Short Research Article

Reliability and Convergence of Conflict Effects

An Examination of Evidence for Domain-General Attentional Control

Peter S. Whitehead[1], Gene A. Brewer[2], and Chris Blais[2]

[1]Center for Cognitive Neuroscience, Duke University, Durham, NC, USA
[2]Department of Psychology, Arizona State University, AZ, USA

Abstract. Recent work in attentional control has suggested that conflict effects measured across different tasks are not reliable and by extension unrelated. The lack of correlation between these conflict effects is in juxtaposition not only to theoretical predictions of a domain-general attentional control mechanism but also to a large body of individual differences research that has used these tasks to show evidence for an attentional control construct and its relatedness to other psychological constructs. In an effort to address this, we fit hierarchical models to each task that modeled trial-to-trial variability in response times to assess the extent to which the parameter estimates for the conflict effect correlated across tasks. We compared this method of assessing shared variance to more traditional summed difference score estimates of the conflict effect by analyzing data from a large-scale individual differences experiment, in which N = 582 subjects completed a Stroop, Flanker, and Simon task. Across tasks, we found that while the reliability of the conflict was sufficiently high and the between-task conflict effect significantly correlated, the magnitude of the between-task correlation was low. We discuss the implications of these results as providing more support for a domain-specific than domain-general attentional control mechanism.

Keywords: attentional control, conflict effect, individual differences, cognitive control

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

As of June 25, 2020, Google Scholar reports that the original Stroop task paper has been cited 19,667 times (Stroop, 1935), the original Flanker task paper 6,567 times (Eriksen & Eriksen, 1974), and the original Simon task paper 1,255 times (Simon & Rudell, 1967). Furthermore, a search of titles and abstracts for the terms "Stroop," "Simon task," or "Flanker task" using the PubMed database (accessed on June 25, 2020) returns 9,203 papers, of which 1,167 were classified as "clinical trials" and 3,300 were published in the last 5 years. These tasks have also been and continue to be widely used not only in the experimental domain but also in the individual differences literature (for highly cited examples, see Friedman & Miyake, 2004; Kane & Engle, 2003). Finally, in assessing the classic articles of the psychological literature, the 5th and 8th most cited articles in the literature either establish or use the Stroop task, respectively (Ho & Hartley, 2016).

However, recent correlational research in the attentional control literature suggests that inhibitory effects measured across tasks are not reliable. Specifically, an emerging literature suggests that conflict effects (incongruent minus congruent trials) derived from common attentional control tasks such as the Simon, Flanker, and Stroop tasks may not provide evidence for a unitary attentional control or inhibition mechanism when taking a differential psychological approach in studying the shared variance between these tasks (Hedge et al., 2018; Rey-Mermet et al., 2018; Rouder & Haaf, 2019). An increasingly frequent finding is that both the within-task reliability of attentional control and inhibition measures derived from performance in the Simon, Stroop, and Flanker tasks are shockingly low, all the more surprising given their widespread use in individual differences research (Enkavi et al., 2019; Feldman & Freitas, 2016; Hedge et al., 2018; Paap & Sawi, 2016).

Furthermore, an emerging literature suggests that conflict effects (incongruent minus congruent trials) derived from common attentional control tasks such as the Simon, Flanker, and Stroop tasks may not provide evidence for a unitary attentional control or inhibition

© 2020 Hogrefe Publishing

mechanism when taking a differential psychological approach in studying the shared variance between these tasks (Hedge et al., 2018; Rey-Mermet et al., 2018; Rouder & Haaf, 2019). These new findings call into question a large number of theoretical positions related to the nature and possibly the existence of a domain general attentional control mechanism.

Measurement Reliability

One reason for obtaining low measures of reliability in each of these tasks is because the standard method of obtaining a conflict effect is to compute a difference score for each participant by aggregating across trials. This method fails to account for the fact that each participant has a different amount of trial-level variability (Hedge et al., 2018; see also Paap & Sawi, 2016). Specifically, when researchers use an aggregate score, they lose information about systematic changes in trial-to-trial variability within the task. Furthermore, multiple aggregate scores (i.e., summed response times over congruent and incongruent trials for each participant) are then used to calculate a difference score to isolate a cognitive process of interest (i.e., inhibition). Not only does this combine the error from the two original components (i.e., congruent and incongruent trials), but the use of a difference score also reduces between-participant variance while increasing the proportion of measurement error (Hedge et al., 2018; see also Lord, 1958). That is, the use of difference scores created from the aggregate of the dependent measure over multiple trials in experimental tasks as an index of attentional control or inhibition means that the effect size will be dependent on the number of trials due to compounding error (i.e., the summed error of each trial incorporated into an aggregate Stroop effect, basically the standard deviation within a condition). This violates the commonly held notion that an effect size should be independent from the length of the measurement test (Green et al., 2016) – a principle referred to as the portability of measurement (Green et al., 2016; see also Rouder & Haaf, 2019) – as the degree of measurement error contributing to the conflict effect is dependent on the length of the experiment.

To avoid issues induced by the use of aggregate scores, Rouder and Haaf (2019) used model parameters derived from the use of hierarchical linear models as indices of attentional control, which led to an improvement in measurement reliability. This is because trial-by-trial variation was explicitly modeled instead of being as a source of unmeasured variance as in traditional methods. Modeling the trial-by-trial variation using a hierarchical linear model reflects the regularization of parameter estimates through the reduction in overall error typically incorporated into a summed subject-level Stroop effect, which, as demonstrated by Rouder and Haaf (2019), allows for a better estimate of the true conflict effect, one that is purportedly not as reliant on the length of an experiment. This approach not only allows for proper interpretation of reliability of conflict measures, solving a statistical issue, but also provides a tool to answer substantive theoretical questions, allowing for a more "process-pure" assessment of shared variance in performance across tasks.

Measurement Validity: Are All Attention Control Tasks the Same?

Although the issues surrounding low measurement reliability within some attentional control and inhibition tasks are currently a focus of active research, it is important to note that the theoretical issue regarding whether these tasks should in fact correlate with one another is also under active investigation (see Hedge et al., 2018; Rey-Mermet et al., 2018; Rouder & Haaf, 2019).

The lack of between-task correlations of the conflict effect in this recent research, however, is in stark contrast to a large body of individual difference studies, showing a relation between reaction time difference scores from inhibition tasks and other measures derived from attention tasks such as antisaccade, psychomotor vigilance, and sustained attention to response task (Redick et al., 2016; Unsworth & Robison, 2017; Unsworth et al., 2009). Taken together, these measures have been statistically combined and argued to provide evidence consistent with a domain-general attentional control construct. This work shows that not only do conflict effects from various tasks correlate at the latent level with an attentional control construct, but this construct is also correlated with other psychological constructs such as working memory (Engle et al., 1999; Kane & Engle, 2003; Unsworth & Spillers, 2010) and general fluid intelligence (Unsworth et al., 2009).

It is important to remember that "attentional control" has been measured and described relatively independently from two complementary psychological traditions. From an experimental perspective, the most common measures of attention control are the conflict effects: the difference in response times between incongruent and congruent trials. These effects are extremely reliable in the sense that nearly every subject will show an effect (e.g., between trial *t*-test comparing incongruent to congruent trials will be significant for each subject), but the stability of the rank

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

 © 2020 Hogrefe Publishing

ordering of individuals on that effect is very low. From an individual differences perspective, the most common measures of attentional control are response times and/or accuracy from different tasks that have been successfully combined together to create latent factors and isolated from other constructs such as fluid intelligence and working memory capacity.

The Current Study

In the current paper, we used the hierarchical modeling method proposed by Rouder and Haaf (2019) to measure the extent to which Simon, Flanker, and Stroop conflict effects are correlated across subjects. This approach allows us to model trial-by-trial variability, which should improve the stability of the rank ordering of effects across subjects. Here, it seems that there are two logical and opposing hypotheses. It could be that there is indeed a domain-general attentional control mechanism driving the production of conflict effects (for reviews, see Miyake & Friedman, 2012; Bugg & Crump, 2012; Botvinick et al., 2001) and that the "other variables" distorting evidence for this conclusion are largely a measurement issue, not necessarily a substantive theoretical concern. Alternatively, one explanation for performance in the three attentional control tasks could be a domain-specific attentional control, where the "dimensional overlap" between different attentional control tasks underlies the production of conflict effects and the relationship between these effects in differing tasks (Kornblum et al., 1990; Kornblum, 1992, 1994). That is, the dimension-overlap taxonomy proposes three potential sources of conflict effects, of which conflict in each of these tasks results from a combination of these sources: (1) an overlap between irrelevant and relevant stimulus dimensions, (2) an overlap between irrelevant stimulus dimensions and response dimension, and (3) an overlap between relevant stimulus dimension and response dimension. An ensemble of tasks can be grouped together based on similar combinations of these three sources of conflict. Within the taxonomy from Kornblum (1994), there are eight different ensemble types, each encompassing a similar group of stimulus–response tasks. This allows us to view tasks and the similarities between them from a point of view of the relevant and irrelevant stimulus–response and stimulus–stimulus overlap. Importantly, some of the ensemble types are more theoretical than practical; Kornblum (1994) noted that ensemble six tasks would have a relevant stimulus–response dimension and a stimulus–stimulus dimension, but no irrelevant stimulus–response dimension, a set of conditions that is difficult to envision. Critical to the present study, however, are ensemble eight tasks. These are tasks in which there are a relevant stimulus–response dimension, an irrelevant stimulus–response dimension, and a stimulus–stimulus dimension (Kornblum, 1994). Type eight ensembles of tasks include Stroop tasks and other tasks similar to that. However, traditionally, the Stroop, Flanker, and Simon tasks all produce conflict from a different combination of sources, classified as coming from different ensembles, and according to the dimensional-overlap hypothesis, we would not expect the conflict produced in each of these tasks to be related.

To this end, we analyzed a dataset collected by Whitehead et al. (2018), in which ensemble eight variants of each of the Simon, Flanker, and Stroop tasks were constructed to maximize similarity between the processes generating the conflict effects (i.e., using a word-location Simon task). We report two analyses comparing the estimates of Simon, Flanker, and Stroop conflict effects measured using (1) the standard approach and (2) the hierarchical approach that will allow us to control the variability across trials for each participant. If attentional control is domain-general, we should see conflict effects correlate between all tasks. However, if attentional control is domain-specific, according to a "dimensional overlap" theory of conflict production, conflict effects would not correlate between all pairs of tasks. To preview our results, we show that the conflict effects from a Simon, Flanker, and Stroop task do correlate between-task, across all pairs, albeit weakly (r_s = .16–.29). Furthermore, given the large number of participants and trials in our tasks, we also simulated the between-task difference score correlations for factorial combinations of differing trials counts and participant numbers in order to provide guidelines on study design for conflict tasks in individual differences research.

Methods

The data and analysis scripts can be downloaded from Open Science Foundation (https://osf.io/7hp85/) in R Markdown format. We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. We analyzed the data first reported by Whitehead et al. (2018). Specifically, we combined three datasets in which participants completed a Simon, Flanker, and Stroop task, resulting in a total N = 582. These data were originally collected for Whitehead et al. (2018), and it is important to note that (1) these tasks were designed to have maximal overlap with one another and (2) the duration of each task (i.e., the number of trials) was much longer than in typical studies examining individual differences.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

© 2020 Hogrefe Publishing

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

To the former task design for maximal overlap point, (1) the Flanker task presented participants with a string of five letters and asked participants to identify the middle one (D, F, J, or K), ignoring the flanking letters using those keys on a standard QWERTY keyboard. (2) The Simon task presented participants with a directional word (RIGHT, LEFT, UP, DOWN) appearing to the right, left, top, or bottom of a fixation cross, and participants were instructed to respond to the word, not its location, using the arrow keys. Thus, this made the Simon task used more like a spatial-Stroop task, and importantly, a Type 8 ensemble tasks in which there is a relevant stimulus–response, irrelevant stimulus–response, and stimulus–stimulus conflict. (3) The Stroop task participants were presented with color words RED, BLUE, GREEN, and YELLOW in those same colors and were instructed to respond to the color of the text. This allows us to categorize all tasks as Type 8 ensembles, according to Kornblum (1994). That is, in the Simon task, the stimulus–stimulus overlap is in the word "LEFT" or "UP," for example, presented on the right side or bottom, respectively, of the screen, and the irrelevant stimulus–response dimension is created through a left/up button press to a stimulus on the right/bottom, and the relevant stimulus–response dimension is the meaning of the word "LEFT" or "UP" to the left/up button response. In the Stroop task, the relevant stimulus–response dimension is created via the button press response to the color, the irrelevant stimulus–response dimension through the mapping of color-response buttons and the identity of the color word used, and the stimulus–stimulus conflict through the conflict, or lack thereof, between the identity of the word used and the color of the text of that word. In the Flanker task, the relevant stimulus–response dimension is the mapping between central letter stimuli and responses, the irrelevant dimension created by the flanking letters and corresponding potential responses, and the stimulus–stimulus dimension through the congruency, or incongruency, between the central and flanking letters in the presented stimuli.

While, in the Stroop task, stimulus responses were randomly mapped to the d, f, j, and k keys, 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. To the latter point regarding the number of trials in each task, this important feature will allow us to provide guidelines on selecting an appropriate number of trials for a task as well as deciding whether employing the hierarchical model is advantageous in this scenario.

As all group-level conflict effects were significant in the original analysis (for descriptive results, see Table 1), these are not reported here (see Whitehead et al., 2018). However, the results from Whitehead et al. (2018) indicate the conflict effect is significant in all data used here and effect sizes quite large ($\eta^2_{pp} > .67$). In order to show group-level similarity between our tasks, δ plots for each conflict effect were constructed by dividing response times for congruent and incongruent trials into five quartiles and then subtracting (incongruent minus congruent) the mean response from each quartile (Pratte et al., 2010; Burle et al., 2014).

Table 1. In the far left two columns, mean response times in milliseconds and *SD*s across participant in parenthesis for congruent and incongruent trials in the Simon, Stroop, and Flanker tasks

	Congruent	Incongruent	Difference score	*t*-Value	*d*
Simon	568 (78)	647 (85)	79 (1.35)	16.56	0.97
Stroop	766 (126)	857 (144)	91 (2.28)	11.53	0.68
Flanker	696 (125)	743 (134)	47 (1.33)	6.21	0.36

The middle column displays the difference score (incongruent minus congruent) for each task, with the standard error across participants in parenthesis for each task. The right two columns display the *t*-value for a *t*-test of the difference score against zero and the associated effect size (Cohen's *d*) for each task.

We then submitted data to a hierarchical model to estimate the individual participant effect sizes for a within-task split-half reliability analysis of the conflict effect (incongruent minus congruent) as well as the between-task analysis of the shared variance for conflict effects (see Rouder & Haaf, 2019). Effect sizes for each were then computed using the Spearman rank-order correlation coefficient. We used lme4 (Bates et al., 2015, p. 4) to implement the following hierarchical models to assess psychometric task performance accounting for trial-by-trial variability proposed by Rouder and Haaf (2019):

$$Y_{ikl} \left| \alpha_{ij}, \theta_{ij}, \sigma^2 \right. \sim \text{Normal}\left(\alpha_{ij} + \chi_k \theta_{ij}, \sigma^2\right). \quad (1)$$

In Equation 1, l indexes the lth trial, i indexes the individual, k indexes the conflict condition (incongruent versus congruent), and j indexes the task (Simon, Stroop, or Flanker). Equation 1 was used to estimate the conflict effects in each task for the between-task conflict analysis. To estimate split-half reliability (conditioned on two levels, 0 or 1), Equation 1 was used for each task, except that j indexed the split-half condition not task.

Additionally, for comparison to the effect sizes derived from the hierarchical models, we also applied the traditional summed difference score approach to estimate conflict effects and split-half reliability for each participant in each task. For the conflict effect, we summed response times (RT)s over trials for each participant and then subtracted the mean congruent from incongruent response to create a conflict score measure for each

 © 2020 Hogrefe Publishing

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

participant. For the split-half reliability analysis, we assigned each trial as odd or even and then averaged over the odd/even trials by congruency for each participant, subtracting the congruent from incongruent response for each odd/even group and thus creating a summed odd–even split-half reliability score for each participant. We did not apply the Spearman–Brown prophecy correction formula to the split-half reliability difference score estimates, as the outcomes are practically equivalent to the hierarchical modeling approach (Rouder & Haaf, 2019).

Finally, to determine to point at which these two methods diverge in power to find an effect, we subsampled from the present data to simulate how lower trial counts and fewer participants impact the measurement of the between-task correlations in the more commonly used difference score approach. We randomly selected between N = 50–550 (in N = 50 increments) participants (from a total of 582) and between N_T = 50–450 (in 50 trial increments) correct trials (from a maximum of 1,187). We did this for a total of 1,000 simulations per factorial combination in order to determine the expected between-task correlation for each task pair using summed difference scores.

Results

Figure 1 shows the delta plots for the conflict effect measured in each task. The patterns are similar, which supports our expectation that all three tasks are Type 8 of the dimension-overlap framework (Zhang, Zhang, & Kornblum, 1999). Importantly, these plots replicate previous findings, showing that the Flanker and Stroop tasks both demonstrate the same pattern of an increasing conflict effect magnitude as response times become slower (Pratte et al., 2010). Critically, the same pattern is also seen here for our implementation of the Simon task in which the semantic meaning of the location was contrasted against the physical location of a word. This is in contrast to a traditional Simon task that shows a decrease in conflict effect magnitude as response times become slower (Pratte et al., 2010). As the three tasks were designed to have maximal overlap according to a dimension overlap taxonomy, the group-level delta plot analysis provides evidence of these tasks' similarity.

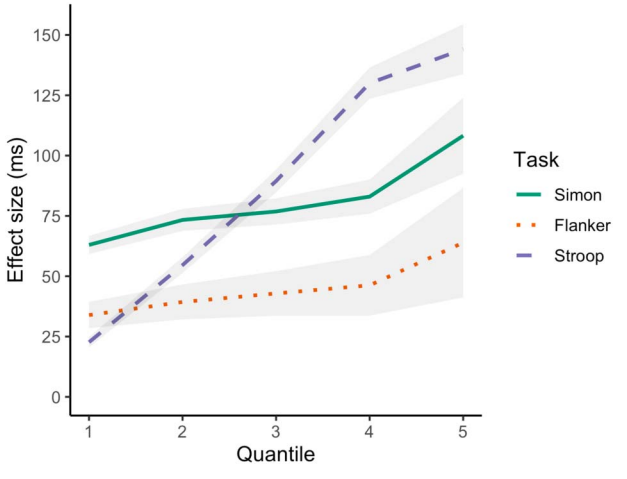

Figure 1. Delta plots for the conflict effect in each task. Quantiles are computed through the classification of each participant's response times for each task into five groups, or quantiles, applying equal probabilities per quantile. The standard error of the mean is plotted in the gray shading surrounding each line.

In applying a correlational approach using effects derived from hierarchical models, we found the conflict effect (parameter θ_{ij} in Equation 1) of each task to be moderately reliable (all split-half reliabilities $r_s > .56$, $p < .001$; Table 2; Figure 2). To evaluate the hypothesis that these three tasks index a common attentional process, we tested whether there were stable individual differences in the magnitude of the conflict effects across similar tasks by measuring the rank-order correlation of conflict effects across tasks. The between-task correlations for the conflict effects between all tasks (Table 2; Figure 2), the Simon and Stroop task ($r_s = .29, p < .001$), Stroop and Flanker task ($r_s = .16, p < .001$), and the Simon and Flanker task ($r_s = .22, p < .001$), were all significant. We will address the extent to which these small correlations support a general-level attentional mechanism in the discussion.

Both the split-half and between-task correlations using conflict effects from the summed difference score over-trials method came to similar conclusions as conflict effects derived from the hierarchical models (Table 2), illustrating that both methods yield unbiased estimates.

Table 2. Comparison of Spearman correlation coefficients derived from (a) hierarchical models and (b) the common over-trials summation and difference score computation

	Simon	Stroop	Flanker
(a) Model effects			
Simon	.65* (19.92)		
Stroop	.29* (7.31)	.65* (19.70)	
Flanker	.22* (4.84)	.16* (4.10)	.57* (18.13)
(b) Difference scores			
Simon	.61* (17.95)		
Stroop	.27* (7.01)	.50* (13.57)	
Flanker	.18* (4.09)	.13⁺ (3.46)	.31* (8.81)

Note. *$p < .001$; ⁺$p = .001$. The middle diagonal is split-half reliability in the tasks, and the lower diagonals are the correlation between the tasks. In parenthesis are t-values from Spearman correlations.

© 2020 Hogrefe Publishing

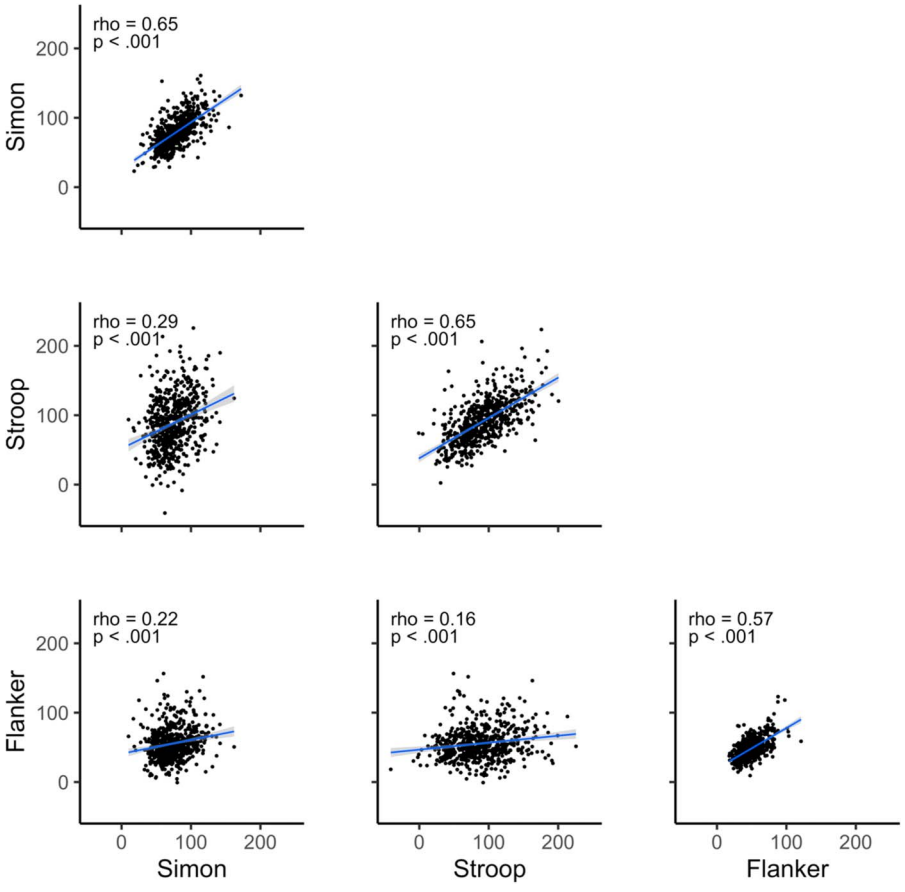

Figure 2. Correlations for the conflict effect, derived from a hierarchical model, for the combined data from Experiments 1, 2, and 3 of Whitehead et al. (2018). In the upper left corner of each graph are the Spearman correlation coefficient and the corresponding *p*-value. The graphs on the diagonal are the split-half reliabilities for each effect, and the graphs on the bottom diagonal are the between-task conflict correlations. The line in each is the line of best fit.

All split-half reliabilities were significant (r_s = .31–.61), and all between-task conflict effect correlations were significant (r_s = .13–.27). The magnitude of this correlation is an improvement over what Hedge et al. (2018) and Rouder et al. (2019) observed in their data, but these weak correlations hardly offer strong evidence for a domain-general attentional control. This is particularly so given that the tasks were designed a priori to be maximally similar.

Assessing the Boundary Conditions: Trial and Subject Counts

As an exercise in determining why previous studies have both failed and succeeded in finding between-task correlations for the conflict effects from these tasks using the current dataset (N = 582), we simulated the same between-task correlations (using the over-trials, summed difference scores) using various different participant sample sizes and task lengths. We found, on average, that significant correlations between tasks usually arose from large sample sizes (N > 350; Figure 3) in conjunction with tests of substantial length (trials > 350; Figure 3), with notable exceptions (see Figure 3 for the Simon and Stroop simulations). The clear interpretation that emerges from Figure 3 is that with a large enough sample (~N = 350 subjects) and a long enough experiment (~n = 350 trials), it is possible to find significant between-task correlations for conflict effects defined via difference scores – which seem to have an upper bound of around r = .30.

In addition to obtaining estimates for the conflict effects from the hierarchical models, the random effects of participant-level intercepts can also be estimated, allowing for a measure of overall performance in each task and as a function of the split-half reliability. That is, the intercept provides a baseline performance measure, akin to an overall response time measure, irrespective of experimental conditions, for each subject. These intercept estimates are extremely stable with a task (r_s > .96 for all tasks; Figure 4) and correlate across tasks at a more modest r_s > .48 (Figure 4). Thus, if one were to consider the overall performance in a task designed to measure attentional control to serve as an index of attentional control, this might serve as a reliable, domain-general measure.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

 © 2020 Hogrefe Publishing

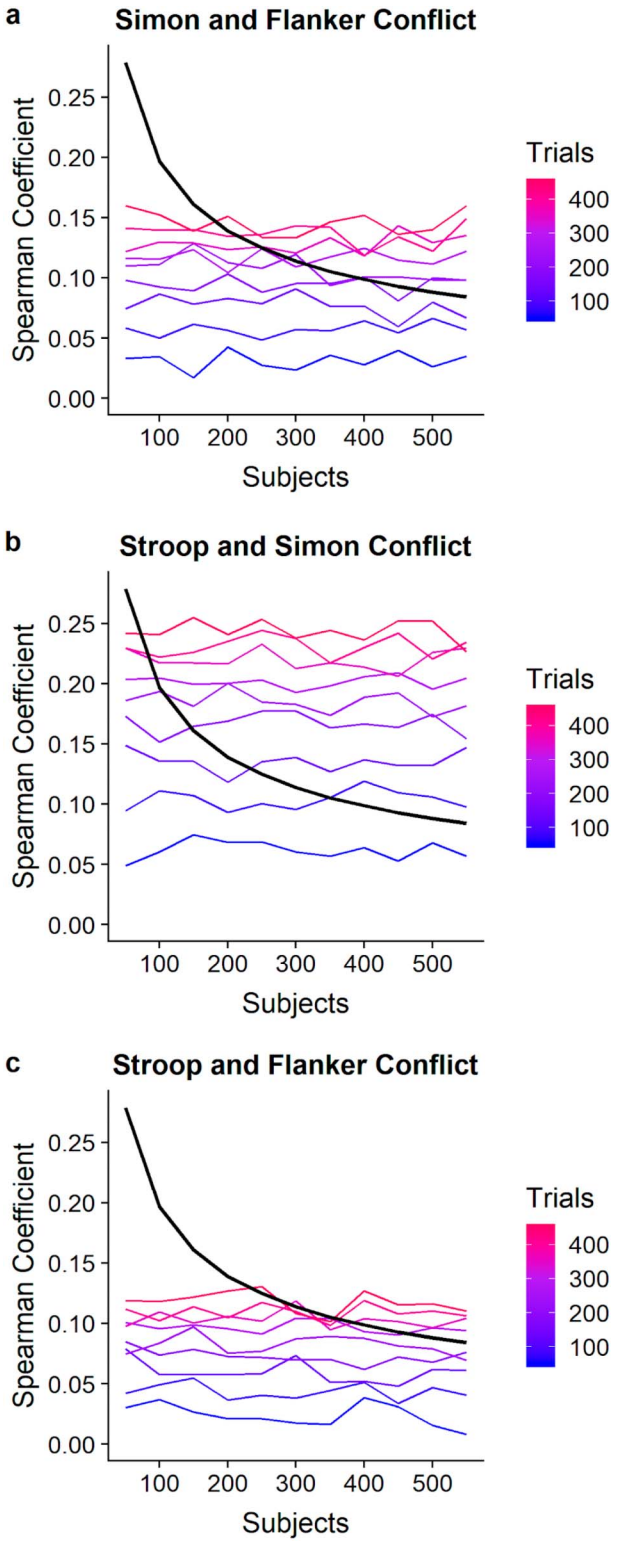

Figure 3. (a, b, and c) Simulations of the between-task Spearman's ρ correlations for the conflict effect, derived from the common over-trial summation difference score computation, as a function of the number of trials and the number of subjects included in a task. The black line in each visualization of the number of subjects needed for a significant Spearman's ρ correlation as a function of the magnitude of the correlation.

Discussion

In analyzing within-task reliability and between-task correlations across three tasks commonly used to study attentional control – the Simon, Stroop, and Flanker tasks – we have demonstrated that (1) at a group level, these tasks are quite similar (Figure 1); (2) using hierarchical linear models, as proposed in Rouder and Haaf (2019), we demonstrate significant, albeit small, correlations in all between-task conflict effect pairs (Figure 2), however; (3) the benefit of hierarchical models is reduced when using a large dataset (here, $N = 582$) and similar conclusions can be made using over-trials summed difference scores (Table 2) and similarly large datasets (Figure 3); and (4) stable between-task correlations are present at the random intercept level.

Although these between-task correlations of conflict effects meet conventional statistical standards (i.e., $ps < .05$; Figure 2), suggesting that our findings diverge from those reported by Rouder and Haaf (2019) and Hedge et al. (2018), these results do not show strong evidence for a domain-general attentional control. In fact, given that our ability to observe these effects rested solely on a particularly large number of subjects and number of trials, these results indicate a likely domain-specific deployment of attentional control in each task. Given the degree of group-level similarity for these tasks (Figure 1), these weak correlations are problematic for more domain-general theories of attentional control. In addition, from a more practical standpoint, these results also suggest that one reason previous research has not found evidence for an attentional control construct when using difference scores in these tasks is that most studies are limited to a much smaller number of subjects and trials due to time constraints (see Figure 3 for a simulation).

The argument for a domain-specific attentional control is largely based on the dimensional-overlap framework (Kornblum et al., 1990; Kornblum, 1992, 1994; but also see Egner, 2008, 2014). This taxonomy of stimulus–response compatibility tasks proposed by Kornblum et al. (1990) allows for the categorization of common performance tasks according to the overlap between stimulus and response sets, on both the task-relevant and task-irrelevant dimensions, under a unified framework that accounts for the bottom-up differences in the creation of "compatibility" (e.g., conflict) between tasks. That is, the multiple combinations in which the mapping of a stimulus to a response option is accomplished can be categorized by the specific dimensions of overlap between a stimulus and a response. Depending on the dimensional overlap of stimulus and response features of a task, the processing of compatibility occurs at a different stage, thus producing

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

© 2020 Hogrefe Publishing

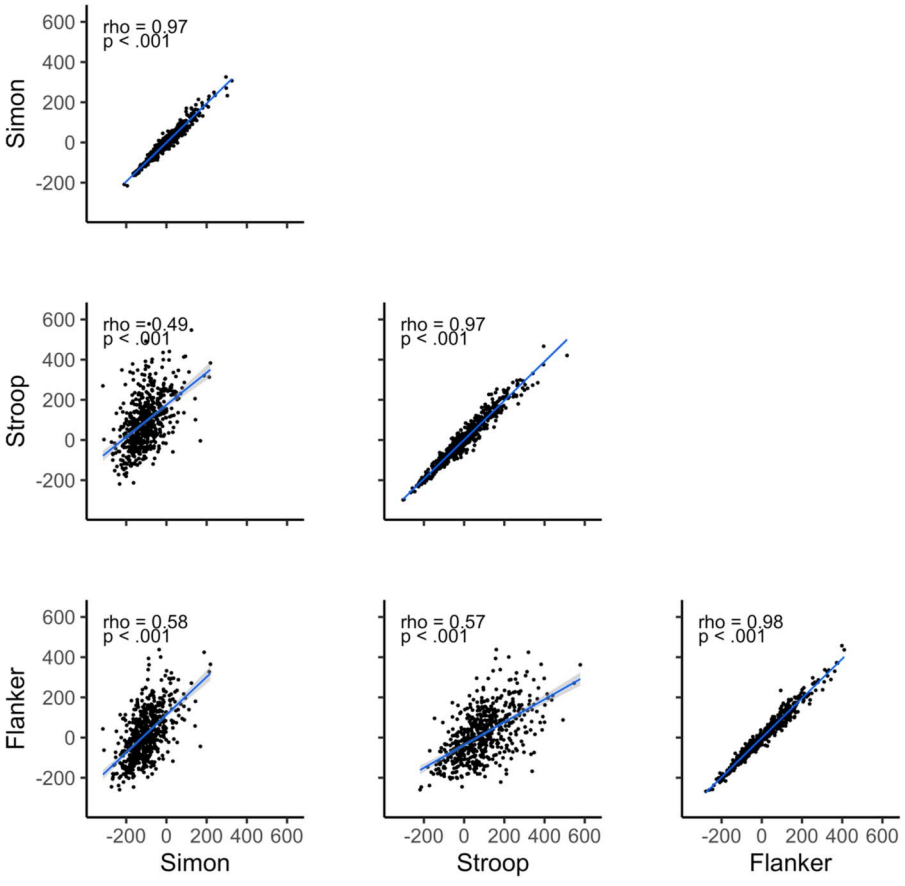

Figure 4. Correlations for the random effect of each participant's intercept (i.e., overall base response time), derived from a hierarchical model, for the combined data from Experiments 1, 2, and 3 of Whitehead et al. (2018). In the upper left corner of each graph are the Spearman correlation coefficient and corresponding *p*-value. The graphs on the diagonal are the split-half reliabilities for each effect, and the graphs on the bottom diagonal are the between-task conflict correlations. The line in each is the line of best fit.

domain-specific conflict effects for each type of task (Kornblum, 1992, 1994; Kornblum et al., 1990). The more the matching dimensions of stimulus and response overlap between two tasks, the more the similar conflict is produced in a similar manner between those tasks.

Here, while we designed each task to provide maximum dimensional overlap (see Figure 1), it is possible that the differences between each task are still salient enough to result in different loci of conflict for each task. These perceptual differences may lead to recruiting of different cognitive mechanisms between tasks, such as automatic word reading in a Stroop task or the narrowing of spatial attentional in a Flanker, which are not directly accounted for by the dimensional-overlap framework. Like a domain-general attentional control, the domain-specific taxonomy proposed in the dimensional-overlap framework still adds a layer of abstraction to task components, categorizing design elements of a task in order to theoretically equate loci of conflict. Categorization, by definition, is the grouping of different items along a known axis (see Anderson, 1991). The salience of one axis necessitates the ignoring of other known axes for potential categorization (or even unknown axes) in order to accomplish the original goal. Thus, the low between-task correlations in these tasks designed for maximal stimulus and response conflict overlap within a dimensional-overlap framework indicates that attentional control is not just unlikely to be deployed in a domain-general manner, but potentially in domain-specific manner driven by the low-level task differences between tasks unaccounted for by the dimensional-overlap framework.

That there is low reliability across tasks is consistent with prior theoretical work positing that congruency effects are driven by multilevel, domain-specific mechanisms (Egner, 2008). Due to the domain specificity of the underlying conflict effect in each task, it is likely that cognitive control is applied in a parallel, domain-specific manner (Egner, 2008, see also 2014). Empirical support for this domain-specific position can be found in a lack of correlation between adaptive control effects produced by these same tasks, while shared variance across error-detection effects demonstrates support for domain-general error processing (Funes et al., 2010; Whitehead et al., 2018). While these prior investigations and theoretical accounts have focused on the cognitive control domain and the production of associated control effects

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

 © 2020 Hogrefe Publishing

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

derived from conflict effects, it would follow that this domain-specific theoretical position could also be applied to the basic conflict effect indexing attentional control that underlies these adaptive control effects.

Furthermore, this view would suggest that findings from prior individual differences research showing a relationship between these conflict effects and other attention control measures (i.e., the antisaccade, psychomotor vigilance, and sustained attention to response tasks; Engle et al., 1999; Kane & Engle, 2003; Redick et al., 2016; Unsworth & Robison, 2017; Unsworth & Spillers, 2010; Unsworth et al., 2009) are not necessarily inconsistent with other results demonstrating the lack of shared variance between these conflict effects (Hedge et al., 2018; Rouder & Haaf, 2019; see also Paap & Sawi, 2016). Specifically, prior latent variable modeling research has aimed to extract common variance across a set of measures that are thought to reflect some underlying ability (in this case, the ability to control attention across a set of domains). The current study highlights that given a large enough sample size, and long enough test, evidence for a domain-general attentional control can be found, confirming the results seen in the prior individual differences literature.

Although the conflict effect does not strongly correlate across tasks, overall performance ($r_s > .48$; Figure 4) and performance on incongruent trials ($r_s > .48$; see Figure 1 in Electronic Supplementary Material, ESM 1) are highly related between tasks. This result is similar to the common approach in individual differences research (e.g., working memory complex span tasks) where an overall task performance measure, not within-task contrasts, is utilized and indeed shows high levels of convergence with a theoretical construct. This suggests the possibility that a focus on measuring "process pure" measures of attentional control could contribute to the lack of shared variance between-task measures. That is, while these tasks may have high internal reliability in what they measure – a process pure measure of attentional control in a Flanker task, for example – the relative increase of internal validity for measuring a task-specific construct may then decrease the external validity of these tasks to measure a common construct, such as attentional control (Cronbach & Meehl, 1955).

Future studies that focus on general performance measures in each of these conflict tasks could partition the variance not using difference scores but with other attention tasks (e.g., vigilance and antisaccade). The performance measure used for each task on its own (i.e., the intercept estimate for each participant or overall processing speed) would not be able to differentiate between attentional control and overall processing. However, the use of other attentional control tasks to partition the variance from this measure in order to isolate a common attentional control construct, demonstrated in the shared variance between multiple tasks, using partial correlation matrices, Bayesian network analysis (see Epskamp et al., 2018; Epskamp & Fried, 2018), or confirmatory factor analysis (Unsworth et al., 2009), is far more likely to succeed given the large differences in reliability for each of these measures (i.e., ~.30 for difference scores under ideal conditions to >.95 for overall performance measures from model estimates under less than ideal conditions; see Figure 2 and 4, and Table 2).

The implications of the low-shared variance in these tasks providing evidence for a domain-general attentional control mechanism on clinical work and other psychological domains remain unclear. These tasks are often used to index performance on other measures or to distinguish between various clinical disorders under the assumption that there exists a domain-general attentional control mechanism (for reviews, see Barch & Sheffield, 2017; Chaarani et al., 2017; de Zeeuw & Durston, 2017). The low correlation in the size of the conflict effect across tasks indicates the need for more theoretically guided use of attentional control tasks in this research, with special consideration given to the low-level features of a task that lead to the production of conflict. Further work and consensus on the theoretical and practical implications of the dimensional overlap view and other data should be undertaken by the attentional control community at large in order to resolve recent and ongoing discrepancies between theoretical positions on attentional control and their corresponding experimental findings, individual differences work, and application in clinical research.

In summary, using the hierarchical linear models proposed by Rouder and Haaf (2019), we were able to show the presence of a moderate correlation in the size of the conflict effect measured across Simon, Flanker, and Stroop tasks reanalyzed from Whitehead et al. (2018). Furthermore, when analyzing the combined data gathered in Whitehead et al. (2018), we show tentative convergence with a domain-general theory of production of conflict effects for measuring attentional control, as all between-task correlations of conflict effects showed a significant relationship. However, we also show that this weak correlation is highly dependent on the sample size and length of task. Therefore, we propose that while these results do not speak against a domain-specific attentional control mechanism, these low correlations showing some shared variance are consistent with the idea of a domain-general mechanism. However, given its relatively weak support, perhaps it no longer deserves its place as a critical node in the information processing stream.

© 2020 Hogrefe Publishing

Electronic Supplementary Material

The electronic supplementary material is available with the online version of the article at https://doi.org/10.1027/1618-3169/a000495

ESM 1. Figure 1: Correlations for the mean of incongruent trials for each subject for combined data from Experiments 1, 2, and 3 from Whitehead et al. (2018). Figure 2: Correlations for the conflict effect, derived from a hierarchical model, for combined data from Experiments 1, 2, and 3 from Whitehead et al. (2018).

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429. https://doi.org/10.1037/0033-295X.98.3.409

Barch, D. M., & Sheffield, J. M. (2017). Cognitive control in schizophrenia: Psychological and neural mechanisms. In T. Egner (Ed.), *The Wiley handbook of cognitive control* (pp. 556–580). Wiley. https://doi.org/10.1002/9781118920497.ch31

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

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. https://doi.org/10.1037/0033-295X.108.3.624

Bugg, J. M., & Crump, M. J. C. (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. https://doi.org/10.3389/fpsyg.2012.00367

Burle, B., Spieser, L., Servant, M., & Hasbroucq, T. (2014). Distributional reaction time properties in the Eriksen task: Marked differences or hidden similarities with the Simon task? *Psychonomic Bulletin & Review*, *21*(4), 1003–1010. https://doi.org/10.3758/s13423-013-0561-6

Chaarani, B., Spechler, P. A., Hudson, K. E., Foxe, J. J., Potter, A. S., & Garavan, H. (2017). The neural basis of response inhibition and substance abuse. In T. Egner (Ed.), *The Wiley handbook of cognitive control* (pp. 581–601). Wiley. https://doi.org/10.1002/9781118920497.ch32

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

de Zeeuw, P., & Durston, S. (2017). Cognitive control in attention deficit hyperactivity disorder. In T. Egner (Ed.), *The Wiley handbook of cognitive control* (pp. 602–618). Wiley. https://doi.org/10.1002/9781118920497.ch33

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

Egner, T. (2014). Creatures of habit (and control): A multi-level learning perspective on the modulation of congruency effects. *Frontiers in Psychology*, *5*, 1247. https://doi.org/10.3389/fpsyg.2014.01247

Engle, R. W., Kane, M. J., & Tuholski, S. W. (1999). Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. In A. Miyake (Ed.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 102–134). Cambridge University Press. https://doi.org/10.1017/CBO9781139174909.007

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences*, *116*(12), 5472–5477. https://doi.org/10.1073/pnas.1818430116

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, *50*(1), 195–212. https://doi.org/10.3758/s13428-017-0862-1

Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods*, *24*(4), 617–634.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149. doi: 10.3758/BF03203267

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. https://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. https://doi.org/10.1037/0096-3445.133.1.101

Funes, M. J., Lupiáñez, J., & Humphreys, G. (2010). Analyzing the generality of conflict adaptation effects. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(1), 147–161. https://doi.org/10.1037/a0017598

Green, S. B., Yang, Y., Alt, M., Brinkley, S., Gray, S., Hogan, T., & Cowan, N. (2016). Use of internal consistency coefficients for estimating reliability of experimental task scores. *Psychonomic Bulletin & Review*, *23*(3), 750–763. https://doi.org/10.3758/s13423-015-0968-3

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

Ho, Y.-S., & Hartley, J. (2016). Classic articles in psychology in the science citation index expanded: A bibliometric analysis. *British Journal of Psychology*, *107*(4), 768–780. https://doi.org/10.1111/bjop.12163

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. https://doi.org/10.1037/0096-3445.132.1.47

Kornblum, S. (1992). *Dimensional overlap and dimensional relevance in stimulus–response and stimulus–stimulus compatibility*. North-Holland.

Kornblum, S. (1994). The way irrelevant dimensions are processed depends on what they overlap with: The case of Stroop- and Simon-like stimuli. *Psychological Research*, *56*(3), 130–135. https://doi.org/10.1007/BF00419699

Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional overlap: Cognitive basis for stimulus–response compatibility – A model and taxonomy. *Psychological Review*, *97*(2), 253–270.

Lord, F. M. (1958). The utilization of unreliable difference scores. *ETS Research Bulletin Series*, *1958*(1), i-6. https://doi.org/10.1002/j.2333-8504.1958.tb00077.x

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science*, *21*(1), 8–14. https://doi.org/10.1177/0963721411429458

 © 2020 Hogrefe Publishing

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

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. https://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. https://doi.org/10.1037/xlm0000450

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

Simon, J. R., & Rudell, A. P. (1967). Auditory S–R compatibility: The effect of an irrelevant cue on information processing. *Journal of Applied Psychology*, *51*(3), 300–304. https://doi.org/10.1037/h0020586

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

Unsworth, N., & Robison, M. K. (2017). The importance of arousal for variation in working memory capacity and attention control: A latent variable pupillometry study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(12), 1962–1987. https://doi.org/10.1037/xlm0000421

Unsworth, N., & Spillers, G. J. (2010). Working memory capacity: Attention control, secondary memory, or both? A direct test of the dual-component model. *Journal of Memory and Language*, *62*(4), 392–406. https://doi.org/10.1016/j.jml.2010.02.001

Unsworth, N., Spillers, G. J., & Brewer, G. A. (2009). Examining the relations among working memory capacity, attention control, and fluid intelligence from a dual-component framework. *Psychological Test and Assessment Modeling*, *51*(4), 388–402.

Whitehead, P. S., Brewer, G. A., & Blais, C. (2018). Are cognitive control processes reliable? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(5), 765–778. https://doi.org/10.1037/xlm0000632

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

History
Received January 7, 2020
Revision received June 26, 2020
Accepted September 10, 2020
Published online December 4, 2020

Acknowledgments
We thank Christina Bejjani for helpful comments on an earlier version of this draft. We also 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.

Open Data
All data from these experiments can be downloaded at https://osf.io/7hp85/.

Funding
This work was supported by the National Science Foundation grant number 1632291 awarded to Gene A. Brewer.

Peter S. Whitehead
Center for Cognitive Neuroscience
Duke University
Box 90999 LSRC
Durham, NC 27710
USA
peter.whitehead@duke.edu

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.