D., Wilhelm, O., Oberauer, K., & Stürmer, B. (2013). Individual differences in response conflict adaptations. *Frontiers in Psychology, 4,* 947. http://dx.doi.org/10.3389/fpsyg.2013.00947

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,* 762–776. http://dx.doi.org/10.1007/s00426-008-0188-9

Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional overlap: Cognitive basis for stimulus-response compatibility—A model and taxonomy. *Psychological Review, 97,* 253–270. http://dx.doi.org/10.1037/0033-295X.97.2.253

Laming, D. (1979). Autocorrelation of choice-reaction times. *Acta Psychologica, 43,* 381–412. http://dx.doi.org/10.1016/0001-6918(79)90032-5

Larson, M. J., Clayson, P. E., Keith, C. M., Hunt, I. J., Hedges, D. W., Nielsen, B. L., & Call, V. R. (2016). Cognitive control adjustments in healthy older and younger adults: Conflict adaptation, the error-related negativity (ERN), and evidence of generalized decline with age. *Biological Psychology, 115,* 50–63. http://dx.doi.org/10.1016/j.biopsycho.2016.01.008

Lie, C.-H., Specht, K., Marshall, J. C., & Fink, G. R. (2006). Using fMRI to decompose the neural processes underlying the Wisconsin Card Sorting Test. *NeuroImage, 30,* 1038–1049. http://dx.doi.org/10.1016/j.neuroimage.2005.10.031

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95,* 492–527. http://dx.doi.org/10.1037/0033-295X.95.4.492

Logan, G. D. (1990). Repetition priming and automaticity: Common underlying mechanisms? *Cognitive Psychology, 22,* 1–35. http://dx.doi.org/10.1016/0010-0285(90)90002-L

Lumsden, J. (1976). Test theory. *Annual Review of Psychology, 27,* 251–280. http://dx.doi.org/10.1146/annurev.ps.27.020176.001343

MacDonald, A. W., III, Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science, 288,* 1835–1838. http://dx.doi.org/10.1126/science.288.5472.1835

Maurer, J. M., Steele, V. R., Cope, L. M., Vincent, G. M., Stephen, J. M., Calhoun, V. D., & Kiehl, K. A. (2016). Dysfunctional error-related processing in incarcerated youth with elevated psychopathic traits. *Developmental Cognitive Neuroscience, 19,* 70–77. http://dx.doi.org/10.1016/j.dcn.2016.02.006

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

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

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24,* 167–202. http://dx.doi.org/10.1146/annurev.neuro.24.1.167

Murphy, K. R., & Davidshofer, C. O. (1988). *Psychological testing*. Englewood Cliffs, NJ: Prentice-Hall. Retrieved from http://www.academia.edu/download/31102694/058927786.pdf

Nieuwenhuis, S., Stins, J. F., Posthuma, D., Polderman, T. J. C., Boomsma, D. I., & de Geus, E. J. (2006). Accounting for sequential trial effects in the flanker task: Conflict adaptation or associative priming? *Memory & Cognition, 34,* 1260–1272. http://dx.doi.org/10.3758/BF03193270

Nigbur, R., Ivanova, G., & Stürmer, B. (2011). Theta power as a marker for cognitive interference. *Clinical Neurophysiology, 122,* 2185–2194. http://dx.doi.org/10.1016/j.clinph.2011.03.030

Norman, D. A., & Shallice, T. (1986). Attention to action. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.), *Consciousness and self-regulation* (pp. 1–18). New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4757-0629-1_1

Notebaert, W., Gevers, W., Verbruggen, F., & Liefooghe, B. (2006). Top-down and bottom-up sequential modulations of congruency effects. *Psychonomic Bulletin & Review, 13,* 112–117. http://dx.doi.org/10.3758/BF03193821

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

Pratte, M. S., Rouder, J. N., Morey, R. D., & Feng, C. (2010). Exploring the differences in distributional properties between Stroop and Simon effects using delta plots. *Attention, Perception & Psychophysics, 72,* 2013–2025.

Puccioni, O., & Vallesi, A. (2012). Sequential congruency effects: Disentangling priming and conflict adaptation. *Psychological Research, 76,* 591–600. http://dx.doi.org/10.1007/s00426-011-0360-5

Rabbitt, P. M. (1968). Three kinds of error-signaling responses in a serial choice task. *The Quarterly Journal of Experimental Psychology, 20,* 179–188. http://dx.doi.org/10.1080/14640746808400146

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., & Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science, 306,* 443–447. http://dx.doi.org/10.1126/science.1100301

Riesel, A., Weinberg, A., Endrass, T., Meyer, A., & Hajcak, G. (2013). The ERN is the ERN is the ERN? Convergent validity of error-related brain activity across different tasks. *Biological Psychology, 93,* 377–385. http://dx.doi.org/10.1016/j.biopsycho.2013.04.007

Schmidt, J. R. (2013). Questioning conflict adaptation: Proportion congruent and Gratton effects reconsidered. *Psychonomic Bulletin & Review, 20,* 615–630. http://dx.doi.org/10.3758/s13423-012-0373-0

Schmidt, J. R., Crump, M. J. C., Cheesman, J., & Besner, D. (2007). Contingency learning without awareness: Evidence for implicit control. *Consciousness and Cognition, 16,* 421–435. http://dx.doi.org/10.1016/j.concog.2006.06.010

Schmidt, J. R., & De Houwer, J. (2011). Now you see it, now you don't: Controlling for contingencies and stimulus repetitions eliminates the Gratton effect. *Acta Psychologica, 138,* 176–186. http://dx.doi.org/10.1016/j.actpsy.2011.06.002

Schmidt, J. R., & Weissman, D. H. (2014). Congruency sequence effects without feature integration or contingency learning confounds. *PLoS ONE, 9,* e102337. http://dx.doi.org/10.1371/journal.pone.0102337

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15,* 72–101. http://dx.doi.org/10.2307/1412159

Töllner, T., Wang, Y., Makeig, S., Müller, H. J., Jung, T.-P., & Gramann, K. (2017). Two independent frontal midline theta oscillations during conflict detection and adaptation in a Simon-type manual reaching task. *The Journal of Neuroscience, 37,* 2504–2515. http://dx.doi.org/10.1523/JNEUROSCI.1752-16.2017

Torres-Quesada, M., Funes, M. J., & Lupiáñez, J. (2013). Dissociating proportion congruent and conflict adaptation effects in a Simon-Stroop procedure. *Acta Psychologica, 142,* 203–210. http://dx.doi.org/10.1016/j.actpsy.2012.11.015

Ullsperger, M., Bylsma, L. M., & Botvinick, M. M. (2015). The conflict adaptation effect: It's not just priming. *Cognitive, Affective, and Behavioral Neuroscience, 5,* 467–472.

Underwood, B. J. (1975). Individual differences as a crucible in theory construction. *American Psychologist, 30,* 128–134. http://dx.doi.org/10.1037/h0076759

Ward, G., Roberts, M. J., & Phillips, L. H. (2001). Task-switching costs, Stroop-costs, and Executive Control: A Correlational Study. *The Quarterly Journal of Experimental Psychology Section A,, 54,* 491–511.

Received August 7, 2017
Revision received April 20, 2018
Accepted April 22, 2018 ■