American Psychological Association

Journal of Experimental Psychology: General

ISSN: 0096-3445

2021, Vol. 150, No. 2, 242–275
http://dx.doi.org/10.1037/xge0000783

A Toolbox Approach to Improving the Measurement of Attention Control

Christopher Draheim, Jason S. Tsukahara, Jessie D. Martin, Cody A. Mashburn, and Randall W. Engle
Georgia Institute of Technology

Cognitive tasks that produce reliable and robust effects at the group level often fail to yield reliable and valid individual differences. An ongoing debate among attention researchers is whether conflict resolution mechanisms are task-specific or domain-general, and the lack of correlation between most attention measures seems to favor the view that attention control is not a unitary concept. We have argued that the use of difference scores, particularly in reaction time (RT), is the primary cause of null and conflicting results at the individual differences level, and that methodological issues with existing tasks preclude making strong theoretical conclusions. The present article is an empirical test of this view in which we used a toolbox approach to develop and validate new tasks hypothesized to reflect attention processes. Here, we administered existing, modified, and new attention tasks to over 400 participants (final N = 396). Compared with the traditional Stroop and flanker tasks, performance on the accuracy-based measures was more reliable, had stronger intercorrelations, formed a more coherent latent factor, and had stronger associations to measures of working memory capacity and fluid intelligence. Further, attention control fully accounted for the relationship between working memory capacity and fluid intelligence. These results show that accuracy-based measures can be better suited to individual differences investigations than traditional RT tasks, particularly when the goal is to maximize prediction. We conclude that attention control is a unitary concept.

Keywords: individual differences, attention control, measurement

Executive attention is studied across many disciplines of psychology and plays a central role in most models of higher-order cognition (Atksinon & Shiffrin, 1968; Baddeley & Hitch, 1998; Botvinick et al., 2004; Egeth & Yantis, 1997; Norman & Shallice, 1986; Posner & DiGirolamo, 1998; Shipstead, Harrison, & Engle, 2016). Broadly defined, executive attention guides the control of thoughts and behavior in a goal-driven manner and is particularly important when there is conflict between more automatic processes and one's intentions. It has been shown that individual differences in the ability of executive attention, which we will refer to as attention control, predict higher-order cognitive abilities (Engle, 2002) and are important for many every-day behaviors, including self-control (Broadway, Redick, & Engle, 2010), emotional regulation (Schmeichel & Demaree, 2010), and task-engagement (Miller & Cohen, 2001; Botvinick, Cohen, & Carter, 2004). Still, there remain many theoretical questions as to the nature of attention control and its relation to other cognitive abilities and every-day behaviors. For instance, a theoretical question relevant to our own research is whether a domain-general factor of attention control is the basis of individual differences in higher-order cognitive abilities such as working memory capacity and fluid intelligence (Shipstead et al., 2016).

Unfortunately, such theoretical questions may need to be put on hold until a more fundamental issue that has long plagued individual differences research on executive attention is resolved (e.g., Friedman & Miyake, 2004). The issue is that many executive functioning measures have poor psychometric properties, and it is only recently that this problem has garnered wide-spread recognition (see Draheim, Mashburn, Martin, & Engle, 2019; Hedge, Powell, & Sumner, 2018; Paap & Sawi, 2016; Rouder & Haaf, 2019; Rouder, Kumar, & Haaf, 2019). This has become a contentious topic in the field of attention control. Problems with existing measures calls into question the capability of researchers to measure individual differences in attentional abilities, which in turn casts doubt on conclusions made from previous research and

This article was published Online First July 23, 2020.

Christopher Draheim, Jason S. Tsukahara, Jessie D. Martin, Cody A. Mashburn, and Randall W. Engle, School of Psychology, Georgia Institute of Technology.

This work was supported by the Office of Naval Research Grant N00014-12-1-1011 to Randall W. Engle.

The data analyzed in this study were part of a larger data collection sample, and we reported data that included some of the same tasks in different publications. The following link has a summary of the larger data collection procedure and a reference list of all publications to come out of this data collection sample with information on which tasks were used for each publication (https://osf.io/s5kxb). We reported data on the relationship between sensory discrimination, fluid intelligence, working memory capacity, and attention control in a separate article (Tsukahara, Harrison, Draheim, Martin, & Engle, 2020). We also reported data on visual arrays tasks in a separate article (Martin et al., 2019) extensively discussing the nature of the tasks and what constructs they measure. We reported data from the attention tasks and a follow-up session Martin et al. (2019) that focuses on predictive validity of the attention measures. The broader issue of measurement concerns in individual differences research, with some discussion of the issues as they pertain to attention control, was discussed in another separate article (Draheim, Mashburn, & Engle, 2018). In addition, data and ideas from the present study were disseminated in various conference presentations (Draheim, Martin, Tsukahara, Mashburn, & Engle, 2018; Draheim, Mashburn, & Engle, 2018; Draheim et al., 2019; Engle, 2017).

Correspondence concerning this article should be addressed to Christopher Draheim, School of Psychology, Georgia Institute of Technology, 648 Cherry Street NW, Atlanta, GA 30313. E-mail: cdraheim3@gatech.edu

renders it difficult to confidently test theoretical hypotheses related to the domain-generality of attention.

Current Views on the Measurement and Assessment of Attention Control

Substantive Versus Methodological Debate

There are competing views for why robust experimental tasks of attention control do not produce reliable and valid individual differences. One view focuses on the methodological issues associated with the tasks. Another argument is that the lack of shared variance among most attention measures is more substantive and reflects the domain-specific nature of what these tasks are thought to measure; that is, there is no domain-general attention control ability that can explain performance across different attention-related tasks (Hedge, Powell, Bompas, & Sumner, 2020; Rey-Mermet, Gade, & Oberauer, 2018; Rouder & Haaf, 2019).

On the methodological side, there are known psychometric issues with using well-established experimental paradigms for correlational research. Studies of individual differences in cognition tend to rely on tasks which come directly out of the experimental tradition. The advantage of this is that experimental tasks are based upon a large body of experimental findings and ostensibly do a good job of isolating specific cognitive processes of interest (see Rey-Mermet, Gade, Souza, von Bastian, & Oberauer, 2019; Rouder & Haaf, 2019; but see Miller & Ulrich, 2013 and Verhaeghen & De Meersman, 1998 for challenges to this view). However, tasks which produce robust and reliable effects at the experimental level are often unreliable and correlate weakly with other theoretically related tasks at the individual differences level. Within the domain of executive functioning, this issue is particularly salient in the measurement of attention control (e.g., Friedman & Miyake, 2004; Goodhew & Edwards, 2019; Hedge et al., 2018; Magnusdottir et al., 2019; Rey-Mermet et al., 2018, 2019; Paap & Sawi, 2016; Rouder & Haaf, 2019; Rouder et al., 2019), though certainly this is a concern in behavioral science as a whole (see Draheim, Mashburn, et al., 2019; Rouder et al., 2019 made a similar observation). The end result is that the integration of experimental tasks for correlational purposes is not as straightforward as one might assume (see Fisher, Medaglia, & Jeronimus, 2018; Goodhew & Edwards, 2019; Hedge et al., 2018; and Logie, Della Sala, Laiacona, Chalmers, & Wynn, 1996).

One likely explanation of why experimental tasks are poorly suited to individual differences research is because experimental tasks are designed to minimize between-subjects variability and be most sensitive to differences between experimental conditions (Hedge et al., 2018). This is ideal for group comparisons, but the presence of between-subjects variability is critical to correlational studies. We argue that difference scores contribute significantly to this problem (for a review, see Draheim, Mashburn, et al., 2019). Difference scores are ideal for experimental research as a way to control for baseline performance. The problem for individual differences research is that a difference score is necessarily less reliable than its components, resulting in less between-subjects variance and therefore attenuated correlations.[1] Difference scores that are low in reliability can counterintuitively result in an increase in power in analysis of variance-based tests so long as the component scores are reliable (Chiou & Spreng, 1996; Overall & Woodward, 1975). In other words, maximizing power in testing for group differences and maximizing power (reliability) in assessing individual differences are at odds with each other. Difference scores are therefore useful and perhaps even necessary for isolating specific cognitive processes within an individual and finding effects in experimental research, but they are not as useful or necessary for isolating cognitive processes between individuals and often demonstrate low reliability and validity in correlational research. Many researchers therefore caution against the use of difference scores, particularly in correlational pursuits (Cronbach & Furby, 1970; Draheim, Hicks, & Engle, 2016; Draheim, Mashburn, et al., 2019; Edwards, 2001; Goodhew & Edwards, 2019; Hedge et al., 2018; Hughes, Linck, Bowles, Koeth, & Bunting, 2014; Lord, 1956, 1963; Paap & Sawi, 2016).

A related methodological explanation for the poor psychometric properties of attention tasks is that there is too much measurement error due to trial-level variability in performance within individuals (Rouder & Haaf, 2019; Rouder et al., 2019). Specifically, Rouder and colleagues argue that the ratio of trial-level noise to true individual variation is too large in part due to small effect sizes in the difference scores of tasks such as Stroop and Simon. Rouder et al. defended the contrast (difference score) approach to isolating variance of interest but also concluded that it might not be possible to increase the ratio of trial noise to true score variance in existing inhibition tasks without substantially increasing the number of trials administered (possibly requiring over 1,000 trials!). Therefore, it appears that the sheer number of trials necessary to provide favorable conditions for the use of difference scores in correlational research effectively precludes their use, at least when using tasks with small effect sizes such as the standard Stroop and flanker tasks.

Yet another methodological argument is that performance in the commonly used Stroop, flanker, and Simon tasks reflect little variance associated with the conflict-resolution processes they are believed to measure, and are overly contaminated with construct-irrelevant variance such as processing speed and speed–accuracy trade-offs (e.g., Hedge et al., 2020). We discuss this idea more in the following section.

On the other hand, there are also researchers who argue that attention measures do not intercorrelate due to theoretically meaningful reasons. Such researchers may recognize various methodological shortcomings with existing attention measures but argue that the issues are overblown and/or that null or conflicting results are due primarily to the nature of the constructs in question. Perhaps attention measures do not correlate because attention control mechanisms are not domain-general and therefore variance in attention control tasks is highly task-specific.[2] For instance, Paap and Sawi (2014) wrote in regard to the weak correlations among attention tasks: ". . . measures of inhibitory control derived from the flanker, Simon, and Stroop tasks have often shown low levels of convergent validity making it highly likely that the

[1] This is assuming that the components are not perfectly reliable or completely independent, which is generally a safe assumption in behavioral research (see Cronbach & Furby, 1970; Lord, 1963).

[2] What we refer to as *attention control* is sometimes called *inhibition* by other researchers. For the present purposes, there is no meaningful difference between these two terms.

conflict resolution mechanisms employed are task specific rather than recruiting general-purpose inhibitory control." Similarly, although Rouder and Haaf (2019) acknowledged several methodological problems with the assessment of attention control/inhibition, their failure to find a meaningful relationship between Stroop and flanker performance using a hierarchical regression approach designed to account for some of these problems led them to conclude that inhibitory mechanisms in Stroop and flanker are task-specific.

Rey-Mermet et al. (2018) reached a similar conclusion after testing a wide array of commonly used attention control measures in a correlational study with a diverse sample of younger and older adults. Performance on their tasks was scored primarily as RT difference scores, and four of their 11 attention measures were standard Stroop and flanker tasks. Only 25% of their task-level correlations were statistically significant, only 11% of the correlations exceeded $r = .20$, and some correlations were in the opposite direction such that better performance on one attention task was associated with worse performance in another. When all tasks were loaded onto a single attention factor, half had factor loadings below .20, which is far from acceptable even if one adopts a liberal tolerance (e.g., Comrey & Lee, 1992; Matsunaga, 2010; Stevens, 2012). They interpreted their results as demonstrating the lack of a unified inhibition factor. We would instead argue that their findings only reinforce the poor psychometric properties of these tasks. For example, only two of their difference score measures had a corrected split-half internal consistency at or above .75, with an average of .64.[3] In a more recent study, Rey-Mermet et al. (2019) administered accuracy-based tasks using a novel calibration procedure but still failed to find a unitary attention control factor. We provide a commentary of their study toward the end of the discussion section with potential reasons for their null results.

Our view is that the weak correlations across various attention control tasks is methodological in nature. We argued in Draheim, Mashburn, et al. (2019) that the use of RT, especially difference scores in RT, is one of the primary reasons for issues with tasks such as Stroop and flanker.[4] In addition to the poor psychometric properties of difference scores, RT measures in general are sensitive to speed–accuracy interactions which can manifest in a number of ways (for reviews, see Draheim, Mashburn, et al., 2019; Heitz, 2014), and these are problematic for differential and developmental studies. No studies in support of the domain-specific view of attention control have appropriately addressed the methodological issues of using RT difference scores, or difference scores more generally. Further, this view relies on the acceptance of the null hypothesis, that nonsignificant correlations between related tasks reflects the absence of common variance. This is particularly problematic in individual differences research in which so many factors can unknowingly contribute to null findings.

Reaction Time Is Contaminated With Speed–Accuracy Trade-Offs and Processing Speed

Many RT tasks are scored without any consideration or control for accuracy. This is an issue because participants who have quick and error-prone responding will appear better on the task than participants who have more deliberate and slower responding but with fewer errors solely due to differences in emphasis on speed and accuracy as opposed to differences in ability. Difference scores are used to account for baseline processing (task fluency and processing speed) but a difference in RT does not account for differences in speed-accuracy tendencies. Another consideration is that researchers have questioned the assumption that RT difference scores wholly control for outside sources of variance, namely those related to general processing speed or ability. Miller and Ulrich (2013) argued that the cognitive processes underlying RT is not as simple as often believed and that interpretations predicated on correlations between RTs are often faulty. The crux of their argument is that RTs are impure, and that observed correlations involving RT measures have multiple influences, some of which map to processes of interest whereas others do not. It has also been established that RT differences do not properly account for processing speed (e.g., Rey-Mermet et al., 2019; Verhaeghen & De Meersman, 1998). Rey-Mermet et al. (2019) pointed out there is a confound between processing speed and cognitive ability in tasks in which performance is scored as a RT difference score. Ergo, even if RT differences in attention control tasks (such as Stroop and flanker) revealed large individual differences that correlated to performance on other tasks, it would not be clear whether these individual differences reflected differences in attentional abilities or merely differences in general processing speed. To that end, Hedge et al. (2020) showed that, although most attention control tasks weakly correlate with one another, processing speed and speed–accuracy trade-offs are shared across these tasks. Specifically, they used diffusion modeling to reanalyze several data sets which used a combination of Stroop, flanker, and Simon tasks and found that despite low correlations of performance data (error and RT costs), nondecision time and variability, drift rate (associated with processing speed), and boundary separation (associated with response cautiousness and therefore speed–accuracy trade-offs) parameters were each strongly correlated within and across tasks. Diffusion parameters associated with conflict resolution (which these tasks are believed to measure) were not correlated within or across tasks. They concluded that these conflict tasks are contaminated with sources of construct irrelevant variance and therefore that a hypothetical positive relationship among them would be only minimally informative due to this contamination. These findings separately support the Rey-Mermet et al. (2019) concern that RT interference effects in these conflict tasks may be contaminated with processing speed, and our argument that speed–accuracy trade-offs and interactions are a concern with executive functioning tasks (Draheim et al., 2016; Draheim, Mashburn, et al., 2019). Hedge et al. also performed a simulation on these tasks and showed that even if the diffusion parameters associated with conflict were very strongly correlated across tasks, this would not necessarily result in strong correlations in the behavioral data. Hedge et al.'s findings call into question the widely held belief that

[3] Although rules of thumb and guidelines are by no means absolute, .80 is often quoted as the threshold for acceptable reliability in basic research (e.g., Nunnally, 1964).

[4] Accuracy-based difference scores are no better than reaction time difference scores (e.g., Hughes et al., 2014; Rey-Mermet et al., 2019), but they are less commonly used in measuring executive functioning, and so it is the reliance upon reaction time difference scores which is our focus here.

difference scores in conflict tasks such as Stroop, flanker, and Simon are process pure and therefore theoretically meaningful.

Reaction Time Irrelevant Tasks May Be a Solution to the Measurement Problem

Although Hedge et al. (2020) found that contamination from processing speed and speed–accuracy trade-offs was present in both RT and accuracy difference scores, it should be noted that the Stroop, flanker, and Simon tasks are generally administered in such a way that both RT and accuracy are important to performance. That is, respondents must balance the two and decide how much emphasis to give to both while performing the task. We argue that it is easier to account for extraneous sources of individual differences variance such as processing speed and speed-accuracy interactions when using accuracy-based measures. Whereas it is difficult to assess executive functioning with a RT task in which accuracy is completely irrelevant, it is much more straightforward to design an accuracy-based executive functioning task in which RT is rendered irrelevant. A participant can respond as quickly or as slowly as they want, provided they are not responding too quickly to make motor errors or too slowly such that the memory representation of the trial is lost. If our reasoning is correct, this is quite a wide time window to respond without appreciably affecting performance. Wickelgren (1977) made a very similar point in reference to ways to account for speed–accuracy trade-offs in cognitive tasks:

> Although the basic fact of speed–accuracy trade-off may make it inadequate to look at RT alone as the dependent variable, it does not invalidate looking at asymptotic accuracy as a dependent variable without reference to RT. The fact that speed–accuracy trade-off functions approach an asymptotic level of accuracy at long RTs means that so long as the response time is sufficiently long to ensure that one is operating at or near the asymptote, enormous differences in RT will be associated with negligible differences in asymptotic accuracy. Thus, there is little opportunity for contamination of asymptotic accuracy by differences in response time, while there is considerable opportunity for contamination of RT by differences in accuracy. (p. 82)

This observation by Wickelgren (1977) implies something rather intuitive. Because speed–accuracy trade-offs are a concern in cognitive tasks, one way to account for them is to render RT irrelevant to the task such that respondents do not have to emphasize one over the other. In other words, one way to account for speed–accuracy trade-offs is to use tasks which apply relatively little demand to respond quickly and/or otherwise make quick responding unnecessary or impossible. If this is accomplished, then the variance of interest should all be in accuracy.

The antisaccade task is an example of one measure in which RT is irrelevant, and therefore variance of interest is entirely in accuracy. On this task, participants must make a saccade in the opposite direction (left or right) of a cued stimulus and identify a target stimulus that is quickly masked. If they do not make the antisaccade in time, then they will miss the target stimulus. They have as long as they want to respond, however they either saw the target stimulus or they did not and therefore their response time is irrelevant. We argue that construct-irrelevant variance such as that from processing speed and speed–accuracy trade-offs are therefore minimal in this task. This could explain Rey-Mermet et al.'s (2019) observations that antisaccade tasks tend to "dominate" latent factors of attention. Our lab similarly finds that loadings for the antisaccade task tend to be much higher than the .20 to .40 observed with traditional Stroop and flanker tasks, which fall below acceptable levels and indicte that our so-called attention factor is comprised mostly of variance from a single task—the antisaccade (e.g., Shipstead, Harrison, & Engle, 2015). Relatedly, we find the antisaccade to be highly reliable (around .90) and the Stroop and flanker tasks to be much less so (.60 – .70), and that correlations involving the antisaccade task and other cognitive measures are much stronger (as strong as $r = .50$), whereas correlations involving Stroop and flanker are numerically half at best, and our Stroop and flanker tasks share only 3% to 4% common variance at the task level (around $r = .10$).

Given that recent attempts to improve the measurement of attention control (e.g., Rey-Mermet et al., 2019; Rouder et al., 2019) have not been successful (but see Paap, Anders-Jefferson, Mikulinsky, Masuda, & Mason, 2019), we believe that the most practical approach to improving the measurement of attention control is to create tasks that are RT irrelevant, accuracy-based, and avoid the use of difference scores. This research endeavor is somewhat exploratory in the sense of task-development, however, is hypothesis-driven because it provides a test of the methodological versus substantive debate as to why common variance is not shared across attention control tasks. Further, we can also ask more specific theoretical questions such as whether a unitary factor of attention control can fully account for the working memory capacity-fluid intelligence relationship.

Goals of the Present Study

The present study was an effort to test whether developing new and modified attention tasks would lead to improvements in the ability to measure individual differences in attention control. Given the issues of RT and difference scores, we reasoned that pushing performance variance away from RT and into accuracy would be a viable approach for improving the psychometric properties of attention tasks. Accuracy-based and adaptive procedures can circumvent the noted issues with difference scores and allow for better control of speed-accuracy interactions and processing speed. Due to the aforementioned concerns regarding contamination from processing speed and other construct-irrelevant sources of variance (Hedge et al., 2020), we also included measures of processing speed to assess discriminant validity.

Formally stated, the primary question of the present study was whether this toolbox approach to developing new attention control tasks would result in improvements in the measurement of attention control. The evaluative criteria used to assess this question were as follows: (1) task reliability, (2) intercorrelations with other attention control tasks, (3) latent factor coherence, (4) the relationship to working memory capacity and fluid intelligence, and, (5) the mediation of the working memory capacity/fluid intelligence relationship. Reliability is the first criterion since a measure cannot be valid without being reliable, and the unreliability of many attention tasks (such as the RT difference score-based Stroop and flanker) has been called into question by us and other researchers. Second, an attention control measure should correlate strongly to other attention control measures if indeed, as we argue, attention control is a

broad, unified, and domain-general construct. Third, another goal for improving the measurement of attention control, if attention control is a unified domain-general ability, should be to find tasks which have a more balanced loading onto a latent factor. Finally, according to our theoretical view of executive attention, attention control measures should strongly relate to working memory capacity and fluid intelligence and attention control should fully mediate the relationship between working memory capacity and fluid intelligence. Predicting real-world behavior is an important criterion as well, however we address this topic in a separate article (Martin, Mashburn, et al., 2019) which focuses on whether attention control can predict incremental variance in multitasking ability above and beyond the Armed Services Vocational Aptitude Battery (ASVAB; see Roberts et al., 2000).

The Toolbox of Attention Control Measures

We compared several different types of attention control measures. The first group consisted of existing measures: the RT difference score-based color Stroop, the RT difference score-based arrow flanker, error rates in a nonadaptive accuracy-based antisaccade task, a RT-based psychomotor vigilance task, and a capacity-based visual arrays (change detection) task requiring selection/filtering of some elements of the display. Stroop, flanker, and antisaccade are tasks our lab has traditionally used to assess attention control. Although the standard Stroop and flanker are known to be poor individual differences measures, they were included for comparative reasons. The psychomotor vigilance task is a RT measure, but it may be an example of one well suited to correlational analysis because it is not scored as a difference and, presumably, speed-accuracy interactions would not be present (there are no right or wrong answers once the counter begins ticking up, only quicker and slower responses). So, just as tasks such as the complex span and antisaccade effectively control for speed (response time is irrelevant), the psychomotor vigilance task effectively controls for accuracy (an inaccurate response is not possible). As for the inclusion of visual arrays as an attention measure, see the following section.

The second group of tasks were threshold-based modified Stroop and flanker tasks that involved an adaptive procedure. We reasoned that modifying these tasks to be adaptive threshold-based and not at all reliant upon differences in RT may be a method of improving the reliability and validity. Further, these tasks were modified in such a way that interference effects due to the intermixing of congruent and incongruent trials should play a role in performance, although said interference was not directly measured with the dependent variable.

The third group of tasks were completely novel tasks. One, sustained attention-to-cue, was designed to be an accuracy analog to the psychomotor vigilance task and with a distractor similar to that of the antisaccade. We reasoned that the requirement for both sustained attention and resisting a distractor should enhance the amount of controlled attention required to perform the task well. The adaptive visual cue was designed to be similar to antisaccade but adaptive and with a larger number of nontarget stimuli. These tasks are available for download at http://englelab.gatech.edu/attentioncontroltasks.

Selective Visual Arrays as a Measure of Attention Control

Including visual arrays as a measure of attention control is controversial and requires explanation. Visual arrays, along with similar change detection measures, is generally conceptualized as a measure of visual working memory capacity and indeed change detection tasks in general do provide a useful estimate of the amount of information stored in working memory (e.g., Conway, Kane, & Engle, 2003; Luck & Vogel, 1997). However, we argue that there is sufficient evidence that individual differences in visual arrays performance is more so a reflection of attentional abilities than storage or capacity. The question of what visual arrays measures is an important one and it would not be feasible to fully detail all lines of evidence for our position here, so instead this issue is discussed in full detail in a separate article which has been submitted for publication and is available to read on PsyArXiv (see Martin, Tsukahara, et al., 2019).

To highlight, several studies by Vogel and colleagues show the importance of attention in variation in visual arrays performance. For example, Fukuda, Woodman, and Vogel (2015) found that capacity scores in visual arrays were significantly smaller for a supracapacity set size (eight) than for a near-capacity set size (four), which supported the attentional-control account of individual differences in visual arrays capacity scores over the memory storage account. Vogel, McCollough, and Machizawa (2005) found that contralateral delay activity (which increases with the amount of stored information and therefore can be used as an indicator of whether irrelevant distractors are processed, thereby unnecessarily consuming memory capacity) for low and high working memory capacity individuals was roughly the same for two items and four items arrays in which distractors were not present, but that there was a strong difference in contralateral delay activity in high versus low spans when two items were presented along with two distractors. Therefore, lower ability individuals perform worse in visual arrays tasks due to the inability to filter distractor items. Finally, Fukuda and Vogel (2011) showed that high and low spans differed in how long it took them to recover from attentional capture when performing a visual arrays task, thus resulting in individual differences in task performance.

Here, a distinction between different types of visual arrays tasks is critical. One type of visual arrays is a more straightforward change detection/primary memory task in which the respondent indicates whether something changed from one display to another (nonselective visual arrays). Another type of visual arrays task places a filtering/selection demand on the respondent such that the respondent must ignore or filter out part of the initial array (a cue presented very briefly prior to the to-be-remembered array indicates which part of the array the respondent should attend to; selective visual arrays). Although Vogel and colleagues have found attentional-related individual differences in nonselective visual arrays, particularly when supracapacity set sizes are used, we would consider performance in these tasks to be more so driven by working memory capacity than attention control. However, we do consider selective visual arrays to primarily reflect attentional processes and, therefore, to be a measure of attention control. Performance on the visual arrays is assessed using a capacity score—a reflection of how many items the respondent could retain in the array. Capacity scores on versions of the visual arrays

without the filtering demand (nonselective visual arrays) are around 3 to 3.5 (similar to tasks of working memory capacity), whereas capacity scores on versions with the filtering demand (selective visual arrays) are about half that (e.g., Shipstead et al., 2015), indicating that the filtering demand alters performance in a meaningful way. Further, data from our lab across four separate studies show that selective visual arrays shares more variance with attention control tasks (namely antisaccade) than with working memory tasks (see Martin, Tsukahara, et al., 2019; Shipstead et al., 2015). At the latent level, selective visual arrays accounts for substantial and unique variance in attention control above and beyond working memory capacity tasks and nonselective visual arrays. Additionally, visual arrays with the filtering demand loads onto the same factor as antisaccade and other attention tasks but not with measures of working memory capacity. In sum, although selective and nonselective visual arrays tasks share substantial variance and load together latently, they can be separated out as measures which primarily reflect attention control and working memory capacity, respectively.

Method

The study at large consisted of approximately 30 cognitive tasks administered over four 2-hr long sessions. Participants were also invited back for subsequent sessions; a 2-hr session to assess test–retest reliability, and two 2-hr long sessions for a separate project from which we present selected data later in the present article. In addition to performance data and demographic information, we also collected eye tracking, pupillometry, and mind wandering data from a number of tasks, which are not reported here.

The data analyzed in this study were part of a larger data collection sample and we reported data that included some of the same tasks in different publications. The following link has a summary of the larger data collection procedure and a reference list of all publications to come out of this data collection sample with information on which tasks were used for each publication (https://osf.io/s5kxb). In addition, we list the various articles and conference presentations in which some ideas and/or data from the present study were disseminated.

Participants

Participants could schedule sessions within designated slots during lab hours and with the restriction that they could not complete more than one session on any single day. Participants received $130 in compensation, which was distributed as $25 for the first session with each subsequent session worth $5 more than the previous. Georgia Institute of Technology (Georgia Tech) psychology students could choose to receive either payment or research participation credit for each session (two credits per session). After completing this study, most participants were invited to return for an additional two sessions for a separate study. After completing that two-session study, they were also invited back for one final session to obtain test–retest reliability estimates on the attention control measures. Participants who frequently rescheduled, failed to show up to appointments, or were otherwise problematic in terms of behavior or performance data (e.g., lack of effort, falling asleep during tasks, failure to understand instructions, or near-floor performance on multiple tasks) were not invited back to the lab for these additional sessions. The experiment room was set up to run up to five participants at a time with partitions between the kiosks such that participants could not see each other. Participants wore headphones while performing each task. An experimenter sat behind them at all times, occasionally observing their performance to ensure participants were following instructions. The experimenter also took notes regarding participants' behavior, alertness, and apparent motivation, answered instruction-related questions, and started the run files for each task. Participants were not told directly that they would be observed but they were aware of the experimenter's presence. Undergraduate research assistants served as the experimenter the majority of the time, with graduate students or post docs filling in as needed. At least one senior lab member (graduate student or post doc) had to be present in the lab to supervise data collection. Most tasks had built-in rest periods that appeared after a set number of trials for each task, designed to occur approximately every 10 min. Participants could advance the rest screen at their convenience to continue performing the task. Participants were asked to avoid getting out of their chair while in the middle of a task if possible and were encouraged to take short breaks between tasks when needed. The Georgia Tech Institutional Review Board approved the protocol for this study.

Data collection took place at the Georgia Tech School of Psychology. A total of 482 participants enrolled in the study, with 403 finishing all four sessions (60.0% female; 39.5% male; 0.5% other). Just under 94% of participants reported being a former or current college student (46% Georgia Tech; 22% Georgia State University; 25% other institution). To be eligible for the study, participants had to report being a native (learned at age 5 or earlier) and fluent English speaker age 18 – 35 years with normal or corrected-to-normal vision, no history of seizure, and having never participated in a previous study in our lab. Participants were recruited via the Georgia Tech participant pool, flyers around the Georgia Tech campus, flyers off-campus, Reddit and Craigslist postings, various advertising efforts (Facebook, Creative Loafing, campus publications), and from in-person recruiting in the downtown and midtown Atlanta areas.

Tasks

Working memory capacity. We measured working memory capacity using three complex span tasks: advanced operation span, advanced symmetry span, and advanced rotation span. The complex span tasks consist of alternating memory storage and processing subtasks (Unsworth, Heitz, Schrock, & Engle, 2005). The advanced versions of the tasks included larger set-sizes of memory items (Draheim, Harrison, et al., 2018). For each task, to help ensure that participants were attending to the processing task and not rehearsing the to-be-remembered information during this time, participants were allotted up to their individual mean RT + 2.5 standard deviations from a block of practice trials consisting of the processing task only. Additionally, participants were asked to maintain a minimum level of performance on the processing tasks (85% accuracy) and their cumulative percent correct was shown to them throughout the task in the upper-left corner of the screen. These tasks can be downloaded at http://englelab.gatech.edu/taskdownloads.html.

Operation span (Kane et al., 2004; Turner & Engle, 1989). This task required participants to remember a series of letters presented in alternation with simple math equations which they were required to solve (see Figure 1). On each trial, participants first solved a simple math equation (e.g., [2 × 2] + 1 = 5) or not (e.g., [3 × 4] − 3 = 8) followed by the presentation of a single letter. This alternation repeated until a variable set-size of letters to-be-remembered had been presented, at which point participants attempted to recall the letters they had seen in serial order. There were a total of 14 trials (two blocks of seven trials), set-sizes ranged from three to eight, and each set-size occurred once in each of two blocks. The dependent variable was the partial span score, which is the total number of letters recalled in proper serial position (Conway et al., 2005).[5]

Symmetry span (Unsworth, Redick, Heitz, Broadway, & Engle, 2009). On each trial, participants were first presented with a 16 × 16 matrix of black and white squares and were required to decide whether the pattern was symmetric on the vertical midline. Followed by the symmetry judgment, a 4 × 4 matrix of squares with one square highlighted in red was displayed. Participants had to remember the location of the red square within the 4 × 4 matrix (Figure 1). This alternation continued until a variable set-size of spatial locations had been presented. Participants then attempted to recall the locations of the red square in correct serial order. There were a total of 12 trials (two blocks of six trials), set-sizes ranged from two to seven, and each set-size occurred twice (once in each block). The dependent variable was the partial span score, which is the total number of square locations recalled in proper serial position.

Rotation span (Kane et al., 2004). This task required participants to remember a series of directional arrows of varying size in alternation with a mental rotation task in which they had to decide whether or not a letter was mirror reversed. On each trial, participants first solved a mental rotation problem followed by the presentation of a single arrow with a specific direction (i.e., eight possible directions, the four cardinal and four ordinal directions) and specific size (small or large). Both the direction and size of the arrow were the to-be-remembered features (see Figure 1). This alternation continued until a variable set-size of arrows had been presented, at which point participants attempted to recall the set of arrows in their correct serial position. There were a total of 12 trials (two blocks of six trials), set-sizes ranged from two to seven, and each set-size occurred twice (once in each block). The dependent variable was the partial span score, which is the total number of arrows recalled in proper serial position.

Fluid intelligence.

Raven's advanced progressive matrices—Odd problems (Kane et al., 2004; Raven & Court, 1998). Participants were presented abstract shapes in a 3 × 3 matrix. The bottom-right shape was absent, and the participant had to select which item, from the eight answer options, best completed the overall pattern by clicking that option on the screen. Participants had 10 min to complete 18 problems. The number of correct responses was the dependent variable.

Number series (Unsworth et al., 2009; Thurstone, 1938). Participants were presented a sequence of numbers and needed to identify the response option that was the next logical number in the sequence by clicking the correct number from five total response options. Participants had 5 min to answer 15 problems. The number of correct responses was the dependent variable.

Letter sets (Ekstrom, French, Harman, & Dermen, 1976). On each problem, the participant was presented five sets of letters, each containing four letters that follow a particular rule. Instructions were to find the rule that applied to four of the five letter sets, and then indicate the set that violates the rule by clicking that set on the screen. Participants had 10 min to complete 30 problems. The number of correct responses was the dependent variable.

Nonadaptive attention control tasks.

Antisaccade (Hutchison, 2007; Kane, Bleckley, Conway, & Engle, 2001). Participants saw a central fixation cross lasting a random amount of time between 2,000 ms to 3,000 ms followed by an alerting tone for 300 ms. After the alerting tone, an asterisk appeared for 300 ms at 12.3° visual angle to the left or the right of the central fixation followed immediately by a target *Q* or an *O* for 100 ms on the opposite side of the screen from the asterisk. The location of the asterisk and target letter were both masked for 500 ms by ##. The participant's goal was to ignore the asterisk and instead look away to the other side of the screen to catch the target *Q* or *O*. Participants had as much time as needed to respond to which letter appeared by pressing the associated key on the keyboard. After responding, accuracy feedback was displayed for 500 ms, followed by a blank intertrial interval of 1,000 ms. Participants completed 72 trials, and the dependent variable was the number of correctly identified target letters.

Selective visual arrays (Luck & Vogel, 1997; Shipstead, Lindsey, Marshall, & Engle, 2014). Participants saw an array of blue and red rectangles differing in orientation. Prior to each trial, the participant was cued to attend to either the red or blue rectangles by a 300 ms flash of *RED* or *BLUE*. Next, the array was presented for 250 ms after a delay of 900 ms, and then the array was presented again with only the target rectangles (either red or blue). One of the target rectangles contained a white dot and changed orientation from the original array on 50% of the trials. The participant was asked whether or not the rectangle cued with the white dot had changed orientation from the initial presentation. Each array contained either five or seven rectangles per color (10 and 14 total), and 40 trials were presented for each array size for a total of 80 trials. The dependent variable was a capacity score (*k*) calculated using single probe correction (see Cowan et al., 2005; Shipstead et al., 2014). This calculation is $N \times$ (Hits + Correction Rejections − 1), where N is the set-size for that array. This calculation results in two separate *k* scores, one for set size five and one for set size seven, and the final dependent variable was the average *k* for these two set-sizes. See Figure 2.

Psychomotor vigilance task. In this task, participants were presented with a row of five zeros in black font at the center of the screen. After a variable wait time (equally distributed among 2, 4,

[5] Due to an error in programming of the advanced operation span task, trials in which the set-size was supposed to be nine only displayed a set-size of eight. This resulted in the set-size of eight occurring twice as often as intended, a total of four trials compared with two trials for each other set-size.

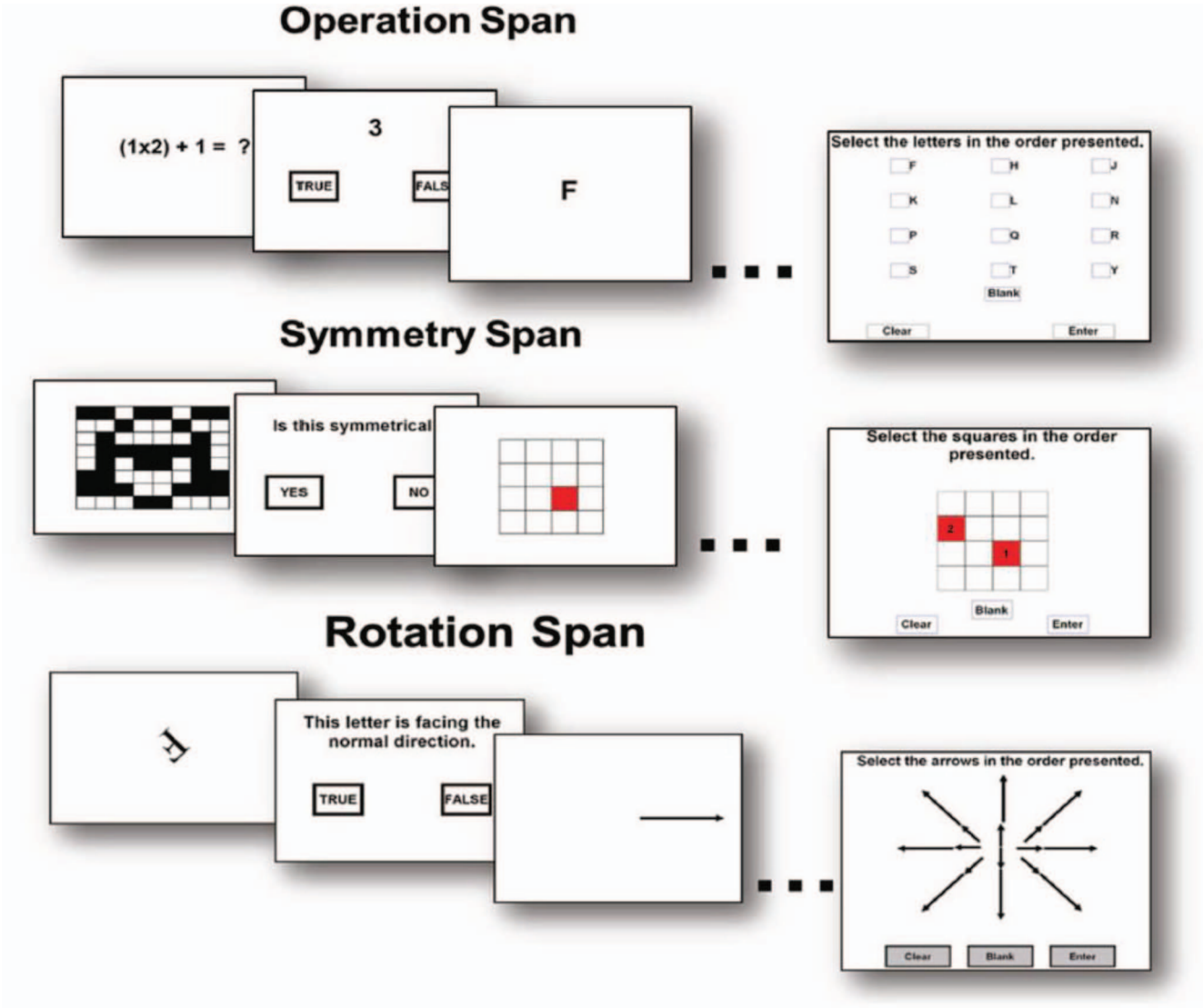

Figure 1. Demonstration of the complex span tasks. In each task, participants respond true/false or yes/no to a processing (distractor) task prior to the presentation of each to-be-remembered stimulus. After a variable amount of presentations (depending on the set-size for that trial), a recall screen appears asking the participant to recall the to-be-remember stimulus in order of presentation. The dependent variable is the partial span score, which is the total number of items recalled in the correct position. See the online article for the color version of this figure.

8, and 12 s) the zeros began to count up by 17 every 17 ms. The participants were instructed to press the spacebar as quickly as possible once the numbers started counting up. After they pressed the spacebar, their RT was left on the screen for 1,500 ms to provide feedback. Eighty trials were administered. The dependent variable was the RTQ20, which is the average RT on each participant's slowest 20% trials (Dinges & Powell, 1985; Unsworth & Robison, 2016).

Arrow flanker (RT flanker; Eriksen & Eriksen, 1974; Nieuwenhuis et al., 2006; Stoffels & van der Molen, 1988). Participants were presented with a target arrow in the center of the screen pointing left or right along with two flanking arrows on both sides. The flanking arrows were either all pointing in the same direction as the central target (congruent trial: e.g., ← ← ← ← ←) or all in the opposite direction (incongruent trial: e.g., ← ← → ← ←).[6] The participant was asked to indicate the direction of the central arrow by pressing the *Z* (left) or "/" (right) key. These keys had the words *LEFT* and *RIGHT* taped onto them to assist with response mapping. A total of 144 trials were administered: 96 congruent and 48 incongruent, with a randomized 400 ms to 700 ms intertrial interval (ITI). The dependent variable was the flanker interference effect—the RT cost of the incongruent trials calcu-

[6] Note that the arrow flanker task often has three trial types: congruent, incongruent, and a neutral type in which dashes flank the central arrow. Similarly, in the Stroop task a neutral trial type is often present in which the word is not a color. Here we used only incongruent and congruent trials.

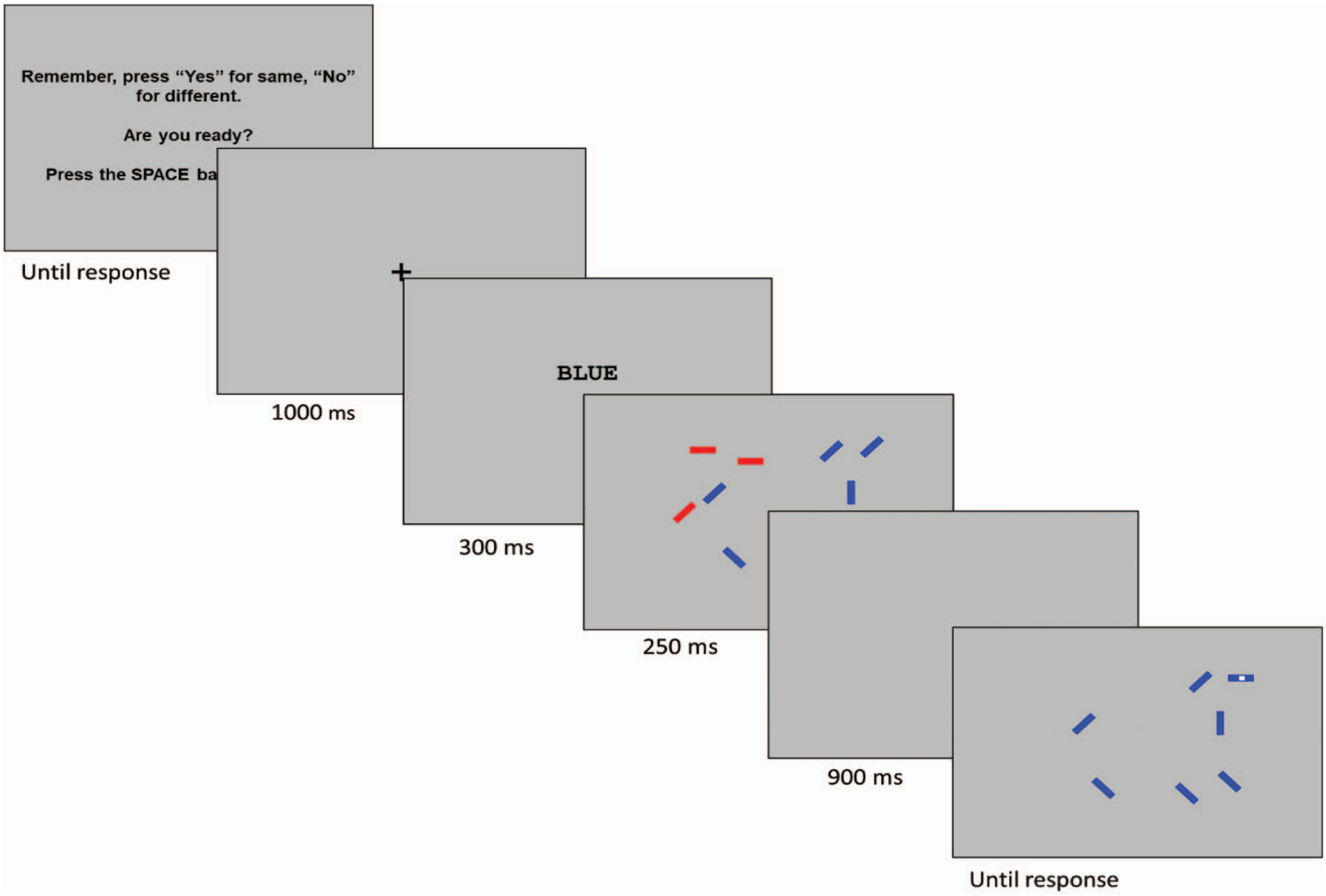

Figure 2. Trial sequence for the selective visual array task. This example shows a trial of set-size seven. Here, the participant needed to indicate whether the blue rectangle in the upper right of the array (indicated by a white dot in its center) had changed orientation from the previous display. In this trial, it did, and so the participant would respond by pressing the key which corresponded to the "yes" response. See the online article for the color version of this figure.

lated by subtracting each participant's mean RT on congruent trials from their mean RT on incongruent trials, excluding inaccurate trials.

Color Stroop (RT Stroop; Stroop, 1935). Participants were shown the word "red", "green", or "blue" in red, green, or blue font. The words were either congruent with the color (e.g., the word *red* in red print), or incongruent with the color (e.g., the word *red* in blue print). The participant's task was to indicate the color in which the word was printed by pressing the *1*, *2*, or *3* key on the number pad. To assist with response mapping, the keys had a piece of paper with the corresponding color taped onto them. A total of 144 trials were administered, 96 congruent and 48 incongruent, with a randomized 400 ms to 700 ms ITI and a 5,000-ms response deadline. The dependent variable was the Stroop interference effect—the RT cost of the incongruent trials calculated by subtracting each participant's mean RT on congruent trials from their mean RT on incongruent trials, excluding all inaccurate trials.

Sustained attention-to-cue task (SACT). This was a novel task designed primarily as an accuracy-based version of the psychomotor vigilance task. In this task, participants needed to sustain their attention on a visual circle cue presented at random locations on the screen and ultimately identify a target letter presented briefly at the center of the cue. The stimuli were presented against a gray background. Each trial started with a central black fixation. On half of the trials, the fixation was presented for 2 s and for the other half the fixation was presented for 3 s. After the fixation, following a 300-ms tone, a large white circle cue was presented in a randomly determined location on either the left or right side of the screen. To orient the participant to the circle cue, the large circle began to immediately shrink in size until it reached a fixed size. Once the cue reached the fixed size, after a variable wait time (equally distributed among 2 s, 4 s, 8 s, and 12 s), a white distracting asterisk appeared at the center of the screen. The asterisk blinked on and off in 100-ms intervals for a total duration of 300 ms (on for 100 ms, off for 100 ms, on for 100 ms). Then, a 3 × 3 array of letters was displayed at the center of the cue. The letters in the array consisted of B, D, P, and R. The central letter was the target letter and was presented in a dark gray font. The nontarget letters were presented in black font with each letter occurring twice in the array and the target letter occurring three times. After 125 ms the central letter was masked with a # for 1,000 ms. Only the central target letter was masked. After the

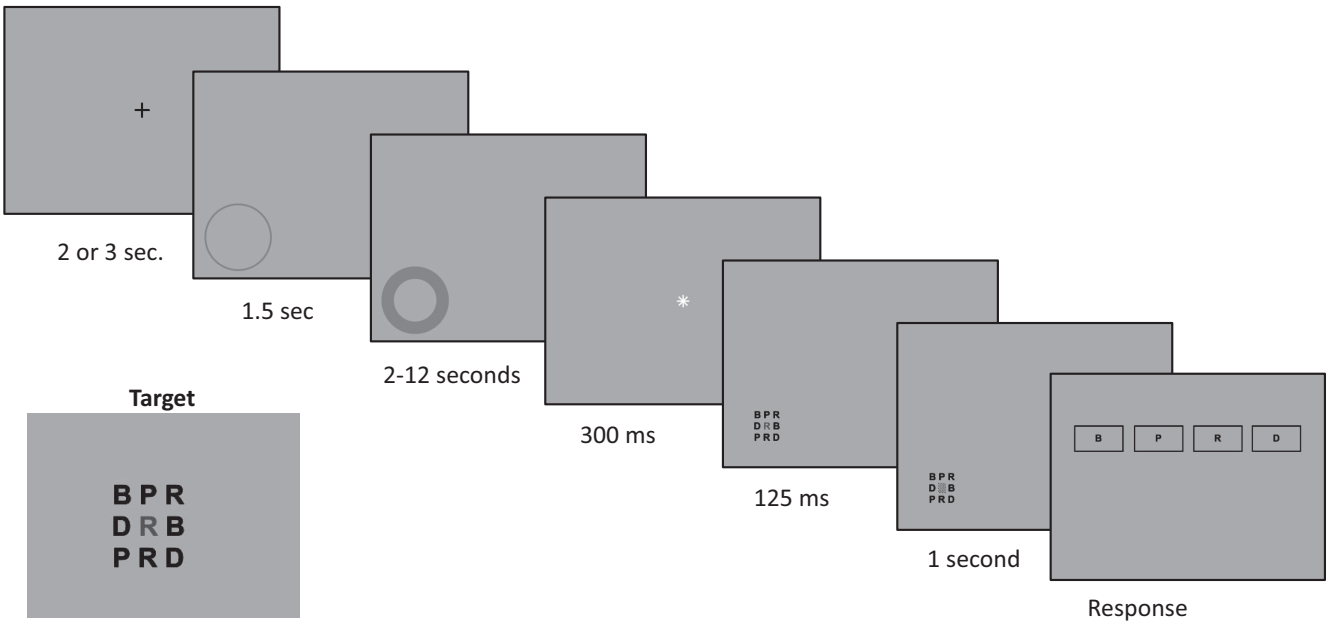

Figure 3. Trial sequence for the sustained attention-to-cue task. Participants saw a fixation for 2 s or 3 s followed by a circle cue indicating the future location of the target letter array. This circle shrunk for 1.5 s and then remained for either 2 s, 4 s, 8 s, or 12 s. After this wait interval, a distracting asterisk then appeared outside of the circle for 300 ms. Then, the target 3 × 3 letter array (see enlarged view at bottom left corner) appeared for 125 ms and was then masked for 1,000 ms. Participants indicated which of four possible letters (i.e., B, D, P, or R) appeared in the center.

mask, the response options were displayed in boxes horizontally across the upper half of the screen. The participant used the mouse to select whether the target was a B, D, P, or R. Feedback was given during the practice trials but not the experimental trials. Sixty-four trials were administered. Accuracy rate was the dependent variable.[7] See Figure 3.

Adaptive attention control tasks.

Flanker presentation rate (Flanker PR). This task was a modified version of the arrow flanker and used an adaptive procedure to estimate the participant's threshold. Eighteen blocks of 18 trials each (total 324 trials) were administered. The duration that the stimuli appeared on the screen (presentation rate) either decreased (shorter presentation time) if the participant was accurate on at least 15 trials within each block or increased (longer presentation time) if their accuracy rate was below that. The first block had a presentation time of 235 ms. For the first six blocks, the presentation time for the following block shortened in duration by 45 ms or increased in duration by 135 ms, again depending on whether the participant was accurate on at least 15 of the 18 trials. For subsequent blocks, the presentation time either quickened by 15 ms or slowed by 45 ms.[8] Participants could take as much time to respond as needed. Each block had 12 congruent and six incongruent trials in random order with a randomized 400 ms to 700 ms interstimulus interval (ISI). Congruent and incongruent trials were treated equally in determining if the presentation rate increased or decreased for the next block (i.e., participants needed to be correct on at least 15 of 18 total trials independent of whether these trials were incongruent or congruent). The dependent variable was the presentation time after the final block of trials (i.e., what the presentation rate would have been on a hypothetical 19th block).

Flanker deadline (Flanker DL). This task was a modified version of the arrow flanker and used an adaptive procedure to estimate the participant's threshold. Eighteen blocks of 18 trials each (total 324 trials) were administered. Each trial had a response deadline that limited how long the participant had to respond before they heard a loud beep and forfeited the opportunity to respond on that trial. This deadline either decreased (less time to respond) if the participant was accurate on at least 15 trials within each block or increased (more time to respond) if their accuracy rate was below that. The first block had a response deadline of 1,050 ms. For the first six blocks, the response deadline either

[7] Although the sustained attention-to-cue task shares an asterisk distractor with the antisaccade, we believe that it does not simply constitute a more complex version of antisaccade as suggested by an anonymous reviewer. First, as the name implies, this task involves a wait period in which the respondent presumably must sustain their attention at the cued location prior to stimulus onset. Second, a critical difference in the distractors in the antisaccade vs. this task is that the respondent has knowledge of the location of the target stimulus) prior to the presentation of the asterisk distractor in the sustained attention-to-cue task. In antisaccade, the respondent does not know where the target will be until the distractor is presented and must use the distractor to look to the opposite side of the screen. Therefore, in antisaccade the goal is to use the distractor as a cue to look in the opposite direction of said cue. In sustained attention-to-cue, it is to sustain attention over time on a particular precued location while avoiding distraction.

[8] If the presentation time were to be set below 10 ms, it was set to exactly 10 ms instead.

decreased by 90 ms or increased by 270 ms for the next block, depending on whether the participant was accurate on at least 15 of the 18 trials. For subsequent blocks, the response deadline either decreased by 30 ms or increased by 90 ms.[9] The stimuli remained on the screen up until the response deadline. Each block had 12 congruent and 6 incongruent trials in random order with a randomized 400 ms to 700 ms ISI. The response deadline was the same for incongruent and congruent trials. Further, congruent and incongruent trials were treated equally in determining whether the response deadline increased or decreased on the following block (i.e., participants needed to be correct and respond before the response deadline on at least 15 of 18 total trials for the response deadline to decrease, independent of whether these trials were incongruent or congruent). The dependent variable was the response deadline after the final block of trials (i.e., what the deadline would have been on a hypothetical 19th block).

Stroop deadline (Stroop DL). This task was a modified version of the Stroop and used an adaptive procedure to estimate the participant's threshold. Eighteen blocks of 18 trials each (total 324 trials) were administered. Each trial had a response deadline that limited how long the participant had to respond before they heard a loud beep and forfeited the opportunity to respond on that trial. This response deadline either decreased (less time to respond) if the participant was accurate on at least 15 trials within each block or increased (more time to respond) if their accuracy rate was below that. The first block had a response deadline of 1,230 ms. For the first six blocks, the response deadline either decreased by 90 ms or increased by 270 ms for the next block, again depending on whether the participant was accurate on at least 15 of the 18 trials. For subsequent blocks, the response deadline either decreased by 30 ms or increased by 90 ms. The stimuli remained on the screen until the response deadline.[10] Each block had 12 congruent and six incongruent trials in random order with a randomized 400 ms to 700 ms ISI. The response deadline was the same for incongruent and congruent trials. Further, congruent and incongruent trials were treated equally in determining whether or not the deadline increased or decreased for the next block (i.e., participants needed to be correct and respond before the deadline on at least 15 of 18 total trials for the response deadline to decrease, independent of whether these trials were incongruent or congruent). The dependent variable was the response deadline after the final block (i.e., what the deadline would have been on a hypothetical 19th block).

Adaptive visual cue task (AVC). This was a new task. A valid cue was presented where a target line was to be located within a larger search array that was presented for a brief duration. Participants needed to identify whether the target line was of vertical, horizontal, or diagonal orientation. The search array was composed of 124 distractor lines of vertical, horizontal, or diagonal orientation. The orientation and location of each distractor line was randomly determined for each trial. The duration of the cue adaptively changed on a trial-by-trial basis. The trial sequence was as follows. Following a 300 ms tone, a white asterisk appeared on either the far left or far right side of the screen. The asterisk blinked on and off in 100-ms intervals for a total duration of 300 ms (on for 100 ms, off for 100 ms, on for 100 ms). A blue circle cue was then displayed on the opposite side of the screen for a variable duration (determined by the adaptive procedure). Then, the search array of the target and distractor lines was presented for 150 ms, after which each stimulus on the display was masked for 300 ms. The participant then used the mouse to select whether the target line was of vertical, horizontal, or diagonal orientation. Feedback was given during the practice trials but not the experimental trials. See Figure 4.

We used a weighted-up-down adaptive procedure with 64 trials to estimate a difference threshold to converge on a certain level of performance (Kaernbach, 1991). If participants made a correct response, the duration of the cue decreased on the next trial (less time to orient to the visual cue). If the participant made an error, the duration of the cue increased on the next trial (more time to orient to the visual cue). The ratio to decrease/increase the duration was 1:2. This weighted-up-down method with a step ratio of 1:2 converges around a threshold value of 66% (Kaernbach, 1991). To estimate this threshold value, we employed a method of averaging the difference value for the last four reversals (Hairston & Maldjian, 2009). A reversal is each instance of which the accuracy on the current trial was different from the accuracy on the previous trial. The critical dependent measure was the average difference value of the last four reversals.

Data Preparation

Data processing steps were implemented at multiple stages to prepare the data for statistical analysis. To outline, data were first trimmed at the task level depending on task-specific characteristics and we removed participants' scores on a task if they showed poor performance indicating they did not understand the instructions or were not truly performing the task (the criteria for each task is described in the following paragraphs). Next, separately for each task we removed outlier scores (± 3.5 *SD*). These two steps can result in a participant having missing data on some tasks and not others. Missing data also occurred for a variety of other reasons including lost data files, experimenter error, or tasks crashing, which accounted for less than 1% of total scores. Finally, for the working memory capacity and fluid intelligence tasks, if a participant had missing data on more than one task for either construct, then they were completely removed from any further analyses. After removing scores based on these criteria, a total of seven participants were completely removed from data analysis for a total sample size of 396. Out of the 396, no single task had more than 5% scores missing.

For the working memory tasks, if a participant's mean accuracy on the processing portion of the task was 3.5 standard deviations below the mean, then their score on the task was removed.

For the standard Stroop and flanker tasks, we removed trials that had extreme RT values. The purpose of this was to remove trials with too short of RTs to accurately reflect task processing. If RTs were shorter than 200 ms, then the trial was removed. If the participants mean accuracy on congruent or on incongruent trials was 3.5 standard deviations below the mean across all participants, then their score on the task was removed.

For the psychomotor vigilance task, we removed trials that had extreme RT values. Again, the purpose of this was to remove trials

[9] If the response deadline were to be set below 150 ms, it was set to exactly 150 ms instead.

[10] If the response deadline were to be set below 150 ms, it was set to exactly 150 ms instead.

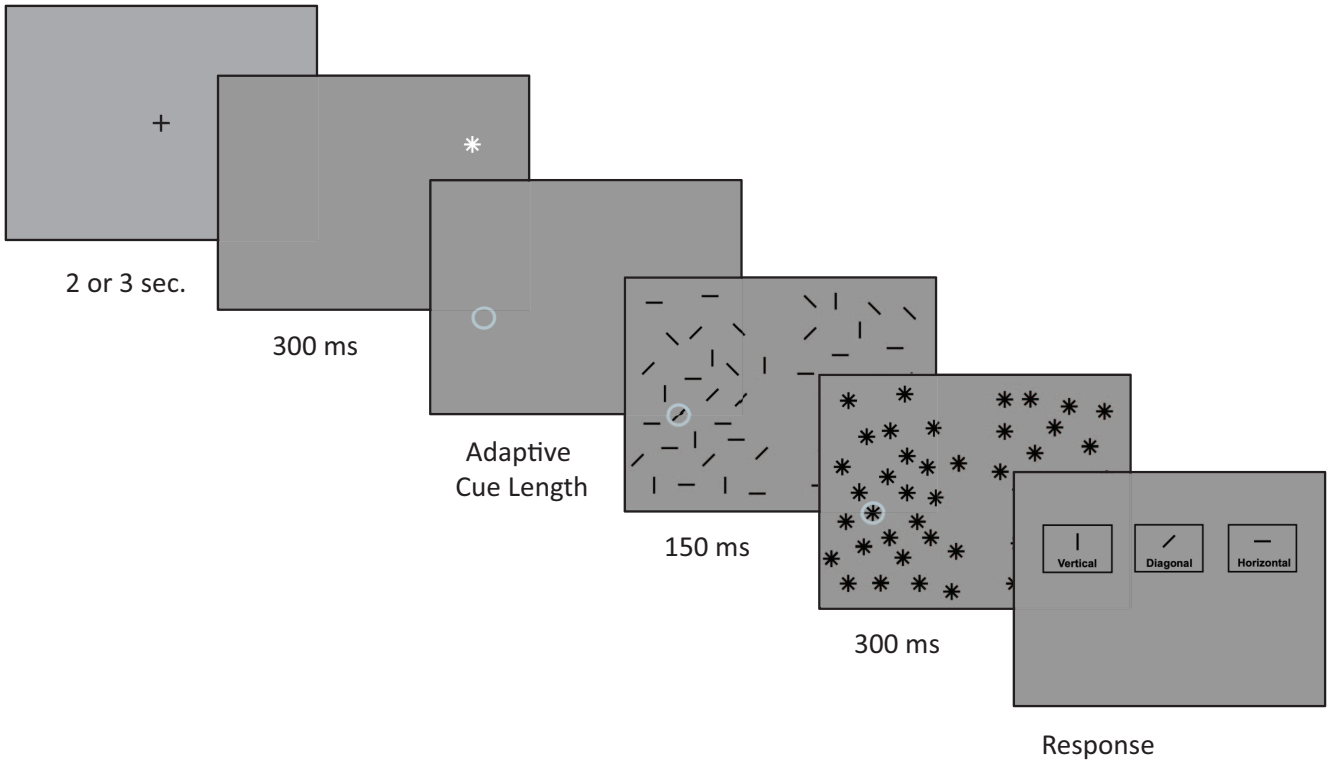

Figure 4. Trial sequence for the adaptive visual cue task. A central fixation appears for 2 s or 3 s, followed by the appearance of an asterisk at a random location on either the left or right side of the screen for 300 ms. Following the asterisk, a cue appears on the opposite side of the screen for a variable amount of time. Then the visual search array appears, with a target line at the center of the cue in one of three orientations, vertical, diagonal, or horizontal. After 150 ms the visual search array is masked for 300 ms. The participant must select which orientation the target had. The length of the cue is determined on a trial-by-trial basis using an adaptive up-down procedure. After a trial with a correct response, the length of the cue is shorter on the following trial. After a trial with an incorrect response, the length of the cue is longer on the following trial.

with too short or too long of RTs to accurately reflect task processing. If RTs were shorter than 200 ms or longer than 10 s, the trial was removed.

For the visual arrays, the adaptive Stroop and flanker tasks, and adaptive visual cue tasks, if a participant's mean accuracy on the task was 3.5 standard deviations below the mean their score on that task was removed.

Data Analysis

We used five criteria to evaluate the strength of each attention control task (1) reliability of each task, (2) intercorrelations with other attention control tasks, (3) latent factor loadings, and (4) external validity to working memory capacity and fluid intelligence. The fifth criterion, mediation of the working memory capacity/fluid intelligence relationship, was not used to assess tasks individually but rather whether the new and modified tasks were improvements in aggregate.

For reliability, we assessed internal consistency using split-half reliability (even/odd trials) corrected using the Spearman-Brown prophecy formula. Internal consistency was not calculated for adaptive and threshold-based measures, as this did not seem appropriate given the nature of these tasks. We assessed test–retest reliability for all attention tasks by bringing back a subset of participants for an additional 2-hr session. Attention measures were administered in the same relative order to one another in the retest session as the initial testing sessions. All attention measures could not fit within one 2-hr session, so we created two separate conditions for test–retest reliability. Antisaccade, visual arrays, and sustained attention-to-cue were included in both conditions whereas the other tasks were only included in one, resulting in a large discrepancy in terms of test-retest sample sizes across the tasks. In hindsight, it was perhaps a mistake not to include all new and modified measures in both test-retest conditions as it limited our ability to assess reliability in the adaptive tasks. That is, because internal consistency could not be calculated for the adaptive tasks (namely the modified Stroop and flanker tasks), the only estimate of reliability we have for these tasks is based on test–retest reliability for 58 to 66 participants with, on average, 194 days between initial administration of the tasks and retest. Readers are therefore urged to exercise caution in interpreting the reported test–retest reliabilities, particularly for the tasks with relatively few participants.

The factor loadings are based on an exploratory structural equation modeling approach. The rationale for this was that in addition

to including a model with all attention tasks loaded onto a single attention control factor, we wanted to evaluate what combination of three tasks formed the strongest attention control factor. The use of at least three tasks on a latent factor is standard as it helps avoid issues with underdefined models and it is not always practical for researchers to include more than three indicators for a latent variable given time and resource constraints of behavioral research. For the 10 attention control tasks, there are 120 possible combinations of three tasks. Therefore, we ran 120 structural equation models with an attention control factor predicting working memory capacity and fluid intelligence.

For external validity, we used structural equation modeling to test whether attention control fully mediated the relationship between working memory capacity and fluid intelligence. We present several models in comparison to our typical attention control factor with the antisaccade, flanker, and Stroop (RT difference score versions).

For all structural equation models, solid paths represent statistically significant paths ($\alpha = .05$) and dotted lines represent statistically nonsignificant paths. The chi square values, degrees of freedom, and chi square significance are reported. The chi-square assesses overall fit and discrepancy between the sample and generalized population wide fitted covariance matrices (i.e., how far apart the covariance matrix implied by the model is when compared to the observed covariance matrix). Although a nonsignificant chi-square value is preferred, indicating that the model is not statistically different from a general population model, it is very sensitive to sample size and degrees of freedom, and the chi-square test is based on assumptions which are rarely met in practice (see Schermelleh-Engel, Moosbrugger, & Müller, 2003). As such, the chi-square value alone is not sufficient to accept or reject a model and is often instead used as one source of evidence for model fit (Anderson & Gerbing, 1988; Schermelleh-Engel et al., 2003). Models must therefore be considered in holistic terms based on multiple fit indices, namely the confirmatory fit indices (CFI) and the root mean square error of approximation (RMSEA). The CFI compares model fit to a null model. Models with a CFI $>$ .90 are considered an acceptable fit, and with a CFI of .95 or higher considered to be a good fit (Hu & Bentler, 1999). The RMSEA is a parsimony adjusted fit index. Models with an RMSEA $<$ .08 are considered to be an acceptable fit, with an RMSEA of .05 or lower considered to be a good fit (Browne & Cudeck, 1993; Kenny, 2015) although Hu and Bentler recommended .06 as the cutoff instead. Any models that are not considered as an acceptable or good fit are thereby considered to have poor model fit, indicating that the model does not sufficiently recreate the observed covariance matrix.

A nuance to assessing model fit in structural equation modeling is the balance between model parsimony/generality and a well-fitting model. Most software used to perform structural equation modeling reports some combination of the Wald test, Lagrange multiplier test, and likelihood ratio test which all assess how specifying the model differently would alter model fit and list changes in descending order of the impact of each change. Using these tests to adjust parameters and constraints to achieve a better fitting model is controversial and can lead to a practice known as overfitting. An overfit model will have excellent fit indices but also a large number of post hoc parameters and constraints that are specific to the particular data set and will likely not generalize to other data sets or populations. In addition, it can often be difficult to theoretically justify certain post hoc modifications, such as why two seemingly poorly related indicators should have correlated error terms for any reason other than it improved model fit. It is therefore our practice to report the most parsimonious models (e.g., ones without a number of cross-loaded indicators and correlated error terms) and to engage in post hoc respecification of the model only when absolutely necessary or justified.

Finally, in some analyses we report whether a path value from one model is significantly different than a path value in another. This is done using the model comparisons approach (Loehlin, 1987) in which nested models are tested against one another. If the model with fewer constraints (more degrees of freedom) is not significantly different from the nested (more constrained) model in terms of fit, then the model with more degrees of freedom is preferred due to parsimony. However, if the nested (more constrained) model is statistically a better fit to the data than the less constrained, the nested model is preferred as it does a better job of recovering the covariance matrix of the observed data.

All analyses were conducted in R statistical software (R Core Team, 2018). The R package *lavaan* (Rosseel, 2012) was used for all structural equation analyses, treatment of missing values was set to full-information maximum likelihood.

Results

Descriptive statistics for all relevant tasks, including time of administration for the attention measures, are provided in Table 1.

Criterion 1—Reliability

Internal consistency and test-retest reliabilities for the attention control measures are presented in Table 2. The psychomotor vigilance, sustained attention-to-cue, and antisaccade tasks were highly reliable in terms of internal consistency (.92–.95). The visual arrays had lower internal consistency (.75) but ranked higher than the RT difference score versions of the flanker (.74) and Stroop (.69). In terms of test–retest reliability, the antisaccade, visual arrays, response deadline Stroop, and sustained attention-to-cue ranked highest (.63–.73). While the psychomotor vigilance task showed high internal consistency, the test–retest reliability was .55. The adaptive response deadline flanker had poor test–retest reliability (.54) though this was higher than the RT Stroop (.46) and RT flanker (.23). The adaptive visual cue had very poor test–retest reliability (.39) as did the adaptive presentation rate flanker task (.32).

The second administration of these tasks occurred, on average, over six months after initial administration ($M = 194$ days, $SD = 131$; min = 6, max = 419). Because of this and the large variance in number of days elapsed between test and retest, we were concerned that number of days elapsed could have affected test–retest reliability. A moderation analysis with number of days as the interaction term showed no statistically significant moderation of number of days elapsed on test–retest reliability for any of the measures (accounting for only around 1% of the variance), indicating that number of days did not interact with the stability of the test–retest estimate. This was surprising given that the test–retest reliabilities were overall very low for most tasks, but results may have been different if the retest session had occurred within a week or so of initial administration and if the sample size was larger.

Table 1
Descriptive Statistics of Dependent Variables From Each Task

Task	*M*	*SD*	Skew	Kurtosis	Administration time (in min)
Raven	10.26	3.22	−.41	−.27	
Letter sets	17.09	4.34	−.32	−.31	
Number series	9.81	3.09	−.40	−.48	
Operation span	55.76	15.58	−.73	.08	
Symmetry span	27.90	10.35	−.06	−.40	
Rotation span	24.49	9.30	−.07	−.29	
Visual arrays	1.86	1.17	.19	−.59	13.0
Antisaccade	.79	.15	−.90	.03	9.8
Psychomotor vigilance	889.44	754.68	2.77	7.72	26.8
Sustained attention-to-cue	.70	.19	−.76	−.06	24.8
Adaptive visual cue	58.83	42.92	4.91	38.39	12.0
Stroop DL	1013.01	345.33	1.97	5.56	10.1
Flanker DL	674.61	212.90	2.33	6.97	10.5
Flanker PR	101.35	93.4	1.73	2.80	14.5
RT flanker effect	80.69	43.18	1.08	1.98	9.1
RT Stroop effect	131.62	85.53	.73	.83	5.7

Note. Administration time refers to the 95th percentile of time it took participants to complete the task. That is, 95% of participants finished at or before that time. Stroop DL = modified Stroop with adaptive response deadline; Flanker DL = modified flanker with adaptive response deadline; Flanker PR = modified flanker with adaptive presentation rate; RT = reaction time.

Criterion 2—Intercorrelations Among Tasks

The full correlation matrix for the attention control measures is presented in Appendix A (Table A1) and Table 3 shows a summarized version of the average intercorrelations among these measures. The RT flanker and Stroop tasks (assessed as difference scores) had the lowest correlations with other attention measures (r = .15 and .11, respectively), which is consistent with previous research. On the other hand, the antisaccade and sustained attention-to-cue tasks had the strongest average intercorrelations (r = .35 and .32, respectively), with all other tasks in the r = .18–.27 range. Because we consider the antisaccade to be a hallmark measure of attention control, we also present in Table 3 the correlation of each task to the antisaccade. All correlations involving antisaccade were statistically significant ($p < .05$). Performance on the selection visual arrays and sustained attention-to-cue tasks correlated quite strongly to antisaccade (r = .45 and .40, respectively). The psychomotor vigilance task, adaptive visual cue task, and response deadline versions of the Stroop and flanker also had relatively strong correlations to the antisaccade (rs = .31 to

Table 2
Internal Consistency (IC) and Test–Retest Reliabilities for Attention Tasks

Task	IC initial (*n*)	IC retest (*n*)	Test–retest	Retest without outliers
PVT	.95 (388)	.92 (63)	.55	.55
SACT	.93 (383)	.95 (126)	.52	.63
Antisaccade	.92 (390)	.92 (126)	.71	.73
Visual arrays	.75 (397)	.71 (125)	.67	.69
RT flanker	.74 (385)	.70 (71)	.23	.23
RT Stroop	.69 (386)	.62 (67)	.46	.46
Stroop DL	(390)	(66)	.55	.67
Flanker PR	(386)	(58)	.32	.32
Flanker DL	(382)	(66)	.31	.54
AVC	(386)	(65)	.31	.39

Note. IC was calculated using an even–odd split procedure and was corrected using the Spearman-Brown prophecy formula. IC Initial is the internal consistency for the first administration of the task. IC Retest is the internal consistency for the second administration (retest). Test–retest is the correlation between performance on the initial and retest administrations. Retest without outliers is this correlation after removing outliers for each task. Outliers are defined as participants with a *z*-score difference >3 between their rank order in the first and second administration. Average days elapsed between Administration 1 and Administration 2 was 194 (*SD* = 131). We do not present reliability estimates for working memory capacity or fluid intelligence tasks, as the psychometric properties of these tasks have been established (e.g., Conway et al., 2005), but all these measures had internal consistencies in the .80 range, except for the rotation span partial score (.77). PVT = psychomotor vigilance task; SACT = sustained attention-to-cue; RT flanker = reaction time flanker effect; RT Stroop = reaction time Stroop effect; Stroop DL = modified Stroop with adaptive deadline; Flanker PR = modified flanker with adaptive presentation rate; Flanker DL = modified flanker with adaptive response deadline; AVC = adaptive visual cue.

Table 3
Intercorrelations Among the Attention Control Measures

Task	Average correlation	Correlation to antisaccade
Antisaccade	.35	
SACT	.32	.40
Visual arrays	.27	.45
PVT	.26	.35
Flanker DL	.25	.34
Flanker PR	.25	.25
Stroop DL	.24	.31
AVC	.18	.33
RT flanker effect	.15	.16
RT Stroop effect	.11	.19

Note. N = 382–390. Correlations involving one task in which lower scores indicate better performance (e.g., reaction time) were multiplied by −1 such that a positive correlation indicates that individuals who performed better on one task also performed better on the other. All tasks statistically significantly correlated to antisaccade. RT Stroop and RT flanker are reaction time difference scores (Stroop and flanker effect, respectively). SACT = sustained attention to cue; PVT = psychomotor vigilance task; Flanker DL = modified flanker with adaptive response deadline; Flanker PR = modified flanker with adaptive presentation rate; Stroop DL = modified Stroop with adaptive response deadline; AVC = adaptive visual cue; RT = reaction time.

.35). The adaptive presentation rate flanker had a moderate correlation to antisaccade (r = .25), and the RT Stroop and flanker effects had the weakest correlations to antisaccade (r = .16 and .19, respectively).

Criterion 3—Latent Coherence

We ran an exploratory factor analysis using principal axis factoring with a varimax rotation and not specifying a set number of factors (see Table 4). Visual inspection of the scree plot suggested that two factors were sufficient: one factor for fluid intelligence and working memory measures and another for the attention measures. Five factors had eigenvalues above 1. Whereas using the eigenvalue above one criterion is likely not the best method of choosing numbers of factors (e.g., Osborne, Costello, & Kellow, 2008), in this case it was more informative than forcing two or even three factors. We ran additional models either specifying a certain number of factors and/or with oblique, rather than orthogonal, rotations. The models converged on the following conclusions: (1) The working memory capacity tasks and fluid intelligence tasks formed their own separate factors, which, in the orthogonal models, accounted for the most variance; (2) Antisaccade, sustained attention-to-cue, psychomotor vigilance task, adaptive visual cue task, and visual arrays loaded together on a separate factor that accounted for the third most variance among the factors; (3) all three flanker measures loaded onto their own factor, driven by the adaptive response deadline flanker task (though any flanker task could be removed from the analysis and the other two still loaded together and separately from the other tasks); (4) performance on the RT Stroop task loaded onto a separate factor with the adaptive response deadline Stroop task (which also equally loaded onto the attention control factor); (5) performance on the visual arrays loaded more strongly with working memory capacity and fluid intelligence than other attention tasks did, generally with a similar magnitude as the visual arrays' loading with the other attention measures; and (6) if we forced fewer factors, the working memory capacity and fluid intelligence measures would form a single g factor but the three flanker tasks would still load together and separate from the other attention tasks. As a validity check, we also added sensory discrimination measures to the exploratory model (data not reported here), which resulted in an additional factor comprising just these three discrimination tasks having an eigenvalue above 1.

We tested the attention control measures in all possible combinations of a three-indicator factor (120 unique models, each task

Table 4
Exploratory Factor Analysis—Rotated Factor Loadings

Task	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Raven	*.26*	**.55**	.13	−.12	−.15
Letter sets	.20	**.67**	.05	−.09	−.09
Number series	.18	**.73**	.07	−.05	−.01
Operation span	**.53**	*.30*	−.04	−.06	.06
Symmetry span	**.86**	.19	.07	−.16	−.08
Rotation span	**.62**	*.31*	.19	−.17	−.06
Visual arrays	*.30*	*.27*	**.37**	−.03	−.20
Antisaccade	.20	*.25*	**.50**	−.15	−.18
Psychomotor vigilance	.03	−.02	**.60**	.12	−.08
Sustained attention-to-cue	.12	−.02	**.55**	−.13	−.01
Adaptive visual cue	−.03	.04	**.38**	.00	−.05
Stroop DL	.14	.12	**.23**	.16	**.24**
Flanker DL	−.05	.16	.09	**.90**	−.01
Flanker PR	.19	−.01	*.25*	**.40**	.08
RT flanker effect	−.06	−.05	.05	**.36**	.09
RT Stroop effect	.01	.12	.02	.14	**.86**

Note. Extraction done via principal axis factoring with varimax rotation. The strongest loading for each task is presented in boldface type. Loadings at or above .25 are shown in italic type when the loading is not the strongest for that particular task. For ease of interpretation, some loadings were multiplied by −1 such that positive loadings reflect better performance for that task was positively related to the factor. Stroop DL = modified Stroop with adaptive response deadline; Flanker DL = modified flanker with adaptive response deadline; Flanker PR = modified flanker with adaptive presentation rate; RT = reaction time.

appeared in 36 of them) and when predicting both fluid intelligence and working memory capacity. Table 5 shows a summary of the average loading for each task across these 120 models. The antisaccade and visual arrays had very strong average loadings (.75 and .69, respectively), whereas the adaptive visual cue task, RT flanker, and RT Stroop had, on average, very low loadings (.28 to .31). The other tasks were in the middle, ranging from .40 to .59.

As it is not practical to show all 120 tri-indicator attention control factors, we instead present a model with all attention tasks loaded onto a single factor to summarize the latent results (see Figure 5). The rank-ordering of the factor loadings for this model is generally consistent with Table 5, and the loadings are quite similar to the average loading for each task across all 120 combinations of tri-indicator attention control factors (also Table 5). Note that this model is only for illustrative purposes to show the relative loadings of the attention tasks on the attention control construct. It is not a particularly good model to assess attention control because a number of poor indicators are present and it is likely that model fit would be improved by loading a subset of the attention tasks onto subfactors of attention. As a result, model fit is relatively poor (e.g., CFI = .90) and the mediation of attention control on the working memory capacity/fluid intelligence relationship is worse than when fewer indicators are used instead (see Figure 6).

Criterion 4—Relationship to Working Memory Capacity and Fluid Intelligence

The final criterion for individual task improvement is attention control's relationship to fluid intelligence and working memory capacity. According to our theory of working memory capacity as executive attention, the strong link between working memory capacity and fluid intelligence is primarily due to individual differences in attention control (Engle, 2002; Engle & Kane, 2004). According to this theory, attention control underlies all higher-order and goal-directed behavior, and hence it should strongly associate with, potentially even fully mediate, the relationship between any two executive functions such as working memory capacity and fluid intelligence. However, potentially due to aforementioned problems with assessing individual differences in attention control, a full mediation of attention control on the working memory capacity/fluid intelligence relationship has not been observed in previous studies. It has been argued that abilities such as secondary memory and memory updating are required, along with attention control, to fully explain the relationship between working memory capacity and fluid intelligence, and yet even when these other constructs are tested there is still variance unaccounted for (e.g., Unsworth & Spillers, 2010). We would argue that attention control alone should be sufficient to explain the variance in the working memory capacity and fluid intelligence relationship, provided that stronger attention control measures are used than the typical ones which are scored using RT and difference scores.

Table 6 shows the first-order correlations between all the attention control tasks with composite z scores of working memory capacity and fluid intelligence. Note that the working memory capacity and fluid intelligence composites correlated at $r = .55$ and their latent scores correlated at $r = .67$ (45% of their reliable variance was shared at the latent level). Table 6 shows that the visual arrays and antisaccade tasks quite strongly correlated with working memory capacity and fluid intelligence (rs = .41 to .46), and that the adaptive response deadline flanker had the next strongest correlation ($r = .29$ with working memory capacity, $r = .33$ with fluid intelligence). The only attention measure that did not correlate significantly with working memory capacity was the RT Stroop ($r = .10$, $p = .054$), although this correlation was not statistically different than the psychomotor vigilance's ($r = .12$) or adaptive visual cue task's ($r = .10$) correlation to working memory capacity. All measures significantly correlated with fluid intelligence and had correlations at or above $r = .18$. The flanker presentation rate, Stroop response deadline, sustained attention-to-cue task, and RT flanker correlated similarly with working memory capacity ($r = .18$ to .23) and fluid intelligence ($r = .16$ to .23) composite scores.

Table 5
Average Factor Loadings for Each Attention Task

Task	Average loading (*SD*)
Antisaccade	.76 (.08)
Visual arrays	.69 (.06)
Flanker DL	.59 (.12)
Sustained attention-to-cue	.51 (.11)
Flanker PR	.48 (.08)
Stroop DL	.46 (.08)
Psychomotor vigilance	.40 (.09)
Adaptive visual cue	.31 (.07)
RT flanker	.31 (.08)
RT Stroop	.28 (.07)

Note. The average loading was calculated by taking the average absolute value of each task's loading across all 120 possible tri-task factors in which each task appeared 36 times. Flanker DL = modified flanker with adaptive response deadline; Flanker PR = modified flanker with adaptive presentation rate; Stroop DL = modified Stroop with adaptive response deadline; RT flanker = reaction time flanker effect; RT Stroop = reaction time Stroop effect.

Criterion 5—Mediation of the Working Memory Capacity/Fluid Intelligence Relationship

Another way to test the relationship between attention control and working memory capacity/fluid intelligence is through latent mediation models with attention control mediating the working memory capacity/fluid intelligence relationship. One issue with this approach is that latent analyses require multiple indicators per factor, and so comparing one task to another is less straightforward. Nonetheless, these mediation models can provide a better understanding of how the new and modified attention tasks performed in aggregate.

We used results reported thus far to inform our decision of which models to test, and these four models are presented in Figure 6. The first model we tested (see Figure 6a) was the baseline model consisting of three attention measures used in our previous studies: the antisaccade, RT Stroop, and RT flanker. In this model the RT flanker and Stroop tasks had poor loadings (.28 and .30, respectively), whereas the antisaccade had a strong loading (.68). This indicates that the antisaccade dominated the factor (a common finding as we discussed previously and was also reported by Rey-Mermet et al., 2018) such that the attention control factor is mostly composed of variance from this single task. Also, of note is

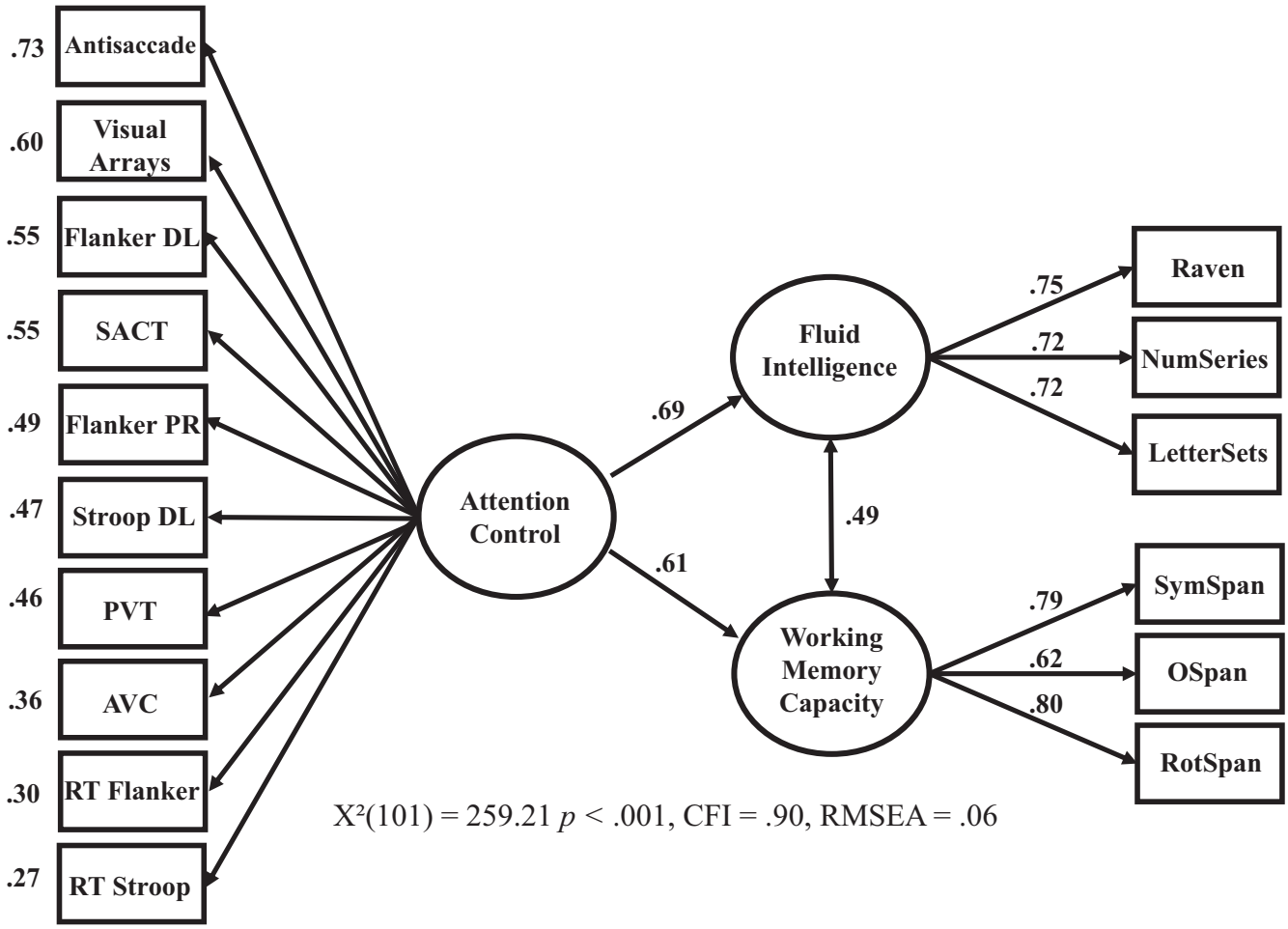

Figure 5. Structural equation model with all attention tasks loaded onto a single factor. Factor loadings for the attention control measures are on the left. Loadings and paths involving tasks in which lower scores indicate better performance (e.g., reaction times) were multiplied by −1. The correlation between working memory capacity and fluid intelligence is the correlation between their disturbance terms. Flanker DL = adaptive flanker response deadline; SACT = sustained attention-to-cue; Flanker PR = adaptive flanker presentation rate; Stroop DL = adaptive Stroop response deadline; PVT = psychomotor vigilance task; AVC = adaptive visual cue; RT Flanker = reaction time flanker effect; RT Stroop = reaction time Stroop effect; NumSeries = number series; SymSpan = symmetry span; OSpan = operation span; RotSpan = rotation span. N = 396.

that the attention control factor did not fully mediate the relationship between working memory capacity and fluid intelligence, as the path between the two (.33) was statistically significant even after accounting for attention control.

The second model (see Figure 6b) has the best performing attention tasks per the previous criteria (antisaccade, sustained attention-to-cue, and adaptive response deadline flanker). Here, the antisaccade and visual arrays had statistically equal loadings (.67 and .64, respectively), and the flanker deadline task loaded at .46. The attention factor had a strong relationship to working memory capacity (.75) and fluid intelligence (.65). Importantly, attention control *did* fully mediate the working memory capacity/fluid intelligence relationship in this model (the path from working memory capacity to fluid intelligence is .21 and nonsignificant). However, setting this path to .33 (the value from the first model) and using the model comparisons approach this .21 value was not statistically different from the mediation value in the first model. As such, attention control in this model did not statistically significantly mediate the working memory capacity/fluid intelligence relationship more than when the antisaccade, RT Stroop, and RT flanker were used.

The antisaccade was an indicator for attention control in the models from Figure 6a and Figure 6b. Although we argue that the antisaccade is a great measure of attention control because of its simplicity and high validity, it is not ideal to rely on one task to measure a construct. Also, it is possible that the antisaccade is anchoring the results of these two models. So, in the third model (see Figure 6c), the antisaccade was excluded and instead we tested the other best-performing tasks (visual arrays, modified flanker with adaptive response deadline, and sustained attention-to-cue). In this model, factor loadings to the attention factor were worse than in the second model (.64 for visual

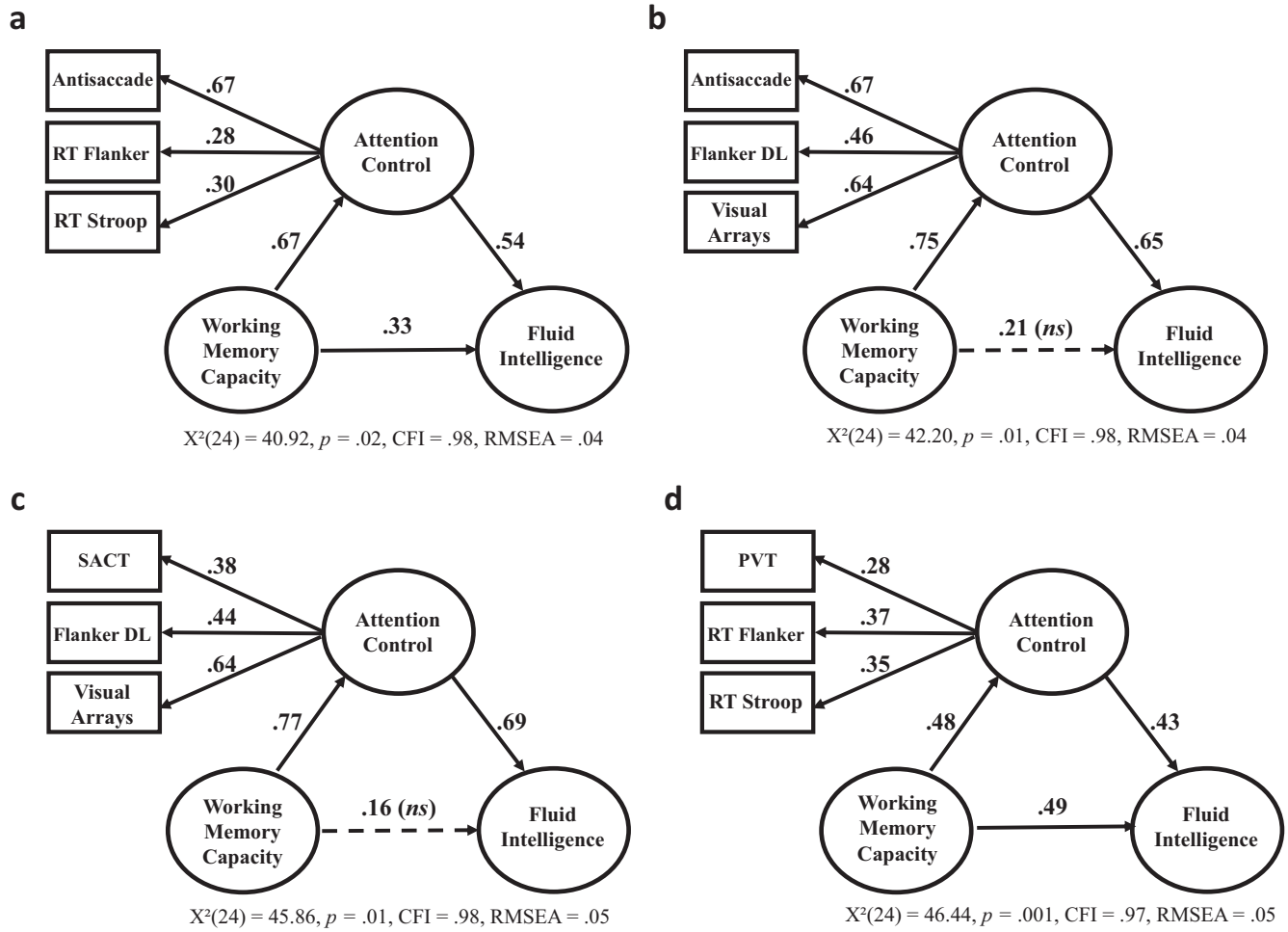

Figure 6. Structural equation models with attention control mediating the working memory capacity/fluid intelligence relationship. Different models are compared in which attention control comprises (a) traditional indicators—antisaccade, RT flanker, and RT Stroop; (b) the best three indicators based on reliabilities, intercorrelations, and relationship to working memory capacity and fluid intelligence—antisaccade, Flanker DL, and visual arrays; (c) the best three indicators excluding antisaccade—SACT, Flanker DL, and visual arrays; and (d) indicators with RT and RT difference scores as the dependent variable—PVT, RT flanker, and RT Stroop. Working memory capacity is comprised of symmetry span, operation span, and rotation span. Fluid intelligence is comprised of Raven's advanced, number series, and letter sets. RT flanker = reaction time flanker effect; RT Stroop = reaction time Stroop effect; SACT = sustained attention-to-cue; Flanker DL = adaptive flanker response deadline; PVT = psychomotor vigilance task. Nonsignificant paths are shown with a dashed line (N = 396).

arrays, .44 for modified flanker with adaptive response deadline, and .38 for sustained attention-to-cue), however these were still improvements over the RT Stroop and flanker tasks from the first model. Further, these three tasks fully mediated the working memory capacity/fluid intelligence relationship (the path was .16 and nonsignificant). Using the model comparisons approach, this was statistically significantly lower than the .33 from the first model. This indicates that the visual arrays, flanker deadline, and sustained attention-to-cue did statistically significantly improve attention control's mediation of the working memory capacity/fluid intelligence relationship over the more traditional attention tasks used in Figure 6a (antisaccade, RT Stroop, RT flanker).

In the fourth model (see Figure 6d) we tested attention control's mediation of the working memory capacity/fluid intelligence relationship with attention control comprised of the RT measures—the psychomotor vigilance task and standard RT difference score Stroop and flanker. This model was conducted to alleviate potential concerns that part of the reason these tasks performed relatively poorly compared with most of the accuracy-based ones is because the number of accuracy-based measures far outweighed the number of RT ones (7:3). But, these three RT measures cohered quite poorly, to the extent that, without imputation and depending on how data filtering was performed, there were issues with statistically nonsignificant factor loadings and model convergence. However, we were able to get the model to converge. Still,

Table 6
Correlations Between Attention Measures and Working Memory Capacity/Fluid Intelligence

Task	Correlation to WMC	Correlation to Gf
Visual Arrays	.43*	.46*
Antisaccade	.41*	.46*
Flanker DL	.29*	.33*
Flanker PR	.23*	.22*
Stroop DL	.21*	.26*
SACT	.21*	.23*
RT flanker effect	.18*	.16*
PVT	.12*	.20*
AVC	.09	.18*
RT Stroop effect	.09	.21*

Note. $N = 378–397$. Correlations involving one task in which lower scores indicate better performance (e.g., reaction time) were multiplied by −1 such that a positive correlation means individuals who performed better on one task also performed better on the other. Working memory capacity (WMC) and fluid intelligence (Gf) are *z*-score composites. Flanker DL = modified flanker with adaptive response deadline; flanker PR = modified flanker with adaptive presentation rate; Stroop DL = modified Stroop with adaptive response deadline; SACT = sustained attention-to-cue; RT = reaction time; PVT = psychomotor vigilance task; AVC = adaptive visual cue.
* $p < .05$.

the loadings were low (.29 to .42), and the paths from working memory capacity to attention control and from attention control to fluid intelligence were much weaker than in the other models (.48 and .43, respectively). This model was the weakest in terms of mediating the working memory capacity/fluid intelligence relationship, as the path between these two was .49 and statistically significant. Using model comparison, this .49 mediation value was statistically worse than in any other mediation model shown in Figure 6, indicating that the RT attention tasks had by far the weakest statistical mediation of the working memory capacity/fluid intelligence relationship.

Additional Analyses

We also performed three sets of post hoc analyses designed to address unresolved concerns of our own or raised during the review process. The first was informed by the exploratory factor analysis results in which we tested whether the three flanker tasks formed a coherent latent factor separate from the rest of the attention measures. The second was a test of the extent to which processing speed or general processing could account for the improvement of the new and modified attention tasks. The third set of analyses was conducted to explore whether using pure RTs on congruent and incongruent trials of the Stroop and flanker as the dependent variable would be a viable alternative to the Stroop and flanker tasks with adaptive response deadlines or presentation rate.

Exploratory factor analysis indicated that the flanker tasks could load onto their own factor separate from the other attention measures. We tested a structural equation model with the flanker tasks loaded onto a separate factor from the other attention measures, and both these factors predicting working memory capacity and fluid intelligence (see Figure 7). The model fit the data well and indeed the flanker tasks formed a coherent separable factor. However, this flanker factor contributed no statistically significant unique variance to working memory capacity or fluid intelligence above and beyond other attention measures. This indicates that the flanker factor reflected flanker-specific variance that is separable from other attention tasks but is not important for explaining variance in the working memory capacity/fluid intelligence relationship.

The second question was whether the tasks we have labeled as attention control reflect important variance related to attentive processes, or if instead they are contaminated with construct-irrelevant variance. For instance, individual differences in processing speed or general fluency might explain the high degree of shared variance among the attention tasks, as well as the attention measures to fluid intelligence and working memory capacity (see Hedge et al., 2020). It therefore was necessary to include some measures of discriminant validity to test whether attention control provided unique contribution to criterion measures above and beyond some baseline performance.

We tested this directly using data from a follow-up study in which we administered processing speed tasks to a subset of the same participants ($N = 173$). The processing speed measures were computerized versions of existing paper-and-pencil tasks; letter string comparison, digit string comparison, and digit symbol substitution (these are described in Appendix B, and the full correlation matrix involving these tasks are shown in Appendix C, Table C1). We were interested in the extent to which processing speed mediated the relationship between attention control and working memory capacity/fluid intelligence. If, for instance, processing speed accounted for this relationship entirely (full mediation), this would be a strong indication that the attention measures were indeed not process pure and that our results could potentially be attributed to the attention tasks employed here measuring general processing abilities rather than attention control.

We first tested two models with processing speed mediating attention control's relationship to either working memory capacity or fluid intelligence, reported in Appendix C (see Figures C1 and C2). These models show that processing speed did not mediate the relationship between attention control and either working memory capacity or fluid intelligence. In Figure 8, we report the critical test of whether general processing speed accounts for why our attention tasks correlated so strongly to working memory capacity and fluid intelligence. Here, processing speed and attention control are correlated but separable factors predicting both working memory capacity and fluid intelligence. If general task fluency or processing speed explained why our attention measures correlated strongly to working memory capacity and fluid intelligence, then we would expect to see the paths from processing speed to working memory capacity/fluid intelligence to be high, and the paths from attention control to working memory capacity/fluid intelligence to be low. In other words, we would see that attention control would not predict much variance in working memory capacity or fluid intelligence above and beyond processing speed. Instead, we found the exact opposite. Attention control and processing speed shared roughly 41% of their variance at the latent level, but processing speed accounted for no incremental variance in working memory capacity/fluid intelligence above and beyond attention control, whereas attention control accounted for substantial variance in these two abilities above and beyond processing speed (69% incremental variance to working memory capacity and 38% to fluid intelligence). The substantial incremental variance

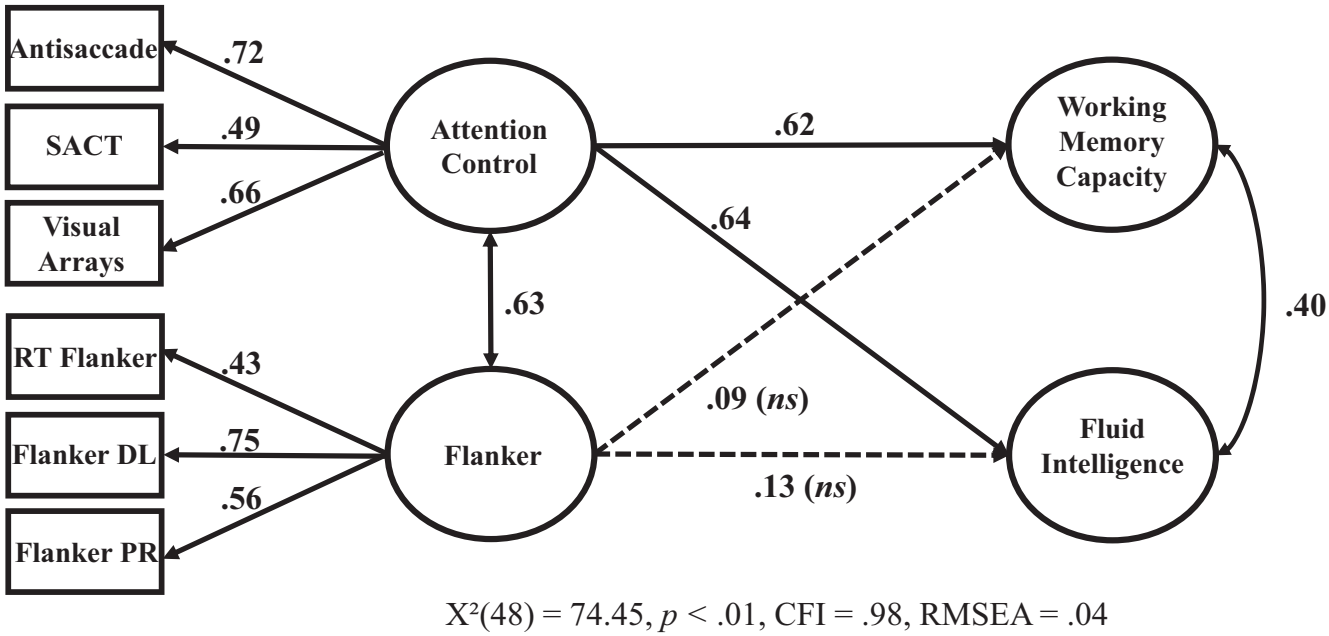

Figure 7. Structural equation model with attention control and flanker as separate factors predicting working memory capacity and fluid intelligence. Working memory capacity is comprised of operation span, symmetry span, and rotation span, with standardized loadings of .62, .79, and .80, respectively. The correlation between working memory capacity and fluid intelligence is the correlation between their disturbance terms. Nonsignificant paths are shown with a dashed line. SACT = sustained attention-to-cue; RT flanker = reaction time flanker effect; Flanker DL = adaptive flanker response deadline; Flanker PR = adaptive flanker presentation rate. N = 383.

contributed by attention control over processing speed shows that processes uniquely captured by the attention tasks are important for explaining the increased relationship to working memory capacity and fluid intelligence, whereas processing speed is not sufficient to account for this relationship.

The final set of analyses was conducted to test the possibility that using incongruent or congruent RT on the Stroop and flanker tasks as the dependent variable would result in similar improvements as the adaptive Stroop and flanker tasks. In other words, if the main problem with the traditional Stroop and flanker is that they are scored using difference scores, then why not avoid this by just using the component scores (mean RT on congruent and/or incongruent trials) instead? There is some precedent to this approach. Kane et al. (2016) reported that residual measures of Stroop and flanker performance (incongruent trials regressed on congruent or neutral) consistently correlated more strongly to within-construct measures than pure RT or accuracy difference scores from those tasks. The exception was their number Stroop, which did not correlate with other attention measures when scored either by pure difference score or the residual. For the number Stroop, they therefore used mean RT on incongruent trials and observed some improved correlations to other measures. In another study, Kane and McVay (2012) used incongruent RT as the indicator for Stroop performance in their structural equation models over the traditional Stroop effect because "Stroop incongruent RT showed stronger simple correlations with the other attention-control measures and loaded significantly on an attention-control latent factor" (p. 11).

In the present study, incongruent and congruent RTs were highly reliable and had stronger correlations to the other executive functioning tasks than did their resulting difference scores. But there is a problem. In Draheim, Mashburn, et al. (2019) we warned against the use of mean RT of congruent or incongruent trials as the dependent variable in the Stroop and flanker because there is no control for general processing abilities, such as processing speed and task fluency. While contrasts (difference scores) are problematic in correlational research due to unreliability and attenuated correlations, they are used as a control for baseline performance and general processing (though researchers have raised doubts as to the efficacy of difference scores as a control for baseline processing, particular for RT differences; e.g., Hedge et al., 2020; Rey-Mermet et al., 2019; Verhaeghen & De Meersman, 1998). Using instead congruent and/or incongruent trial performance in the Stroop and flanker tasks properly accounts for neither speed-accuracy interactions nor general processing. Our data bear this out as well. Reaction time on the congruent and incongruent trials correlated with processing speed quite strongly, an average of r = .46, compared with the modified Stroop and flanker tasks or RT difference scores from the standard Stroop and flanker (an average of r = .23). We tested the statistical significance of the difference between the correlations (Steiger, 1980) and found that both incongruent and congruent RT on the Stroop and flanker correlated more strongly to processing speed than their corresponding modified Stroop and flanker variant (α = .05; no correction for multiple comparisons). On the other hand, the modified Stroop and flanker tasks did not significantly correlate any stron-

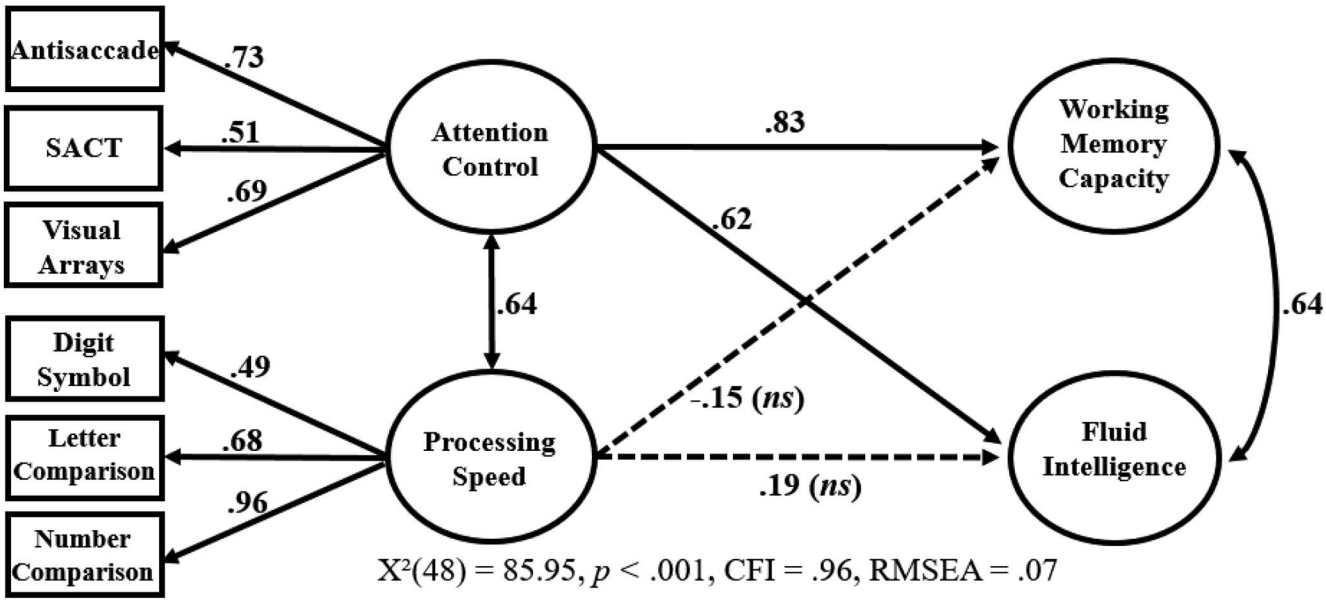

Figure 8. Structural equation model showing the relative contributions of attention control and processing speed on working memory capacity and fluid intelligence. Working memory capacity comprising symmetry span, operation span, and rotation span, with loadings of .83, .63, and .81, respectively. Fluid intelligence comprising Raven's advanced, number series, and letter sets with loadings of .76, .79, and .84, respectively. The correlation between working memory capacity and fluid intelligence is the correlation between their disturbance terms. Nonsignificant paths are shown with a dashed line (N = 173). SACT = sustained attention-to-cue task.

ger to processing speed than the RT difference score versions of the tasks. This analysis highlights why using simple mean performance on Stroop and flanker tasks as the dependent variable is not a good approach—these scores correlate more strongly across the board but not discriminately as there is less control for construct-irrelevant variance. This analysis also suggests that our modified Stroop and flanker tasks were no more contaminated with processing speed than RT difference scores from the standard Stroop and flanker tasks.

A limitation is that the above analysis was based on task-level correlations and not latent relationships. We therefore tested a related hypothesis using structural equation modeling (see Figure 9) which had three factors predicting attention control (operationalized here as sustained attention to-cue, visual arrays, and antisaccade). The model was a test of the relative contributions to attention control from the Stroop and flanker tasks with the adaptive response deadline, processing speed, and mean RT on Stroop and flanker. That is, how much variance each contributes when accounting for (partialing out) variance shared with the other factors. In this model, the adaptive response deadline versions of Stroop and flanker used in the present study contributed substantial unique variance (29%) to attention control above and beyond both processing speed and RTs on congruent/incongruent Stroop and flanker. Further, RTs on congruent and incongruent Stroop and flanker trials contributed no statistically significant variance to attention control beyond processing speed or the adaptive Stroop and flanker response deadline tasks.

The large and positive association between processing speed and attention control even after accounting for the other two factors is also noteworthy. This suggests that processing speed also has independent contributions to attention control that cannot be accounted for by performance in the Stroop or flanker tasks. This would be a cause for concern if performance on the response deadline Stroop and flanker tasks did not also contribute substantial unique variance to attention control. However, both processing speed and the adaptive response deadline Stroop and flanker tasks contributed unique variance to attention control, which shows that processing speed is at play in tasks such as antisaccade and visual arrays. This is not surprising as we would expect general processing/processing speed to be involved in performance in any executive functioning task to some degree, and it highlights the importance of using structural equation modeling and other regression techniques to partial out variance to test the unique contributions of different constructs.

Discussion

To aid discussion, Table 7 has rankings of each task in terms of their performance on the four criteria tested here. On the whole, the accuracy-based measures of attention performed markedly better than the existing RT measures. In particular, the RT Stroop and flanker measures were poor on every criterion. This was expected given the previous literature on the problems with these tasks (e.g., Draheim, Mashburn, et al., 2019; Hedge et al., 2020; Rouder & Haaf, 2019).

The antisaccade, visual arrays, adaptive response deadline flanker and sustained attention-to-cue tasks performed very well in terms of their rank-ordering on all criteria. They were the most stable across two administrations, they were the most strongly intercorrelated, and they best formed a coherent latent factor which

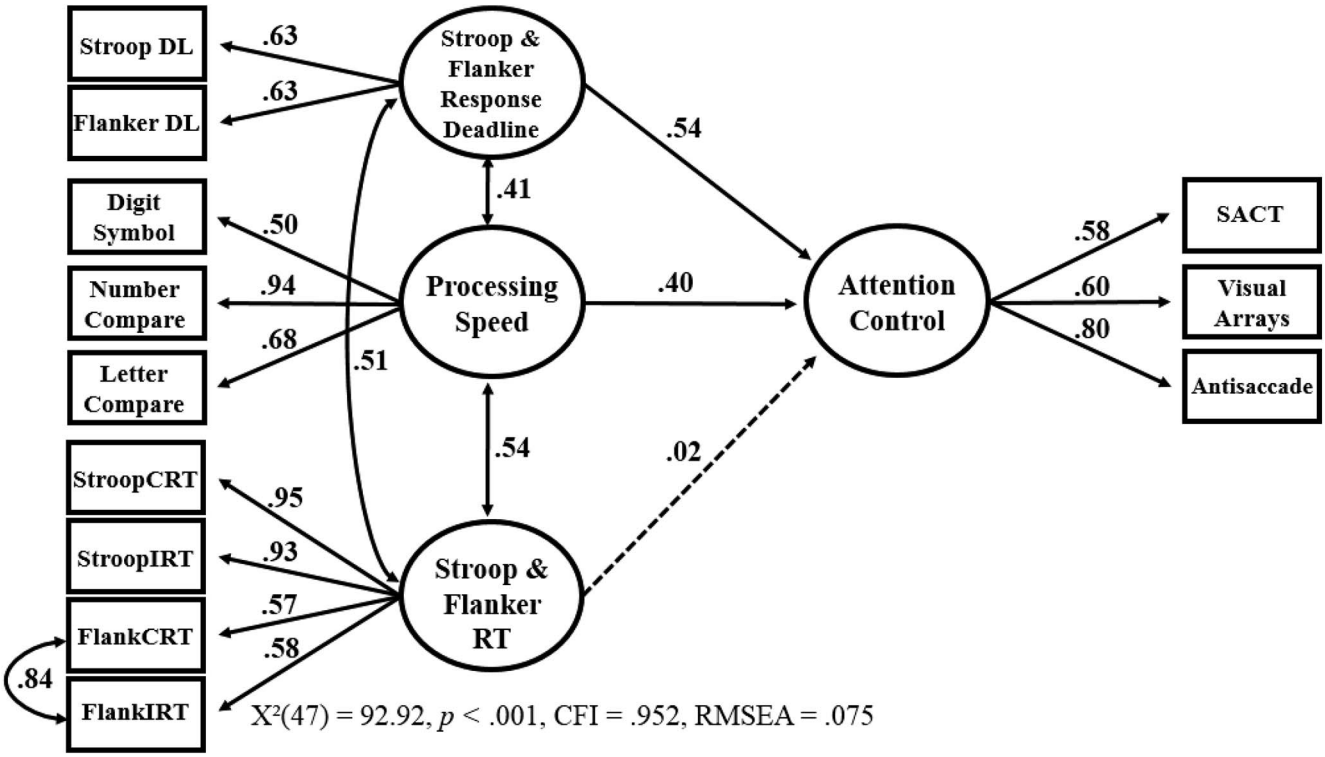

Figure 9. Structural equation model testing the relative contributions of the Stroop and flanker deadlines tasks, processing speed, and mean reaction time (RT) in Stroop and flanker to attention control. Reaction time on flanker incongruent and congruent trials needed to have correlated error terms for the CFI and RMSEA fit indices to reach acceptable levels, hence why the error terms were correlated for these two indicators. The model without this additional path had a CFI of around .70 and an RMSEA > .10, but all other paths were numerically very similar regardless of whether flanker RTs were allowed to correlate. Paths are reported as positive if better performance on one task or factor was associated with better performance on another (i.e., a positive path between the adaptive Stroop and flanker response deadlines and attention control indicates that individuals with larger and better scores in the attention tasks had lower response deadline thresholds in the Stroop and flanker deadline tasks; $N = 173$). StroopCRT = mean RT on congruent trials in the standard Stroop task; StroopIRT = mean RT on incongruent trials in the standard Stroop task; FlankCRT = mean RT on congruent trials in the standard flanker task; FlankIRT = mean RT on incongruent trials in the standard flanker task; SACT = sustained attention-to-cue task.

in turn had a strong relationship with working memory capacity and fluid intelligence. The selection visual arrays and sustained attention-to-cue tasks correlated to antisaccade at or above $r = .40$. This magnitude of correlation is noteworthy because these are superficially quite different tasks that place different sorts of demands on the respondent. Further, correlations this strong are rarely observed among attention or inhibition measures.

Directly comparing the adaptive response deadline flanker to the standard RT flanker interference effect, the adaptive version was better across all criteria. The test–retest reliability, correlation to antisaccade, average factor loading onto an attention factor, and correlation to working memory capacity and fluid intelligence were roughly doubled for the adaptive response deadline flanker over the standard RT flanker. The latent analyses presented in Figure 7 shows that the traditional and modified flanker tasks still loaded together, which is one piece of evidence that the manner in which we modified the flanker task improved its psychometrics without drastically changing the nature of the task. This is noteworthy because a primary concern with the manner in which we modified the Stroop and flanker tasks was that interference effects were not directly reflected in the score.

The response deadline Stroop and presentation rate flanker performed in the middle. These measures were an improvement to the traditional RT Stroop and flanker tasks, but less so when compared with the antisaccade, visual arrays, sustained attention-to-cue, and response deadline flanker (see Table 7).

The psychomotor vigilance task also performed in the middle but was the best RT measure of the three tested. We had reasoned that this might be an example of a RT-based measure better suited to individual differences than most other RT measures because it is not difference score-based and does not appear to be sensitive to speed-accuracy interactions. And indeed the psychomotor vigilance task was highly reliable in terms of internal consistency, had moderately strong intercorrelations with the other attention measures, and loaded with the other high-performing attention measures in an exploratory factor analysis. However, this task had an

Table 7
Ranking of Attention Tasks Based on Criteria Used in Present Study

Task	Internal consistency	Test–retest reliability	Intercorrelation	Factor coherence	Relationship to WMC and Gf	Rank
Antisaccade	3	1	1	1	2	1
Visual arrays	4	2	3	2	1	2
SACT	2	4	2	4	6	3
Flanker DL		6	5	3	3	4
Stroop DL		3	7	6	4	5
PVT	1	5	4	7	8	6
Flanker PR		9	5	5	5	6
AVC		8	8	8	10	8
RT flanker	6	10	9	8	7	8
RT Stroop	5	7	10	10	9	10

Note. Values equivalent to the second decimal were considered a tie. Internal consistency was not factored into total rank since it was not calculated for the adaptive tasks. Test-retest rankings reflect reliability after removing outliers for the task. Factor coherence was defined as the average factor loading for each task across all 120 possible tri-indicator models. Relationship to working memory capacity (WMC) and fluid intelligence (Gf) was defined as the average of each task's correlation to composites of both WMC and Gf. SACT = sustained attention-to-cue; Flanker DL = modified flanker with adaptive response deadline; Stroop DL = modified Stroop with adaptive response deadline; PVT = psychomotor vigilance task; flanker PR = modified flanker with adaptive presentation rate; AVC = adaptive visual cue; RT flanker = reaction time flanker effect; RT Stroop = reaction time Stroop effect.

average factor loading of only .40 in the models we tested (see Table 5) and had a weak relationship to working memory capacity (r = .12). It may therefore be that performance on the psychomotor vigilance task reflects processes related to attention control that are less important for performing higher-order cognitive tasks. However, this question is beyond the scope of the present study.

The adaptive visual cue also performed very poorly across the board. One potential reason is that, even after removing outliers that were 3.5 standard deviations above or below the mean, performance on this task was highly non-normal with skew of 4.91 and kurtosis of 38.39 (see Table 2). However, the results from the adaptive visual cue task did not improve even when we normalized scores using an iterative procedure that continuously removes outliers and calculates a new mean until skewness and kurtosis is under three. As such, there is a fundamental problem with this task conceptually, in its implementation, and/or with its scoring.

In summary, the adaptive response deadline flanker was an improvement to the standard RT flanker; the sustained attention-to-cue with 64 trials was roughly equal in reliability to the psychomotor vigilance task with 80 but an improvement otherwise; and the visual arrays performed very strongly and on par with the antisaccade.

The mediation analyses (see Figure 6) showed that attention control could fully mediate the working memory capacity/fluid intelligence relationship, which is not something we had observed with previous studies when using the standard RT difference score versions of the Stroop and flanker. This full mediation is theoretically interesting because it supports our theory that individual differences in executive functioning arise due to individual differences in executive attention (attention control). However, full mediation of attention control on the working memory capacity/fluid intelligence relationship was only possible when visual arrays was included in the model—the other new and modified attention tasks did not achieve full mediation without the visual arrays. This is worth exploring further, which we do in Appendix D.

Finally, we found evidence that performance on flanker tasks reflected different variance from other attention measures, although flanker performance did not predict unique variance to working memory capacity or fluid intelligence. This did not appear to occur simply due to having more flanker tasks than, say, Stroop tasks or sustained attention tasks, as we found that flanker performance was separable even when the analyses included only two flanker tasks. These results are consistent with views that flanker tasks reflect different processes than other attention tasks (e.g., Paap et al., 2019). For example, in the Friedman and Miyake (2004) taxonomy, flanker performance relies on the ability to resist distractor interference whereas Stroop and antisaccade require inhibition of a prepotent response. Similarly, Kane et al. (2016) argued that tasks such as antisaccade and Stroop require restraining attention to override a prepotent response with a novel and goal-directed one, whereas flanker tasks require constrained attention to identify targets among visual distractors. Another example is Kornblum's (1994) dimensional-overlap model which distinguishes two types of conflict in attention tasks, stimulus-stimulus incompatibility and stimulus-response incompatibility, with flanker tasks often used to assess the former and Simon tasks the latter (cf., Paap et al., 2019). Still, the emergence of a coherent and separable flanker factor was surprising. Although flanker performance is thought to reflect different processes than other attention tasks, a number of studies have reported a lack of reliable and valid individual differences in standard flanker tasks. For example, Salthouse (2010) concluded that while the flanker task was sensitive to attention-related conflict (and therefore suitable for experimental studies), there were no systematic individual differences in the resulting flanker scores and so he questioned whether the flanker should even be used as an individual differences measure. Paap et al. (2019) found rather strong correlations among inverse efficiency (RT/error rate) scores on a Simon and two Stroop tasks, but not with the flanker task. And Rouder et al. (2019) determined that individual variation in arrow flanker performance from Rey-

Mermet et al.'s (2018) data was "so small that it can be accounted for with trial variation alone" (p. 20). From these results and ours, we conclude that there are individual differences in the ability to resolve conflict in flanker-like tasks, but that these differences are often muted due to methodological issues with how flanker performance is typically assessed (RT difference scores).

A Commentary on Rey-Mermet et al. (2019)

Our findings are at odds with many individual differences investigations of attention control/inhibition. We found relatively strong intercorrelations among attention, a coherent attention control factor, and strong criterion validity with other cognitive ability measures; other studies of this type generally report the opposite (see Draheim, Mashburn, et al., 2019; Friedman & Miyake, 2004; Paap et al., 2019; Rey-Mermet et al., 2018, 2019). Rey-Mermet et al.'s (2019) study is of particular relevance because their goal was putatively the same as ours—to test the unity of attention control and its relationship to working memory capacity/fluid intelligence using accuracy-based attention measures. There are important differences between our study and theirs which likely account for the discrepant results and conclusions. These include their use of difference scores to assess performance and increased attentional demand in their baseline trials due to time pressure and carryover effects. These are discussed in more detail in the following paragraphs.

Rey-Mermet et al. (2019) and the present study used different novel approaches to isolate variance of interest. Here, we included tasks designed to demand greater controlled attention and without requiring scoring via differences or contrasts. We also tested whether improvements in the tasks were due to the introduction of construct-irrelevant variance common across the tasks such as processing speed and/or general task fluency (see Figure 8 and Figures C1 and C2 in Appendix C). These analyses showed that, while processing speed and attentional control were correlated, attention control predicted a substantial amount of variance in working memory capacity/fluid intelligence over processing speed. Rey-Mermet et al. used a different approach, including a calibration procedure designed to eliminate variance associated with processing speed, episodic memory, and general task ability. However, they did not test the efficacy of their novel manipulations.

One major difference between Rey-Mermet et al. (2019) and the present study was that Rey-Mermet et al. scored all but one of their attention tasks using difference scores. The problems with scoring performance on executive functioning tasks using difference scores were mentioned in the introduction and have been discussed at length elsewhere (e.g., Draheim, Mashburn, et al., 2019; Hedge et al., 2018; Hughes et al., 2014; Paap & Sawi, 2016). Difference scores are used despite their known problems because there is a widely held belief that difference scores are necessary to account for baseline performance and therefore can help isolate variance of interest. However, there is a strong argument that difference scores are unsatisfactory for this purpose (e.g.,Hedge et al., 2020; Verhaeghen & De Meersman, 1998). Rey-Mermet et al. acknowledged this in regard to RT difference scores:

> Subtracting RTs is premised on the assumption of additive factors. That is, the duration of the processes in the baseline condition and the duration of the executive-control process combine additively to the RT in incongruent trials, and the duration of each process is uniquely affected by its own source of individual differences. However, this assumption is questionable because across various speeded tasks, the RTs of slow individuals are related to those of fast individuals through a constant proportional slowing factor (Zheng, Myerson, & Hale, 2000). This implies that differences between RTs are also proportionally larger for slower than for faster individuals (p. 1339).

Although Rey-Mermet et al. (2019) used accuracy-based difference scores and not RT-based ones, it is an open question as to whether accuracy difference scores suffer from the same shortcomings as RT differences. We would argue that many of the concerns expressed by Draheim, Mashburn, et al. (2019) apply to both RT and accuracy difference scores—we took aim at RT ones because they are used far more often in assessing executive functioning. Further, Hedge et al. (2020) found that both accuracy and RT difference scores in the Stroop, flanker, and Simon tasks are contaminated by processing speed and response cautiousness. Independent of process purity, difference scores are also less reliable than their components, to that end Rey-Mermet et al.'s (2019) measures had an average internal consistency of .70 as opposed to .83 for our nonadaptive measures (see Table 2), although some of their measures were quite reliable. But other methodological factors cannot be ruled out as the reason for the discrepant results, such as differences in population, sample size, number of trials, effect sizes, ratio of baseline trials (congruent or neutral) to incongruent, just to name a few.

Rey-Mermet et al. also employed a within-subject calibration procedure in each of their attention tasks. In this calibration procedure, response deadlines for baseline trials (neutral Stroop and flanker trials, congruent Simon trials, and prosaccade trials) increased or decreased to converge upon a 75% accuracy threshold for each participant. This adaptive response deadline then carried over into the experimental blocks for each task, which, for the Stroop, flanker, and Simon tasks contained the baseline trials plus incongruent trials, and for the antisaccade task were pure antisaccade trials. It is not clear how this adaptive procedure changed the nature of these tasks, and Rey-Mermet et al. (2019) acknowledged some potential pitfalls of this approach. The calibration procedure used by Rey-Mermet et al. may account for their null findings. First, reducing the response deadline of baseline trials (e.g., neutral trials in Stroop, prosaccade in the antisaccade task) during calibration would lead participants to engage in more controlled processing than when not under time pressure. That is, time pressure increases attentional demands. Indeed, studies have shown that task performance is altered under time pressure (Earles, Kersten, Berlin Mas, & Miccio, 2004), that time pressure in Raven's increases its correlation to working memory capacity (Chuderski, 2015), and that the ability to perform under time pressure predicts decision-making performance (Joslyn & Hunt, 1998). Second, Rey-Mermet et al. noted that administering prosaccade trials immediately prior to antisaccade trials could impact antisaccade trials and reduce individual differences (e.g., Kane et al., 2001). Our concern is with the opposite: Kane et al. (2001) showed that there were carryover effects when presenting prosaccade (baseline) trials after antisaccade trials such that the prosaccade trials involved more controlled processing, resulting in performance differences between high and low working memory capacity individuals. Rey-Mermet et al. administered calibration

(baseline) blocks both before and after the participants performed the critical attention trials. Participants having performed the experimental trials shortly beforehand would therefore introduce interference into the calibration trials and require the participants to reinstantiate the new goals of the task (e.g., look toward the cue now instead of away from it as you have been previously). We argue that the baseline trials used in Rey-Mermet et al.'s calibration procedure likely involved a significant degree of controlled processing due to these carryover effects and the aforementioned time their practice of subtracting out calibration trial performance from experimental trial performance reduced not just general construct-irrelevant variance (such as processing speed) as they intended, but also reduced the amount of controlled attention reflected in the dependent variables for these tasks.

Alternatively, it is possible that our results and Rey-Mermet et al.'s (2019) are not as contradictory as they appear. We may have different conceptualizations of what these tasks are designed to measure, and we may therefore be simply referring to different abilities despite using shared terminology and tasks. We hold the view that attention control is a broad and domain-general ability responsible in part for individual differences in most, if not all, higher order cognitive tasks (Engle, 2002, 2018; Kane, Conway, Hambrick, & Engle, 2007). From this perspective, there may be no separate attention control mechanisms, but rather tasks place different demands on attention. Rey-Mermet et al. used the narrower term *inhibition*, ostensibly conceptualized as a specific cognitive mechanism which is required by tasks such as Stroop, flanker, and Simon. Their argument is that inhibitory (or conflict-resolution) processes in these tasks are task-specific. Perhaps both are true, and that Rey-Mermet et al.'s methodology (difference scores and their calibration technique) reduced, even eliminated, variance associated with general attentional control and therefore the leftover reliable variance reflected the very narrow ability to resolve interference in the specific task at hand, whereas our findings are due to influences of domain-general attention control that cannot be attributable to any single mechanism or ability.

Recommendation for Future Research

Future research should focus on further development and improvement of attention measures, and to that end the present study provides just one potential method for doing so. For researchers interested in using the tasks employed here, they were programmed in E-Prime and we will make them available for download on our website at http://englelab.gatech.edu/taskdownloads.

We can in good conscience recommend the antisaccade and the selective-based visual arrays to other researchers interested in assessing individual differences in attentional control. These are established tasks that display good psychometric properties with relatively little time investment. These tasks were the best performing across all criteria with the exception of internal consistency in the visual arrays, though visual arrays retained most of its reliable variance in test-retest even after an average of around six months between administrations.

Two other promising tasks are the sustained attention-to-cue and the adaptive response deadline flanker, with the adaptive response deadline Stroop just behind those two. These tasks performed quite well, particularly when compared with the traditional RT difference score Stroop and flanker tasks. However, these tasks were programmed more as a proof of concept than as finished products. These tasks would very likely benefit from further modification, which we intend to explore in future studies. Researchers interested in using these tasks are encouraged to contact the corresponding author for recommended changes.

We continue to warn against the use of the standard Stroop and flanker tasks in individual differences contexts, as these two tasks expectedly performed the worst among all attention measures. These tasks are known to be problematic. For example, Rouder et al.'s (2019) findings show that typical administrations of these tasks involve a high degree of trial-level noise, and that that hundreds if not more trials may be required to raise Stroop and flanker reliability and correlations to desired levels. Further, the work by Hedge et al. (2020) indicates that these tasks reflect little variance associated with the processes they are intended to measure.

Conclusion

The toolbox approach used here to develop attention control measures appears to be a viable method for moving away from traditional experimental tasks which are not suitable for correlational pursuits (cf., Hedge et al., 2018). Specifically, developing new attention tasks which do not rely on difference scores and in a manner that minimizes influences from speed-accuracy interactions and processing speed is a promising method for establishing reliable and valid assessments of attention control. Our results demonstrate that attention control is indeed a unitary concept and, as such, that attention measures can reflect much more than task-specific variance. This finding is contrary to results with difference score-based tasks (e.g., Friedman & Miyake, 2004; Rey-Mermet et al., 2018, 2019; Rouder & Haaf, 2019; Rouder et al., 2019; but see Paap et al., 2019). Although a coherent latent attention factor emerged in the present study, we also found evidence for specific and separable sources of variance in some of tasks (e.g., flanker).

Statistical and post hoc attempts to improve the measurement of attention control had not consistently yielded positive results. It was clear that, as Friedman and Miyake (2004) concluded, new attention control measures were needed to make any further progress. In the present study, we pushed performance variance into accuracy and developed new and modified accuracy-based attention measures designed to measure similar processes as other attention measures but without the methodological problems. The results show that these tasks are improvements to existing attention measures, which has implications for how attention control can be better measured moving forward. To that end we hope to see more development of new attention measures instead of continued reliance on measures that have proven to be inadequate. At the very least, we hope that investigators assessing individual differences in attention control and related processes will be cognizant of the methodological issues with many existing attention tasks and consider taking steps to circumvent these issues. Doing so will improve theoretical studies involving attention by lowering the likelihood of finding null, misleading, or inconsistent results and thereby reducing the occurrence of misguided conclusions based on these results.

Context of Research

The present research was motivated by the current scientific debate regarding the nature and measurement of attention control. Attention

researchers struggle to find a coherent and unified attention control factor, and the reliability and validity of commonly used attention measures are generally low. We argue that the reasons for these problems are primarily methodological, not theoretical, and that researchers should not give up on the construct. But several previous statistical attempts to solve the methodological problems with attention control had been unsuccessful. This article was therefore a test of whether employing new tasks, modified tasks, and tasks not generally recognized as primarily reflecting attention processes could lead to better results.

More broadly, tasks which are excellent for experimental (group differences) purposes are often poorly suited for correlational (individual differences) purposes. This phenomenon is beginning to receive more attention in the literature due to the increased awareness of reliability and validity issues with a number of ubiquitous cognitive tasks. The hope is that this line of work will raise awareness to these issues as they pertain to attention control specifically but also in the broad sense. And while there are a variety of ways to address the problem, the authors primarily aim to solve it through continued modification and development of tasks to have more desirable psychometric properties.

References

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103,* 411–423. http://dx.doi.org/10.1037/0033-2909.103.3.411

Atksinon, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *Psychology of learning and motivation: Advances in research and theory* (Vol. 2, pp. 89–195). New York, NY: Academic Press. http://dx.doi.org/10.1016/s0079-7421(08)60422-3

Baddeley, A., & Hitch, G. (1998). Recent developments in working memory. *Psychology of Learning and Motivation, 8,* 234–238. http://dx.doi.org/10.1016/S0959-4388(98)80145-1

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

Broadway, J. M., Redick, T. S., & Engle, R. W. (2010). Individual differences in working memory capacity: Control is (in) the goal. In R. Hassin, K. Ochsner, & Y. Trope (Eds.), *Self-control in society, mind, and brain* (pp. 163–174). New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195391381.003.0009

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. *Sage Focus Editions, 21,* 230–258. http://dx.doi.org/10.1177/0049124192021002005

Chiou, J. S., & Spreng, R. A. (1996). The reliability of difference scores: A re-examination. *Journal of Consumer Satisfaction, Dissatisfaction & Complaining Behavior, 9,* 158–167.

Chuderski, A. (2015). The broad factor of working memory is virtually isomorphic to fluid intelligence tested under time pressure. *Personality and Individual Differences, 85,* 98–104. http://dx.doi.org/10.1016/j.paid.2015.04.046

Comrey, A. L., & Lee, H. B. (Eds.). (1992). Interpretation and application of factor analytic results. *A first course in factor analysis* (pp. 240–262). Hillsdale, NJ: Lawrence Erlbaum Associates. http://dx.doi.org/10.4324/9781315827506

Conway, A. R. A., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence, 30,* 163–183. http://dx.doi.org/10.1016/S0160-2896(01)00096-4

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12,* 769–786. http://dx.doi.org/10.3758/BF03196772

Conway, A. R. A., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences, 7,* 547–552. http://dx.doi.org/10.1016/j.tics.2003.10.005

Cowan, N., Elliott, E. M., Scott Saults, J., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology, 51,* 42–100. http://dx.doi.org/10.1016/j.cogpsych.2004.12.001

Cronbach, L. J., & Furby, L. (1970). How should we measure "change"—Or should we? *Psychological Bulletin, 74,* 68–80. http://dx.doi.org/10.1037/h0029382

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

Draheim, C., Harrison, T. L., Embretson, S. E., & Engle, R. W. (2018). What item response theory can tell us about the complex span tasks. *Psychological Assessment, 30,* 116–129. http://dx.doi.org/10.1037/pas0000444

Draheim, C., Hicks, K. L., & Engle, R. W. (2016). Combining reaction time and accuracy: The relationship between working memory capacity and task-switching as a case example. *Perspectives on Psychological Science, 11,* 133–155. http://dx.doi.org/10.1177/1745691615596990

Draheim, C., Martin, J. D., Tsukahara, J. S., Mashburn, C. A., & Engle, R. W. (2018, November). *Measurement of attention control*. Paper presented at the 59th meeting of the Psychonomic Society, New Orleans, LA.

Draheim, C., Mashburn, C. A., & Engle, R. W. (2018, November) *Reaction times do not reliably measure individual differences in cognition: The problem and solutions*. Poster presented at the 59th Annual Meeting of the Psychonomic Society, New Orleans, LA.

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

Draheim, C., Tsukahara, J. S., Martin, J. D., Mashburn, C. A., & Engle, R. W. (2019, November). *Attention control is a unitary concept when measured with accuracy-based tasks*. Poster presented at the 60th Annual Meeting of the Psychonomic Society, Montréal, Québec, Canada.

Earles, J. L., Kersten, A. W., Berlin Mas, B., & Miccio, D. M. (2004). Aging and memory for self-performed tasks: Effects of task difficulty and time pressure. *The Journals of Gerontology Series B, Psychological Sciences and Social Sciences, 59,* 285–293. http://dx.doi.org/10.1093/geronb/59.6.P285

Edwards, J. R. (2001). Ten difference score myths. *Organizational Research Methods, 4,* 265–287. http://dx.doi.org/10.1177/109442810143005

Egeth, H. E., & Yantis, S. (1997). Visual attention: Control, representation, and time course. *Annual Review of Psychology, 48,* 269–297. http://dx.doi.org/10.1146/annurev.psych.48.1.269

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual for kit of factor referenced cognitive tests*. Princeton, NJ: Educational Testing Service.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11,* 19–23. http://dx.doi.org/10.1111/1467-8721.00160

Engle, R. W. (2017, November 9–12). *Working memory capacity and intelligence*. Keynote address at the 58th Annual Meeting of the Psychonomic Society. Vancouver, British Columbia, Canada.

Engle, R. W. (2018). Working memory and executive attention: A revisit. *Perspectives on Psychological Science, 13,* 190–193. http://dx.doi.org/10.1177/1745691617720478

Engle, R. W., & Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 44, pp. 145–199). New York, NY: Elsevier. http://dx.doi.org/10.1016/s0079-7421(03)44005-X

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,* 143–149. http://dx.doi.org/10.3758/BF03203267

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human participants research. *Proceedings of the National Academy of Sciences, 115*(27), E6106–E6115. http://dx.doi.org/10.1073/pnas.1711978115

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

Fukuda, K., & Vogel, E. K. (2011). Individual differences in recovery time from attentional capture. *Psychological Science, 22,* 361–368. http://dx.doi.org/10.1177/0956797611398493

Fukuda, K., Woodman, G. F., & Vogel, E. K. (2015). Individual differences in visual working memory capacity: Contributions of attentional control to storage. *Mechanisms of Sensory Working Memory: Attention and Perfomance, XXV,* 105–119. http://dx.doi.org/10.1016/B978-0-12-801371-7.00009-0

Goodhew, S. C., & Edwards, M. (2019). Translating experimental paradigms into individual-differences research: Contributions, challenges, and practical recommendations. *Consciousness and Cognition, 69,* 14–25. http://dx.doi.org/10.1016/j.concog.2019.01.008

Hairston, W. D., & Maldjian, J. A. (2009). An adaptive staircase procedure for the E-Prime programming environment. *Computer Methods and Programs in Biomedicine, 93,* 104–108. http://dx.doi.org/10.1016/j.cmpb.2008.08.003

Hedge, C., Powell, G., Bompas, A., & Sumner, P. (2020, February 1). Strategy and processing speed eclipse individual differences in control ability in conflict tasks. *PsyArXiv Preprints*. http://dx.doi.org/10.31234/osf.io/vgpxq

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

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience, 8,* 150. http://dx.doi.org/10.3389/fnins.2014.00150

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55. http://dx.doi.org/10.1080/10705519909540118

Hughes, M. M., Linck, J. A., Bowles, A. R., Koeth, J. T., & Bunting, M. F. (2014). Alternatives to switch-cost scoring in the task-switching paradigm: Their reliability and increased validity. *Behavior Research Methods, 46,* 702–721. http://dx.doi.org/10.3758/s13428-013-0411-5

Hutchison, K. A. (2007). Attentional control and the relatedness proportion effect in semantic priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 645–662. http://dx.doi.org/10.1037/0278-7393.33.4.645

Joslyn, S., & Hunt, E. (1998). Evaluating individual differences in response to time-pressure situations. *Journal of Experimental Psychology: Applied, 4,* 16–43. http://dx.doi.org/10.1037/1076-898X.4.1.16

Kaernbach, C. (1991). Simple adaptive testing with the weighted up-down method. *Perception & Psychophysics, 49,* 227–229. http://dx.doi.org/10.3758/BF03214307

Kane, M. J., Bleckley, M. K., Conway, A. R., & Engle, R. W. (2001). A controlled-attention view of working-memory capacity. *Journal of Experimental Psychology: General, 130,* 169–183. http://dx.doi.org/10.1037/0096-3445.130.2.169

Kane, M. J., Conway, A. R., Hambrick, D. Z., & Engle, R. W. (2007). Variation in working memory capacity as variation in executive attention and control. In A. R. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory* (pp. 21–48). New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195168648.003.0002

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General, 133,* 189–217. http://dx.doi.org/10.1037/0096-3445.133.2.189

Kane, M. J., & McVay, J. C. (2012). What mind wandering reveals about executive-control abilities and failures. *Current Directions in Psychological Science, 21,* 348–354. http://dx.doi.org/10.1177/0963721412454875

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

Kenny, D. A. (2015). *Measuring model fit*. Retrieved from http://davidakenny.net/cm/fit.htm

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,* 130–135. http://dx.doi.org/10.1007/bf00419699

Loehlin, J. C. (1987). *Latent variable models: An introduction to factor, path, and structural analysis* (3rd ed.). Hillsdale, NJ: Lawrence Erlbaum. http://dx.doi.org/10.2307/1165058

Logie, R. H., Della Sala, S., Laiacona, M., Chalmers, P., & Wynn, V. (1996). Group aggregates and individual reliability: The case of verbal short-term memory. *Memory & Cognition, 24,* 305–321. http://dx.doi.org/10.3758/BF03213295

Lord, F. M. (1956). The measurement of growth. *ETS Research Bulletin Series, 1956*(1), i-22. http://dx.doi.org/10.1002/j.2333-8504.1956.tb00058.x

Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 21–38). Madison, WI: University of Wisconsin Press.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390,* 279–281. http://dx.doi.org/10.1038/36846

Magnusdottir, B. B., Faiola, E., Harms, C., Sigurdsson, E., Ettinger, U., & Haraldsson, H. M. (2019). Cognitive measures and performance on the antisaccade eye movement task. *Journal of Cognition, 2,* 1–13. http://dx.doi.org/10.5334/joc.52

Martin, J. D., Mashburn, C. A., & Engle, R. W. (2019). *ASVAB, multitasking, and attention control*. Manuscript in preparation.

Martin, J. D., Tsukahara, J. S., Draheim, C., Shipstead, Z., Mashburn, C. A., & Engle, R. W. (2019). The visual arrays task: Visual working memory capacity or attention control? *PsyArXiv Preprints*. https://psyarxiv.com/u92cm/

Matsunaga, M. (2010). How to factor-analyze your data right: Do's, don'ts, and how-to's. *International Journal of Psychological Research, 3,* 97–110. http://dx.doi.org/10.21500/20112084.854

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

Miller, J., & Ulrich, R. (2013). Mental chronometry and individual differences: Modeling reliabilities and correlations of reaction time means and

effect sizes. *Psychonomic Bulletin & Review, 20,* 819–858. http://dx.doi.org/10.3758/s13423-013-0404-5

Nieuwenhuis, S., Stins, J. F., Posthuma, D., Polderman, T. J., 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

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

Nunnally, J. C. (1964). *Educational measurement and evaluation*. New York, NY: McGraw-Hill.

Osborne, J. W., Costello, A. B., & Kellow, J. T. (2008). Best practices in exploratory factor analysis. *Best Practices in Quantitative Methods, 10,* 86–99. http://dx.doi.org/10.4135/9781412995627.d8

Overall, J. E., & Woodward, J. A. (1975). Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin, 82,* 85–86. http://dx.doi.org/10.1037/h0076158

Paap, K. R., Anders-Jefferson, R., Mikulinsky, R., Masuda, S., & Mason, L. (2019). On the encapsulation of bilingual language control. *Journal of Memory and Language, 105,* 76–92. http://dx.doi.org/10.1016/j.jml.2018.12.001

Paap, K. R., & Sawi, O. (2014). Bilingual advantages in executive functioning: Problems in convergent validity, discriminant validity, and the identification of the theoretical constructs. *Frontiers in Psychology, 5,* 962. http://dx.doi.org/10.3389/fpsyg.2014.00962

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

Posner, M. I., & DiGirolamo, G. J. (1998). Conflict, target detection and cognitive control. In R. Parasuraman (Ed.), *The attentive brain* (pp. 401–423). Cambridge, MA: MIT Press.

Raven, J. C., & Court, J. H. (1998). *Raven's progressive matrices and vocabulary scales*. Oxford, UK: Oxford Psychologists Press.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.r-project.org/

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,* 501–526. http://dx.doi.org/10.1037/xlm0000450

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

Roberts, R. D., Goff, G. N., Anjoul, F., Kyllonen, P. C., Pallier, G., & Stankov, L. (2000). The armed services vocational aptitude battery (ASVAB): Little more than acculturated learning (Gc)!?. *Learning and Individual Differences, 12,* 81–103. http://dx.doi.org/10.1016/s1041-6080(00)00035-2

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48,* 1–36. http://dx.doi.org/10.18637/jss.v048.i02

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

Rouder, J., Kumar, A., & Haaf, J. M. (2019, March 25). Why most studies of individual differences with inhibition tasks are bound to fail. *PsyArXiv Preprints*. http://dx.doi.org/10.31234/osf.io/3cjr5

Salthouse, T. A. (2010). Is flanker-based inhibition related to age? Identifying specific influences of individual differences on neurocognitive variables. *Brain and Cognition, 73,* 51–61. http://dx.doi.org/10.1016/j.bandc.2010.02.003

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8,* 23–74. http://dx.doi.org/10.1016/j.tine.2018.11.003

Schmeichel, B. J., & Demaree, H. A. (2010). Working memory capacity and spontaneous emotion regulation: High capacity predicts self-enhancement in response to negative feedback. *Emotion, 10,* 739–744. http://dx.doi.org/10.1037/a0019355

Shipstead, Z., Harrison, T. L., & Engle, R. W. (2015). Working memory capacity and the scope and control of attention. *Attention, Perception & Psychophysics, 77,* 1863–1880. http://dx.doi.org/10.3758/s13414-015-0899-0

Shipstead, Z., Harrison, T. L., & Engle, R. W. (2016). Working memory capacity and fluid intelligence: Maintenance and disengagement. *Perspectives on Psychological Science, 11,* 771–799. http://dx.doi.org/10.1177/1745691616650647

Shipstead, Z., Lindsey, D. R., Marshall, R. L., & Engle, R. W. (2014). The mechanisms of working memory capacity: Primary memory, secondary memory, and attention control. *Journal of Memory and Language, 72,* 116–141. http://dx.doi.org/10.1016/j.jml.2014.01.004

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87,* 245–251. http://dx.doi.org/10.1037/0033-2909.87.2.245

Stevens, J. P. (2012). *Applied multivariate statistics for the social sciences* (5th ed.). http://dx.doi.org/10.4324/9780203843130

Stoffels, E. J., & van der Molen, M. W. (1988). Effects of visual and auditory noise on visual choice reaction time in a continuous-flow paradigm. *Perception & Psychophysics, 44,* 7–14. http://dx.doi.org/10.3758/BF03207468

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

Thurstone, L. L. (1938). *Primary mental abilities*. Chicago, IL: University of Chicago Press.

Tsukahara, J. S., Harrison, T. L., Draheim, C., Martin, J. D., & Engle, R. W. (2020). General discrimination ability: Why should performance on sensory tasks relate to cognitive ability? *Attention, Perception, & Psychophysics*. Advance online publication. http://dx.doi.org/10.3758/s13414-020-02044-9

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language, 28,* 127–154. http://dx.doi.org/10.1016/0749-596X(89)90040-5

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37,* 498–505. http://dx.doi.org/10.3758/BF03192720

Unsworth, N., Redick, T. S., Heitz, R. P., Broadway, J. M., & Engle, R. W. (2009). Complex working memory span tasks and higher-order cognition: A latent-variable analysis of the relationship between processing and storage. *Memory, 17,* 635–654. http://dx.doi.org/10.1080/09658210902998047

Unsworth, N., & Robison, M. K. (2016). Pupillary correlates of lapses of sustained attention. *Cognitive, Affective & Behavioral Neuroscience, 16,* 601–615. http://dx.doi.org/10.3758/s13415-016-0417-4

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,* 392–406. http://dx.doi.org/10.1016/j.jml.2010.02.001

Verhaeghen, P., & De Meersman, L. (1998). Aging and the Stroop effect: A meta-analysis. *Psychology and Aging, 13,* 120–126. http://dx.doi.org/10.1037/0882-7974.13.1.120

Vogel, E. K., McCollough, A. W., & Machizawa, M. G. (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature, 438,* 500–503. http://dx.doi.org/10.1038/nature04171

Wechsler, D. (1991). *WISC-III: Wechsler Intelligence Scale for Children: Manual*. San Antonio, TX: The Psychological Corporation. http://dx.doi.org/10.1038/nature04171

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica, 41,* 67–85. http://dx.doi.org/10.1016/0001-6918(77)90012-9

Zheng, Y., Myerson, J., & Hale, S. (2000). Age and individual differences in visuospatial processing speed: Testing the magnification hypothesis. *Psychonomic Bulletin & Review, 7,* 113–120. http://dx.doi.org/10.3758/BF03210729

Appendix A

Correlation Matrix for Attention Measures

Table A1
Zero-Order Intercorrelations Among the Attention Control Measures

	Threshold DV				Accuracy DV			Reaction time DV		
Task	1	2	3	4	5	6	7	8	9	10
1. Flanker DL	—									
2. Flanker PR	.38*	—								
3. Stroop DL	.26*	.30*	—							
4. AVC	.13*	.06	.08	—						
5. Antisaccade	.34*	.25*	.31*	.33*	—					
6. Visual arrays	.18*	.27*	.20*	.20*	.45*	—				
7. SACT	.25*	.27*	.29*	.27*	.40*	.31*	—			
8. PVT	.20*	.20*	.21*	.20*	.35*	.25*	.42*	—		
9. RT Stroop	.10	.09	.23*	.09	.19*	.18*	.01	.01	—	
10. RT flanker	.33*	.21*	.04	.04	.16*	.15*	.13*	.10*	.17*	—

Note. Pairwaise-deletion method was used for missing values, correlations are based on a range of $N = 382–390$. RT Stroop and RT flanker are reaction time difference scores (Stroop and flanker effect, respectively). For ease of interpretation, a positive correlation indicates better performance on both tasks. DL = deadline; PR = presentation rate; AVC = adaptive visual cue; SACT = sustained attention-to-cue; PVT = psychomotor vigilance task; RT Stroop = reaction time Stroop effect; RT flanker = reaction time flanker effect.
* $p < .05$.

(*Appendices continue*)

Appendix B

Description of Processing Speed Tasks

All processing speed measures were computerized versions of paper-and-pencil tests. In each case, participants were instructed to respond as quickly and accurately as possible, but consistent with standard administration procedures, were not alerted of the time limits of each task in the instruction phase.

Letter String Comparison

In this version of the letter string comparison task (Conway, Cowan, Bunting, Therriault, & Minkoff, 2002), participants viewed strings of three, six, or nine consonants appearing to the left and right of a central line. The letter strings could either be the same or differ by a single letter. If different, the mismatching letter could appear in any location in the string. Participants indicated their response by clicking on a button on the screen labeled *SAME* for identical strings or *DIFF* for mismatching strings. Letters were printed in white size 18-pt. Courier New font on a black background. After completing six practice trials, participants completed two 30-s blocks of the task. The dependent variable was the number of accurate responses across both blocks.

Digit String Comparison

The digit string comparison task was identical to the letter string comparison, except that participants viewed and made judgements about strings containing three, six, or nine digits.

Digit Symbol Substitution

This adaptation of the digit symbol substitution task (Wechsler, 1991) has been modified to make it more amenable to computer administration and response collection via a standard number pad. The symbols used are the same as the paper-and-pencil version of the task, and we endeavored to maintain the same basic demands. However, rather than viewing digits and reporting corresponding symbols, this task required participants to view symbols and report the corresponding digits. On each trial, participants were presented with two boxes stacked one on top of the other in the center of the screen. A symbol appeared in the bottom box. Participants were to consult a key presented at the top of the screen, and to indicate via key press with the digit that belonged in the top box. After 10 practice trials, participants completed 90 s of the task. The dependent variable was the number of correctly reported digits during that 90-s period.

(*Appendices continue*)

Appendix C

Additional Processing Speed Analyses

Table C1
Full Correlation Matrix for Analyses With Processing Speed

Task	RAPM	NumSeries	LetterSets	OSpan	SymSpan	RotSpan	Antisaccade	Flanker	FlankerDL	FlankerPR	Stroop	StroopDL	VA4	PVT	SACT	AVC	LetterComp	NumComp	DigitSymb
RAPM	—	0.520***	0.569***	0.435***	0.519***	0.506***	0.271**	−0.155	−0.247**	−0.126	−0.253**	−0.205*	0.316***	−0.101	0.124	−0.018	0.187*	0.272**	0.446***
NumSeries	0.520***	—	0.615***	0.385***	0.443***	0.482***	0.261**	−0.102	−0.201*	−0.046	−0.127	−0.077	0.282**	−0.020	0.078	0.147	0.178*	0.284**	0.279**
LetterSets	0.569***	0.615***	—	0.374***	0.499***	0.462***	0.292***	−0.170	−0.188*	−0.015	−0.231**	−0.083	0.310***	−0.117	0.245**	0.109	0.396***	0.444***	0.450***
OSpan	0.435***	0.385***	0.374***	—	0.537***	0.411***	0.199*	−0.085	−0.144	−0.047	−0.061	−0.256**	0.222*	0.102	0.049	0.104	0.105	0.123	0.216*
SymSpan	0.519***	0.443***	0.499***	0.537***	—	0.670***	0.370***	−0.234**	−0.296***	−0.264**	−0.217*	−0.255**	0.488***	−0.048	0.245**	0.056	0.257**	0.285**	0.357***
RotSpan	0.506***	0.482***	0.462***	0.411***	0.670***	—	0.359***	−0.233**	−0.327***	−0.277**	−0.253**	−0.159	0.402***	−0.151	0.118	0.104	0.106	0.169	0.266**
Antisaccade	0.271**	0.261**	0.292***	0.199*	0.370***	0.359***	—	−0.179*	−0.242**	−0.263**	−0.219*	−0.310***	0.390***	−0.202*	0.337***	−0.289***	0.281**	0.397***	0.293***
Flanker	−0.155	−0.102	−0.170	−0.085	−0.234**	−0.233**	−0.179*	—	0.450***	0.183*	0.270**	0.020	−0.104	−0.034	−0.134	0.101	−0.196*	−0.217*	−0.146
FlankerDL	−0.247**	−0.201*	−0.188*	−0.144	−0.296***	−0.327***	−0.242**	0.450***	—	0.449***	0.156	0.280**	−0.075	0.069	−0.155	0.052	−0.125	−0.155	−0.250**
FlankerPR	−0.126	−0.046	−0.015	−0.047	−0.264**	−0.277**	−0.263**	0.183*	0.449***	—	0.119	0.391***	−0.174*	0.294***	−0.235**	0.035	−0.025	−0.056	−0.127
Stroop	−0.253**	−0.127	−0.231**	−0.061	−0.217*	−0.253**	−0.219*	0.270**	0.156	0.119	—	0.210*	−0.136	0.003	−0.125	0.038	−0.180*	−0.250**	−0.175*
StroopDL	−0.205*	−0.077	−0.083	−0.256**	−0.255**	−0.159	−0.310***	0.020	0.280**	0.391***	0.210*	—	−0.153	0.092	−0.292***	0.029	−0.086	−0.156	−0.178*
VA4	0.316***	0.282**	0.310***	0.222*	0.488***	0.402***	0.390***	−0.104	−0.075	−0.174*	−0.136	−0.153	—	−0.252**	0.233**	−0.234**	0.315***	0.384***	0.352***
PVT	−0.101	−0.020	−0.117	0.102	−0.048	−0.151	−0.202*	−0.034	0.069	0.294***	0.003	0.092	−0.252**	—	−0.414***	0.268**	−0.198*	−0.176*	−0.151
SACT	0.124	0.078	0.245**	0.049	0.245**	0.118	0.337***	−0.134	−0.155	−0.235**	−0.125	−0.292***	0.233**	−0.414***	—	−0.278**	0.284**	0.246**	0.190*
AVC	−0.018	0.147	0.109	0.104	0.056	0.104	−0.289***	0.101	0.052	0.035	0.038	0.029	−0.234**	0.268**	−0.278**	—	−0.134	−0.145	0.038
LetterComp	0.187*	0.178*	0.396***	0.105	0.257**	0.106	0.281**	−0.196*	−0.125	−0.025	−0.180*	−0.086	0.315***	−0.198*	0.284**	−0.134	—	0.717***	0.255**
NumComp	0.272**	0.284**	0.444***	0.123	0.285**	0.169	0.397***	−0.217*	−0.155	−0.056	−0.250**	−0.156	0.384***	−0.176*	0.246**	−0.145	0.717***	—	0.368***
DigitSymb	0.446***	0.279**	0.450***	0.216*	0.357***	0.266**	0.293***	−0.146	−0.250**	−0.127	−0.175*	−0.178*	0.352***	−0.151	0.190*	0.038	0.255**	0.368***	—

Note. Correlation matrix for the subset of participants who returned for an additional session of tasks, including three processing speed tasks. RAPM = Raven; NumSeries = number series; OSpan = operation span; SymSpan = symmetry span; RotSpan = rotation span; FlankerDL = flanker with adaptive response deadline; FlankerPR = flanker with adaptive presentation rate; StroopDL = Stroop with adaptive response deadline; VA4 = selective visual arrays; PVT = psychomotor vigilance task; SACT = sustained attention-to-cue; AVC = adaptive visual cue; LetterComp = letter string comparison; NumComp = number (digit string) comparison; DigitSymb = digit symbol substitution. $N = 173$.
* $p < .05$. ** $p < .01$. *** $p < .001$.

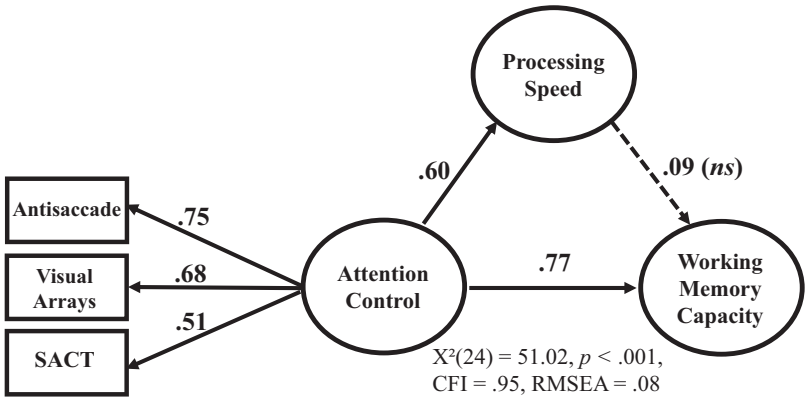

Figure C1. Processing speed mediating the relationship between attention control and working memory capacity.

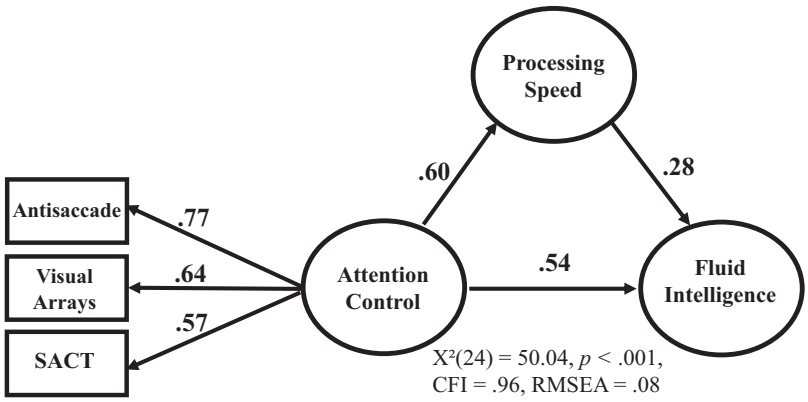

Figure C2. Processing speed mediating the relationship between attention control and fluid intelligence.

We tested whether processing speed mediated the relationship between attention control and either working memory capacity or fluid intelligence separately. Table C1 has the full correlation matrix for the subset of participants who were administered the processing speed tasks. The attention factor consisted of the antisaccade, visual arrays, and sustained attention-to-cue. In both these models (see Figure C1 and Figure C2), attention control predicted a substantial amount of variance in working memory capacity (64%) and fluid intelligence (33%) above and beyond processing speed. Processing speed's paths to working memory capacity and fluid intelligence were also not significant when accounting for attention control, indicating that processing speed did not at all mediate the relationship between attention control and working memory capacity or fluid intelligence. Because previous results indicated that the visual arrays had a strong relationship with working memory capacity and fluid intelligence, we wanted to ensure that this particular task was not unduly influencing the results. So, we tested models using the adaptive flanker response deadline task in place of the visual arrays, and the results were similar.

(*Appendices continue*)

Appendix D

The Visual Arrays was Required to Fully Mediate the Working Memory Capacity/Fluid Intelligence Relationship

Although attention control fully mediated the working memory capacity/fluid intelligence relationship, this is only true when visual arrays was included as one of the indicators of attention. The modified and new attention tasks did not achieve full mediation alone, or even with the antisaccade. There are a number of potential and non-mutually exclusive explanations for this which are explored in further detail in the following text.

One possibility is that the full mediation occurs with selection visual arrays because the visual arrays is in fact a measure of visual working memory capacity and not attention. If so, then the full mediation is simply due to misspecification and is not theoretically meaningful. We cover the extensive lines of evidence for the visual arrays as a measure of attention in Martin, Tsukahara, et al. (2019), but one simple test of whether the visual arrays loads more with attention control or working memory capacity is to compare the visual arrays factor loadings in a model in which it is cross-loaded onto both working memory capacity and attention control. This model is shown in Figure D1, and the results are quite clear—the selective visual arrays used in this study loaded with attention control (namely antisaccade) strongly and significantly (.60) and did not load with working memory capacity (.10, non-significant). This rather straightforward model illustrates that individual differences in the specific version of the visual arrays task that was used here (with a selection demand) reflected attention control more so than working memory capacity.[11] Relatedly, if the full mediation of attention control on the working memory capacity/fluid intelligence relationship required visual arrays because selective visual arrays is a working memory task, then we would see a unique contribution of variance from visual arrays to working memory capacity when the variance common to the attention tasks was partialed out. This model is shown in Figure D2. The selective visual arrays shared 44% of its variance with the attention control factor but it did not predict any statistically significant variance in working memory capacity above and beyond attention control and, conversely, these attention measures contributed strong variance (38%) to working memory capacity above and beyond visual arrays. As such, it does not appear that the mediation of attention control on the working memory capacity/fluid intelligence relationship when the visual arrays is included as a measure of attention control is due to misspecification or redundancy of indicators (i.e., using a working memory measure to mediate a relationship involving working memory capacity).

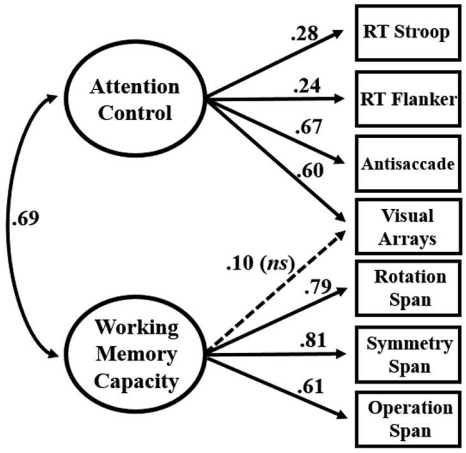

Figure D1. Confirmatory factor analysis with visual arrays crossloaded onto WMC and AC. RT Stroop and RT flanker are the mean differences in reaction time on congruent and incongruent trials. Paths are reported as positive if better performance on one task or factor was associated with better performance on. The factor loadings for the RT Stroop and RT flanker are unacceptably poor, however they were used as indicators of attention control here for the sake of demonstration and because this is a more standard factor of attention control (N = 396).

[11] In Martin, Tsukahara, et al. (2019), we found that these results hold when working memory capacity is measured by tasks other than complex span as well which is replicated over three other large-scale data sets.

(*Appendices continue*)

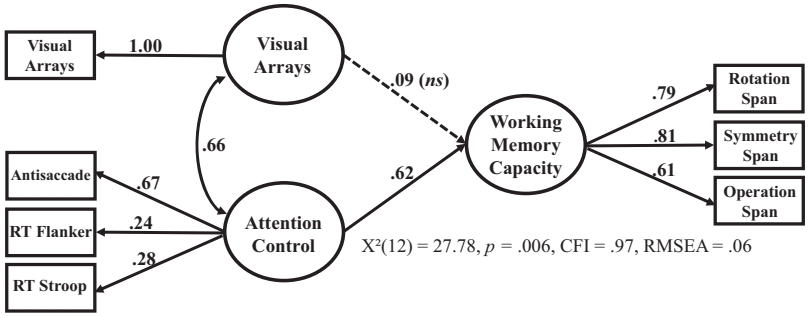

Figure D2. Structural equation model testing whether visual arrays contributes unique variance to working memory capacity above and beyond attention control. RT Stroop and RT flanker are the mean differences in reaction time on congruent and incongruent trials. Paths are reported as positive if better performance on one task or factor was associated with better performance on. The factor loadings for the RT Stroop and RT flanker are unacceptably poor, however they were used as indicators of attention control here for the sake of demonstration and because this is a more traditional factor of attention control ($N = 396$).

One speculative explanation is that the full mediation of the working memory capacity/fluid intelligence relationship requiring visual arrays was not due to specific processes in the visual arrays, but rather was due to visual arrays improving the attention factor as a whole. The question is, is there something special about the visual arrays itself? Or perhaps the selective visual arrays results in a stronger attention factor, as a whole, due to the manifestation of more theoretically relevant shared variance across the attention tasks. Further research is required to better explore this idea, but we did test one additional model in which the selective visual arrays was the sole indicator of attention control and mediated the working memory capacity/fluid intelligence relationship. The logic here was that if there was something special about the visual arrays by itself then the visual arrays should mediate the working memory capacity/fluid intelligence relationship completely (or near completely) by itself. The model (see Figure D3) shows that the selective visual arrays task was a weak mediator of the working memory capacity/fluid intelligence relationship on its own (the direct path between working memory capacity and fluid intelligence is .58), only achieving a partial mediation. We replicated these results in a reanalysis of Shipstead et al. (2014) in which we used two selective visual arrays tasks as the mediator of the working memory capacity/fluid intelligence relationship with working memory capacity measured using two complex span and two running span tasks: The direct path between working memory capacity and fluid intelligence was .70 after accounting for the mediation of the selective visual arrays tasks. This finding speaks not only to the possibility that the selective visual arrays is improving the attention factor in previous models, but again that visual arrays is not the sole cause of the full mediation of attention control onto the working memory capacity/fluid intelligence relationship. The full mediation was not due to any single task individually but rather was due to the common variance across the tasks.

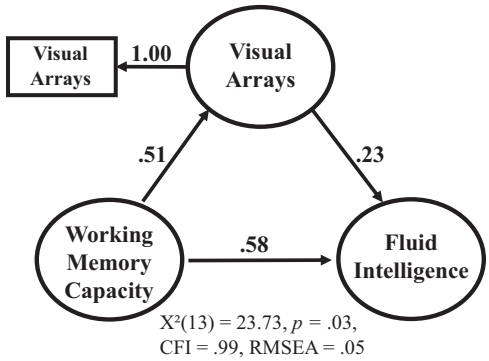

Figure D3. Structural equation model testing whether visual arrays solely mediates the working memory capacity/fluid intelligence relationship. Working memory capacity comprising operation span, symmetry span, and rotation span with respective loadings of .62, .80, and .79; fluid intelligence comprised Raven, number series, and letter sets with respective loadings of .75, .72, and .73 ($N = 396$).

Received April 2, 2019
Revision received March 17, 2020
Accepted March 22, 2020 ■