Spring 2021

Dual Mechanisms of Cognitive Control: A Hierarchical Bayesian Approach to Test-Retest Reliability

Jean-Paul Snijder
Claremont Graduate University

Dual Mechanisms of Cognitive Control:

A Hierarchical Bayesian Approach to Test-Retest Reliability

By

Jean-Paul Snijder

Claremont Graduate University

2021

Approval of the Dissertation Committee

This dissertation has been duly read, reviewed, and critiqued by the Committee listed below, which herby approves the manuscript of Jean-Paul Snijder as fulfilling the scope and quality requirements for meriting the degree of Doctor of Philosophy in Psychology with a concentration in Cognitive Psychology.

Andrew Conway, Chair

Claremont Graduate University

Professor of Psychology

Gabriel Cook

Claremont McKenna College

Associate Professor of Psychology

Megan Zirnstein

Pomona College

Assistant Professor of Linguistics and Cognitive Science

Claudia von Bastian

The University of Sheffield

Lecturer

Abstract

Dual Mechanisms of Cognitive Control:

A Hierarchical Bayesian Approach to Test-Retest Reliability

by
Jean-Paul Snijder

Claremont Graduate University: 2021

Cognitive control, also known as attentional control or executive function, is a set of fundamental processes that are utilized in a wide range of cognitive functioning: including working memory, reasoning, problem solving, and decision making. Currently, no existing theory of cognitive control unifies experimental and individual differences approaches. Some even argue that cognitive control as a psychometric construct does not exist at all. These disparities may exist in part because individual differences research in cognitive control utilizes tasks optimized for experimental effects (i.e., Stroop effect). As a result, many cognitive control tasks do not have reliable individual differences despite robust experimental effects (Hedge, Powell, & Sumner, 2018). In the current study, we examine the efficacy of a new task battery based on the Dual Mechanisms of Cognitive Control theory (DMCC; Braver, 2012) to provide reliable estimates of individual differences in cognitive control. With two sets of analyses, the first traditional (e.g., split-half, ICC, and rho), and the second hierarchical Bayesian, we provide evidence that (1) reliable individual differences can be extracted from experimental tasks, and (2) weak correlations between tasks of cognitive control are not solely caused by the attenuation of unreliable estimates. The implications of our findings suggest that it is unlikely that poor measurement practices are the cause of the weak between-task correlations in cognitive control, and that a psychometric construct of cognitive control should be reconsidered.

Table of Contents

1 Introduction

Cognitive control is a construct used to refer to the set of processes involved in deliberate regulation of information processing to facilitate goal-directed behavior (Miller & Cohen, 2001; Posner & Snyder, 1979). Cognitive control is associated with several important real-world outcomes including psychopathology (Snyder et al., 2015), impulsivity (Sharma et al., 2014), addiction (Hester & Garavan, 2004), and age-related cognitive declines (Hasher et al., 1991). Also, the ability to engage cognitive control is a strong predictor of working memory capacity (WMC), which is associated with a broad range of outcomes, including academic achievement (Alloway & Alloway, 2010; Gathercole et al., 2003), reading comprehension (Daneman & Carpenter, 1980), mathematical ability (Ramirez et al., 2016), and multi-tasking (Redick et al., 2016). Cognitive control also plays an important role in contemporary theories of intelligence. By many accounts, cognitive control is considered to be the primary source of variance in overall cognitive ability (Engle & Kane, 2004; Kovacs & Conway, 2016; Van Der Maas et al., 2006). The association between WMC and intelligence has been suggested to arise due to both processes' reliance on the ability to regulate attention (e.g., cognitive control) in order to ignore distractors (Engle, 2002, 2018; Kane et al., 2007). Improvement in cognitive control ability therefore provides a boost to intelligence and allows for the differentiation of cognitive abilities (Kovacs & Conway, 2016).

However, there are some problems with the construct validity of cognitive control. First of all, there is the inconsistency of definitions. Cognitive control is also referred to as executive function, executive control, controlled attention, and attentional control (Diamond, 2013). Even the postulated processes within the construct are referred to by different names (Rey-Mermet et al., 2018). Dempster (1993) refers to processes within inhibition as *control of perceptual*

interference, *control of motor interference*, and *control of verbal-linguistic interference*. In comparison, Friedman and Miyake (2004) refer to these processes as *resistance to distracter interference*, *inhibition of prepotent responses*, and *resistance to proactive interference*, and Hasher et al. (2007) use *access*, *restraint*, and *deletion*.

Second, cognitive control research is plagued by inconsistencies in hypothesized latent variables and factor structure. Friedman and Miyake (2004) found that their postulated inhibition processes *resistance to distractor interference* (i.e., the ability to ignore distracting external information) and *inhibition of prepotent responses* (i.e., suppressing dominant responses) were highly correlated (r = .67). The authors found the strength of this relationship enough evidence to combine them into a single latent construct, namely, *response-distractor inhibition*. Pettigrew and Martin (2014) found support for this combined construct, however, in the same year Stahl et al. found support for the separability of distracter- and response-related interference. Some studies found that tasks often used to measure inhibition (e.g., Eriksen and Simon tasks) had no common variance, and thus, that the variance was task-specific (Keye et al., 2009; Wilhelm et al., 2013). Other studies were not able to extract an inhibition factor at all (Klauer et al., 2010; Krumm et al., 2009; van der Sluis et al., 2007).

Another common issue with cognitive control is that correlations between measures of cognitive control are often weak. Considering that these tasks are assumed to tap the same psychometric construct, this is counterintuitive. Here is a non-exhaustive but representative sample of studies reporting correlations between tasks of cognitive control. Friedman and colleagues (2016): Stroop and Antisaccade, r = .17; Stroop and Stop-signal, r = .15; Antisaccade and Stop-signal, r = .26. Gustavson and colleagues (2018): Stroop and AX-CPT, r = .16; Stroop and Category Switch, r = .06; AX-CPT and Category Switch, r = .12. Paap and Greenberg

(2013): Antisaccade and Simon effect, r = -.12; Flanker effect and Simon effect, r = -.01. Antisaccade was not run in the same study as the Flanker test, hence no reported correlation. Finally, Rey-Mermet and colleagues (2019) reported 21 correlations between Number Stroop, Arrow Flanker, Letter Flanker, Simon, Antisaccade, and Stop-signal: ranging from r = -.16 to r = .15. Finally, in a large meta-analysis of 70 studies producing 2,114 between-task correlations, von Bastian et al., (2020) report a median correlation of .16, with most studies not surpassing between-task correlations of .30.

Taken together, the inconsistencies in nomenclature and processes, and the weak between-measure correlations have a sparked a debate on whether a coherent psychometric construct of cognitive control even exists. One side argues that poor correlational results between these tasks indicate that cognitive control is perhaps not a coherent psychometric construct (Paap & Sawi, 2016; Rey-Mermet et al., 2018). Another side argues that many of these inconsistent and poor correlational results stem from measurement issues and poor operationalization of the construct (Draheim et al., 2020; Hedge, Powell, & Sumner, 2018). Given the practical and theoretical importance of cognitive control in clinical applications and in theories of working memory and intelligence outlined above, these measurement issues warrant further examination before discarding the construct as a whole. After an encapsulation of important theories that laid the groundwork for research in cognitive control, measurement of cognitive control will be discussed.

2 Theories of Cognitive Control

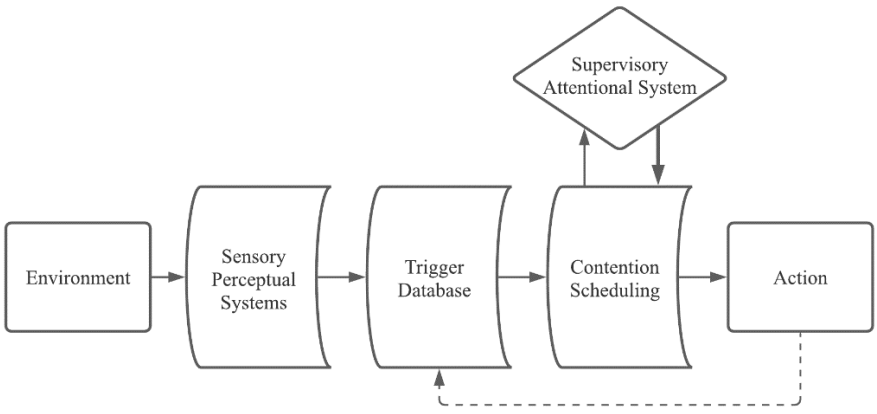
2.1 Norman and Shallice (1986)

Early research on cognitive control focused on inhibition of automatic behavior in novel situations. In one of the first theories to venture outside of standard box models and into dynamic

neurocognitive models, Norman and Shallice proposed a cognitive control model of executive functioning (1986) (See Figure 1). In their seminal work, they outline how schemas are activated to instruct behavior. A schema, in this sense, is a collection of sequential thoughts and actions triggered by perceptual stimuli. Schemas can be subdivided into two groups; schemas for automated and for controlled processes. Automated, or routine, processes are defined here as processes that are engaged when performing a task or action that do not require attentional resources (Norman & Bobrow, 1975); for example, riding a bike after many years of practice. Controlled processes are tasks or actions that require deliberate and sustained attention. These are often employed when important situations, dangerous environments, or novel problems present themselves and navigating behavior successfully is paramount. For example, riding a bike for the first time in new city during rush hour in a country where they drive on the opposite side of the road.

Figure 1

Schematic model of the cognitive control model of executive function



Norman and Shallice (1986) propose that schemas are activated by a process called contention scheduling (CS). CS ensures that the "optimal" schema is activated and inhibits incorrect, or less efficient, schemas from contending. Here, optimal is considered to be

subjective. If a schema is repeatedly activated given a certain event, its threshold for activation is lowered and considered optimal for similar future events. The CS mechanism is suggested to work relatively straightforward; it allows schemas to compete for activation and it activates a schema once it reaches its activation threshold. The main role of CS is to subconsciously monitor for automated and routine situations. When a situation requires a more controlled and conscious decision, for example when it is novel or complex, the supervisory attentional system (SAS) is engaged. When engaged, the SAS controls the CS by influencing schema activation thresholds and activating existing schemas to novel problems (e.g., using existing strategies in situations for which a schema does not exist). The SAS does not directly control action and decision making, but rather the thresholds for activation and inhibition of competing schemas.

As the more controlled construct, the SAS is an early mechanism for what now is dubbed "cognitive control". The account by Norman and Shallice was a frontrunner for future research on cognitive control, however, it was not yet specific enough. For one, it was too broad in its description of processes related to the SAS, which made distinguishing SAS from other constructs, such as intelligence, difficult. It is important to note that the failure to specify processes of cognitive control is a common theme throughout the lifespan of this research.

2.2 Shallice and Burgess (1996)

Shallice and Burgess were the first to explore whether the Norman and Shallice (SAS) is fractionable into different sub-processes (1996). For about two decades, scholars generally agreed that the prefrontal cortex housed an important central ability that influenced multiple domains, but this ability was mainly characterized as a single type of process. Some suggested that this singular process was what underlies general intelligence, or *g*, and most commonly were accounts of working memory add a Duncan reference here (Duncan et al., 1996; Engle et al.,

1995; Kimberg & Farah, 1993). Other unitary accounts existed as well, but the idea that the prefrontal cortex carried only a single key process was the overarching school of thought. Alternatively, Shallice and Burgess suggested that even if the SAS was a single system, it was incorrect to view it as carrying out only a single process. The authors showed evidence for "*the existence of a variety of processes carried out by different subsystems but operating together to have a globally integrated function*". They observed "*very low correlations across patients on more than one measure…*", and argued for the separability of processes stemming from the prefrontal cortex. The hypothesis that a central attention system is a multi-process system is what eventually lead to contemporary theories of cognitive control. However, note that weak correlations between measures of cognitive control played an important role in advancing research of cognitive control, yet currently are also the reason for the suggestion that cognitive control might not be a valid construct.

2.3 Miller and Cohen (2001)

At the beginning of this century, Miller and Cohen published the important *Integrative Theory of Prefrontal Cortex Function* (2001). The authors argue that cognitive control is the main function of the prefrontal cortex (PFC) and consists of different processes such as selective attention, error monitoring, decision-making, and inhibitions of stimuli and response. They not only specified a set of cognitive control sub-processes, but also suggested mechanisms by which those sub-processes are executed, and provided a review of neurobiological evidence that supported their theory.

The integrative theory of prefrontal cortex function states that the PFC is critical for carrying out processes that require top-down processing; that is, when behavior benefits greatly from being intentionally controlled based on internal representations, or mappings, of a goal. For

example, if one's goal is to reach the other side of a busy road safely, then observing and timing bi-directional traffic, locating a pedestrian crossing, or alternatively, determining the presence or absence of law enforcement, are all behaviors that need to be actively controlled in order to complete the goal. Such processes are often aimed to be operationalized by cognitive-behavioral tasks. For example, in the Stroop task (Stroop, 1935) subjects are presented with words that are names of colors (e.g., the word "GREEN") in different colored fonts (e.g., green, yellow, red). Here, correct performance, and hence the internalized goal, is based on the task rule to name the font-color rather than to verbalize the color-word. Generally, reading of a simple word is a bottom-up, automated, and a much practiced process and hence, has strong existing mappings between reading the word, semantic processing, and verbalizing it. According to Miller and Cohen's theory, the PFC is not critical in such processes. However, in an incongruent trial, the Stroop task requires that subjects inhibit the tendency of such reading and then selectively attend to the color of the font. Inhibiting and selectively attending are considered processes with weaker existing mappings than the automated reading, and hence, are suggested by Miller and Cohen to require the PFC to control behavior for a correct performance of the task.

The theory also describes *how* the PFC controls behavior. The theory builds on an earlier principle by Desimone and Duncan (1995) which states that multiple available behaviors exist simultaneously and compete to be executed. Neurobiologically, the executed behavior is the behavior downstream of the most excited neural pathway. According to Miller and Cohen, cognitive control is the voluntary biasing of pathway towards the behavior that best fits current task-goals. The PFC resolves competition by inhibiting pathways of alternative behaviors and exciting the pathways to the preferred behavior. This preferential activation establishes the mappings needed to execute the goal-appropriate behavior. Miller and Cohen's description of

how the PFC controls behavior, can be viewed as the neural implementation of the internal and external rules and goals governing controlled behavior.

Finally, Miller and Cohen review neurobiological evidence for the distinction between a system that regulates routine behavior (contention scheduling) and a system for deliberate conscious control (supervisory attentional system). Their strongest evidence relies on the notion that if the PFC is crucial in the SAS, but not in the CS, then damage to PFC would impact the ability to control behavior, but not routine behavior. In their manuscript, the authors present prior findings following that logic. First, it is generally known in clinical neuropsychology that damage to the prefrontal structures leave execution of basic skills unaffected (Walsh, 1978). Second, performance on WAIS subtests is also relatively unaffected by frontal lesions (McFie, 1960). Third, contrasting evidence can be found in Lhermitte et al., (1972). This classic study showed that two patients with frontal lobe lesions were able to complete the verbal and performance WAIS tasks at normal levels. However, their performance on the WAIS Block Design or the reproduction of a complex figure (i.e., Figure of Rey) was extremely poor. These are tasks that require more controlled processes such as novel programming, planning, and problem solving. For more evidence, see Norman and Shallice (1986).

2.4 Miyake et al. (2000)

Miller and Cohen's theory provided a framework based on neurobiological evidence that supported mechanistically explicit hypotheses about processes of cognitive control. Around the same time, Miyake and colleagues (2000) published their influential (~ 13,000 citations currently) *unity/diversity* framework of cognitive control, though their theory was based on cognitive-behavioral evidence. Their seminal work employed structural equation modeling on tasks of attention and found convergent validity for a model with multiple correlated first-order

constructs. More specifically, three latent factors emerged: *shifting* between task sets, *updating* and monitoring of information, and *inhibition* of prepotent responses. Their continued work (Friedman et al., 2008; Friedman & Miyake, 2017; Miyake & Friedman, 2012) has shown an additional higher-order common factor (i.e., cognitive control) that accounts for covariance across the latent factors. Friedman and Miyake (2017) suggest that this common factor reflects active goal maintenance and top-down biasing of attention. Yet, despite the popularity of this unity/diversity model, it seemingly has one weakness; the latent factor termed inhibition by Miyake and colleagues is not consistently evident.

In addition to the problems with the inhibition factor that have already been mentioned in the introduction (e.g., no common variance between inhibition tasks; no inhibition latent factor found at all), there are some additional concerns with the unity/diversity framework worth mentioning here. One, a meta-analytic review showed that out of the studies that found support for these models, only a few tested other models (Karr et al., 2018). Two, the same meta-analytic review by Karr and colleagues reported that many studies based on this theory suffered from low rates of model acceptance and model selection. These issues were generally attributed to the small sample sizes, high model complexities, and poor reliability of the experiments. Three, even though the factor-loadings of updating and shifting are often found to be acceptable (Ecker et al., 2010; Singh et al., 2018; von Bastian & Druey, 2017), they are found to be weak for the inhibition factor (Friedman & Miyake, 2004; Gustavson et al., 2018; Hedge, Powell, Bompas, et al., 2018; Paap & Greenberg, 2013; Rey-Mermet et al., 2019). And four, latent factors of inhibition are often dominated by a single task (typically, the anti-saccade task (Rey-Mermet et al., 2019)).

In conclusion, Miyake et al. (2000) promote a model of cognitive control with three emerging (latent) processes; shifting, updating, and inhibition. Like Miller and Cohen (2001), Miyake et al. present evidence that cognitive control consists of different processes underlying a central function. However, a wide-scale review of studies employing this unity/diversity model reveals mixed evidence for its construct validity. As previously mentioned, such inconsistent and weak individual differences results perhaps stem from measurement issues.

3 Measurement of Cognitive Control

3.1 Experimental Approach

Cognitive-behavioral tasks are designed to measure whether theorized processes indeed manifest behaviorally. Often, existing tasks are adapted to find behavioral evidence of such processes. Cognitive control is often measured by tasks that through experimental manipulations create two or more trial types. A baseline trial generally presents the subject with low (or no) conflict and correct performance does not require much cognitive control. An experimental trial includes a manipulation which causes interference, and hence, requires cognitive control to resolve this interference, necessary for correct performance. A classic example of this is the Stroop task (Stroop, 1935). There are generally two types of trial in the Stroop task. In the non-interference trial, more commonly known as a congruent trial in the Stroop task, the color-names match the font-color in which they are presented (e.g., “GREEN” in a green font-color). In contrast, in the interference trail, referred to as an incongruent trial, the color-names do not match the font-color (e.g., “GREEN” in a red font-color). Correct performance in an incongruent trial requires the subject to resolve the conflict that arises between the reading of the color-word and the naming font-color. The Stroop effect is considered the decrement in incongruent

performance (lower accuracy and slower reaction time) when compared to congruent performance.

Experimental effects of cognitive control are numerous and varied, including the congruency or conflict effect (Stroop, 1935), response inhibition (Hallett, 1978), error-related slowing (Rabbitt, 1966), sequential congruency effects (Gratton et al., 1992), costs of switching between tasks or completing multiple tasks simultaneously (Koch et al., 2018), and monitoring and updating of information (Miyake et al., 2000). Each of these paradigms reveals robust and reliable experimental effects that are considered signatures of cognitive control. Together, these experimental effects and other "benchmark findings" in psychology and neuroscience help to establish and test theories and models of cognitive control, which in turn help to guide further investigation and research.

Yet, models based on experimental evidence alone are limited; they provide an account of normative behavior but do not explain individual differences in cognitive control. A unified approach to the study of cognitive control would require a combination of experimental and correlational methods (Cronbach, 1957). Ideally, the experimental and correlational approaches inform each other, allowing for a theoretical framework that integrates different kinds of empirical evidence and accounts for inter-individual differences in terms of intra-individual psychological processes.

3.2 Individual Differences Approach

As mentioned in the introduction, the weak correlations between the measures of cognitive control indicate that the individual differences dimension has not been successfully integrated. One difference with other areas of individual differences research such as personality or intelligence, is that those areas use tasks that were designed to measure ability differences

between people. Correlational approaches in cognitive control research have mostly consisted of comparing effects from experimental measures. Because of the robustness of these effects, differences in the size of the effects were assumed to reflect a general cognitive control ability (von Bastian et al., 2020). This brings us to the main concern addressed in this paper; can measures of cognitive control derived from experimental tasks be used to explain individual differences in control ability?

This is not an entirely new concern, in fact, it is based on a longstanding concern in Psychology. In 1957, Cronbach famously quipped, "*Individual differences have been an annoyance rather than a challenge to the experimenter*" (Cronbach, 1957, p. 674). Recently there seems to be an increased awareness of psychometric issues when employing experimental cognitive tasks to measure individual differences. Here, three of those issues are discussed: (a) the poor reliability of existing measures, (b) the use of difference scores in individual differences studies, and (c) the effect of reliability on correlations between measures.

One popular account by Hedge, Powell, and Sumner (2018), aptly titled "[t]he reliability paradox…" examines the phenomenon that robust cognitive-behavioral tasks do not produce reliable individual differences measures. They report the test-retest reliabilities of 7 classic experimental effects (e.g., Stroop, flanker) used in cognitive psychology and neuroscience. To summarize their results, their reliabilities were generally weak with a median ICC = .40. Their investigation clearly illustrates that experimental effects in cognitive control tasks are robust, yet the test-retest reliabilities are weak and in many cases are not reliable or only moderately reliable.

One explanation for this paradox is that the meaning of "reliable" is different in experimental vs. correlational psychology. An experimental manipulation is "reliable" when the

intended effect is replicated across multiple studies (in different labs, with different stimuli, etc.). In contrast, an individual differences measure is considered reliable when it ranks subjects consistently in terms of the effect size. Experimental reliability is best served by *low* between-subject variance (e.g., homogeneous measures) and high within-subject variance (e.g., a large effect due to a manipulation). Contrarily, correlational reliability is best served by *high* between-subject variance (e.g., heterogeneous measures), making it easier to tease apart performance which is preferable for finding individual differences. Critically, between-subject variance is considered measurement error in experimental designs and typically are designed to minimize this noise (Burgess, 1997). As a result, these tasks may be not be suitable as reliable measures in individual differences research (Hedge, Powell, & Sumner, 2018).

Another measurement issue stems from the popular use of difference scores in cognitive-behavioral tasks. Many classic effects (e.g., the Stroop effect) are a simple difference score based on the contrast between the two experimental task trial types (e.g., Stroop effect = incongruent – congruent trial performance). However, from a psychometric perspective, difference scores are notoriously problematic for reliability (Caruso, 2004; Cronbach & Furby, 1970; Lord, 1956). It is common knowledge that the reliability of a difference score is not as robust as the reliability of its components (Edwards, 2001; Rogosa, 1988, 1995; Willett, 1988; Zimmerman & Williams, 1998; Zumbo, 1999). This is a purely psychometric phenomenon. Generally, when taking the difference between two measures, the amount of between-subject variance is lowered, but the measurement error is relatively unaffected. Hence, the ratio of measurement error to between-subject variance increases. Lower between-subject variance increases experimental reliability (less error or "noise"). Critically however, it *decreases* the reliability of individual differences because lower between-subject variability makes it more difficult to separate the performance of

subjects when ranking them. Thus, difference scores are well suited in experimental, but not in individual differences research. For an applied illustration, see Rodebaugh et al., (2016).

Finally, Spearman (1904) noted that measurement error attenuates the maximum attainable correlation between two measures. A "true" between-measure correlation of, for example, .80, can only be attained if both measures are free of measurement error. The reliability coefficient essentially reflects the measurement error in each measure individually, hence, a correlation between two measures is constrained by the average of their individual reliabilities (Hedge, Powell, & Sumner, 2018; Nunnally Jr., 1970; Spearman, 1904).

To summarize three main issues with the measurement of cognitive control: (a) experimental tasks are used to research individual differences, but are designed to minimize between-subject variance, which oftentimes causes poor reliability of the measures; (b) the popular use of difference scores further accentuates the issue of poor reliability, because it increases the ratio of measurement error to between-subject error; and (c) a correlation between two measures is constrained by the reliability of each measure; if individual differences in cognitive control are not reliable, correlational results are difficult to interpret.

4 Reliability

These reliability issues are not just a problem in cognitive science (Parsons et al., 2019); so how is it possible that they have not been noticed on a grand scale? For one, reliability estimates are not always reported; this may lead to task reliability not being considered as one of the suspects of poor correlational results (Flake et al., 2017; Hussey & Hughes, 2018). Consequently, some results may have been erroneously reported as replicable and generalizable, perhaps propagating false standards in the field (e.g., the replication crisis). Furthermore, when reliabilities are reported it is not always accompanied by how they were calculated. Here, some

common pitfalls of calculating reliability are presented and a standard methodological approach in reporting individual differences reliability is discussed.

4.1 Internal Consistency

There are many ways to estimate reliability and currently there is no standardized procedure (Parsons et al., 2019). Additionally, the reliability methods offered in many statistical software packages assume that the data conforms to analysis-specific assumptions. For example, a common and well-known method for estimating reliability is Cronbach's alpha, a measure of internal consistency. Alpha is most commonly derived by averaging the correlations between each item (trial) and the sum of the remaining items (trials). The default method offered in statistical software packages calculates alpha based on the assumption that the order of the items is identical for all subjects. Furthermore, it is assumed that each item measures the same underlying construct, to varying degrees, as a function of item difficulty and discriminability. In survey research, this is often the case. However, in cognitive-behavioral tasks, trial order is often randomized. More concerning, the cognitive processes involved in task performance may vary across trials, as a function of practice, fatigue, or strategy development/deployment, or even due to the experiment's own manipulations. If these issues are ignored, which is typically the case, then alpha reliability estimates may not be accurate nor valid. Hence, standard Cronbach's alpha is generally unsuitable as an index of reliability for tasks designed to measure individual differences in cognitive control.

Alpha can also be calculated as the average of correlations between two halves of the data (e.g., split-half reliability). Most commonly, the data are split into the first and second half or even- and odd-numbered trials. However, it has been demonstrated that split-half reliabilities based on these kinds of simple split methods are unstable (Enock et al., 2014). It is recommend

to apply multiple random splits to the data to generate multiple split-half reliability estimates and then taking the average of all split-half estimates as the overall reliability estimate (Enock et al., 2014; Parsons et al., 2019). Such permutation-based method for calculating split-half reliability approximates Cronbach's alpha (Cronbach, 1951), while simultaneously avoiding the pitfalls described above. Importantly, splitting the number of observations in half leads to underestimation. The Spearman-Brown (prophecy) formula can be applied to correct for this (Equation 1).

$$r_s = \frac{2r}{1+r} \tag{1}$$

4.2 Test-Retest Reliability

When repeated measures are available, it is possible to calculate test-retest reliability estimates. The Intraclass Correlation Coefficient (ICC) provides evidence for or against the measure's stability over time. More specifically, ICC indicates how well two measurements consistently rank-order the subjects. Ten different forms of ICC have been developed (Mcgraw & Wong, 1996), resulting from a combination of three specified parameters; model selection, type, and the definition of the relationship.

When selecting a model based on a research design, three models are available to choose from: (a) one-way random-effects model, in which each subject is rated by a different set of raters; (b) two-way random-effects model, in which random raters are chosen from a population of similar raters. This model is chosen when the reliability results are to be generalized across different raters (e.g., different clinicians). And (c) two-way mixed-effects model, in which the chosen rater is the only rater of interest (e.g., a computerized experiment).

Selecting a type is more straightforward. If the mean from multiple raters is used to calculate the measure of interest, the type "mean of *k* raters" should be selected. Alternatively, if the measure is based on a single rater, then "single rater" should be selected.

In selecting the definition of the relationship, ICC is estimated based on either a consistency or absolute agreement between the two measurements (e.g., the relationship). A consistency relationship is not affected by systematic changes (e.g., practice effects, learning between measurements) and only the consistency of the rank-order is rated. Absolute agreement expects the two measurements to be identical in rank-order *and* in value (e.g., session mean), in other words, this relationship is affected by systematic differences. For example: these two measurements {1,2,3}, {4,5,6} would have a perfect consistent relationship (ICC = 1.00), but the measurements are not in absolute agreement (ICC = .09). This decision is important in calculating test-retest reliability; should ICC consider systematic differences? If one expects their data to have systematic differences between time one and time two (e.g., practice effects, differences in state), then a consistency relationship should be selected. Otherwise, absolute agreement should be selected.

We suggest, as do others (Koo & Li, 2016; Parsons et al., 2019), that when estimating test-retest reliability of measurements from computerized cognitive-behavioral tasks (e.g., single rater), the important decision is the specification of relationship definition. If systematic differences are expected between time one and time two, then the preferred form of ICC is a two-way mixed-effects, consistency, with a single rater/measurement ((3,1) in Shrout and Fleiss (1979) convention). If one expects no such differences, then calculate ICC based on two-way random-effects, absolute agreement, with a single rater/measurement ((2,1) in Shrout and Fleiss convention). For a more in-depth discussion see Koo & Li (2016).

There are many methods to estimate reliability and it seems that not all researchers are aware of all the options or the assumptions underlying each method. Reliability is crucial when making inferences from correlational results, and hence, correctly estimating and reporting reliability should be a top priority in any individual differences research. Reliability estimates of internal consistency (i.e., permutation-based split-half) should always be reported. When possible test-retest reliability should be reported as well; ICC is robust and provides options for different research designs.

5 Interim Summary

In recent years a number of psychometric issues in the field of cognitive control have led to a debate regarding the construct's validity. Commonly reported concerns include poor reliability of measures, especially of those derived from difference scores, and unexpected weak between-measure correlations. Based on this, a number of studies have suggested that cognitive control research, as is, should perhaps be abandoned (e.g., Paap & Sawi, 2016; Rey-Mermet et al., 2018). Yet, others advocate further examining these issues before taking such drastic measures (Draheim et al., 2020; Hedge, Powell, & Sumner, 2018). These issues are prevalent when approaching cognitive control from an individual differences perspective using established experimental tasks (e.g., the Stroop, flanker, etc.). Psychometrically, these tasks are created to produce a variance structure optimal for analyzing experimental effects, however, the same structure turns out to be unsuitable for individual differences (e.g., Rodebaugh et al., 2016). Such a divide between the two research traditions (e.g., experimental and correlational), is nothing new (Cronbach, 1957). The question on whether and how this divide can be bridged remains. Based on the seemingly pivotal role of reliability, future studies should investigate whether improving reliability can provide such a metaphorical bridge. Specifically, *can existing*

experimental tasks provide reliable individual differences estimates of cognitive control? And *do the weak between-task correlations of cognitive control stem from reliability issues?*

Recent work suggests at least two possible approaches that could potentially answer these questions (von Bastian et al., 2020). A first approach is to introduce theoretically motivated task manipulations in order to increase between-subject variance, and hence, reliability. The Dual Mechanisms of Cognitive Control (DMC) account provides a theoretical framework that decomposes cognitive control into two qualitatively distinct mechanisms – proactive control and reactive control (Braver et al., 2007; Braver, 2012). Empirical findings provide compelling evidence in support of these two modes of control and suggest that an important source of variability in control function at both the individual- and group-level is the bias or preference to adopt one control mode over the other mode (Barch et al., 2001; Braver et al., 2001). Gonthier and colleagues (2016) have shown that task manipulations can indeed shift subjects toward either mode of control. Such added between-subject variance should produce more reliable measures of cognitive control.

A second approach has been suggested by Rouder and Haaf (2019) and Haines et al. (2020). They propose that traditional aggregate statistics are not suitable for extracting individual differences estimates. There are many examples of how averaging across individuals while ignoring "uncertainty" (i.e., individual-level variation) leads to faulty inferences (e.g., Davis-Stober et al., 2016; Estes, 1956; Heathcote et al., 2000; Liew et al., 2016; Pagan, 1984; Vandekerckhove, 2014). Additionally, Rouder and Haaf provide evidence that aggregation, or averaging subject-by-task, "greatly" attenuates measures of reliability, and hence, correlation. Traditional approaches assume the mean point-estimates (MPE) per subject represents their "true" ability, or in other words, is free of measurement error. Measurement error stems from

trial-level noise and it is suggested that reliable individual differences *can* be extracted through modeling this individual-level variability.

This dissertation is a study with two sets of analyses aimed at investigating the viability of these two approaches in recovering reliable individual differences estimates from cognitive control tasks. With the first set of analyses, evidence is presented to show that theoretically motivated task manipulations alone do *not* improve reliability markedly. However, results and patterns that emerged are informative nonetheless. For instance, we show that poor reliability indeed bottlenecks, and hence, alters theoretical interpretations of between-task correlations. Furthermore, we show evidence that difference scores derived from experimental tasks with traditional statistics are not suitable for individual differences. The second set of analyses focuses on modeling trial-level variability through Hierarchical Bayesian Modeling (HBM) using the same data. At this stage there is promising preliminary evidence that HBM can extract reliable individual differences test-retest estimates (Haines et al., 2020). Although cautiously optimistic of the efficacy this approach, a null outcome would provide important evidence that individual differences research of cognitive control needs serious restructuring.

6 Dual Mechanisms of Cognitive Control Task Battery

The Dual Mechanisms of Cognitive Control (DMC) framework suggests that cognitive control operates in two qualitatively distinct modes; a proactive control mode and reactive control mode (Braver, 2012; Braver et al., 2007). Proactive control refers to a sustained and anticipatory mode of control that is goal-directed, allowing individuals to actively and optimally configure processing resources prior to the onset of task demands. Reactive control, by contrast, involves a transient mode of control that is stimulus-driven, which relies upon temporarily retrieving task goals and mobilizing processing resources only after the onset of a demanding

event (Braver, 2012; Braver et al., 2007). In other words, proactive control is planning ahead for alternate timelines, while reactive control is "I will deal with it when I need to". Prior research has dissociated these two modes in both healthy and impaired populations (Braver et al., 2005; De Pisapia & Braver, 2006); based on behavioral signatures in young adults (Gonthier et al., 2016); and among different age groups (Bugg, 2014; Paxton et al., 2008).

In response to challenges with reliability of cognitive control paradigms, the DMC group created a cognitive control task battery including proactive and reactive variants of four well-established cognitive tasks; Stroop, AX-CPT, Cued Task-Switching, and Sternberg Working Memory. These tasks are theorized to measure selective attention, context processing, multi-tasking, and working memory, respectively. The variants were theoretically optimized to capture individual variability in proactive and reactive control. Specifically, there were three variants of each task representing different experimental conditions: (a) a baseline condition that maximizes within- and between-subject variability, which does not bias the adoption of proactive or reactive control; (b) a proactive condition that shifts individuals toward proactive control; (c) a reactive condition that independently engages the reactive mode of control. As will be detailed for each task in the method section, specifying a priori behavioral performance patterns across the three variants enabled us to examine whether proactive and reactive control variants did indeed produce the predicted shifts in control. Additionally, below we describe the method of the study and the rationale for the experimental manipulations underlying the theoretical-based variants of all four cognitive control tasks.

7 Method

7.1 Subjects

Subjects were recruited via the Amazon Mechanical Turk (MTurk) on-line platform. The TurkPrime interface was used to post study descriptions, manage recruitment and payment, send out reminder emails and handle all other communication with the subjects. After reading a description of the study that indicated its multi-session nature and time commitment, interested subjects accessed a link which allowed them to review and sign the consent form. After signing the consent, the web-links for the first session of the study were made available over MTurk. Subjects were not restricted with regard to age range, and as such a wide range were included in the sample (22-64, M=37.11, SD=9.90; 82 females, 47 males).

7.2 Design and Procedure

The study protocol consisted of 30 separate testing sessions that subjects completed in a sequential manner (15 for the test phase, and another 15 for retest). Subjects were asked to complete the sessions at a rate of 5 per week, i.e., 6 weeks to complete the full protocol. Each session lasted approximately 20-40 minutes in duration, with the exception of the first session, which was 1 hour in duration (and included a Stroop practice to validate operation of vocal response recording plus a battery of demographic and self-report questionnaires). To both incentivize and prorate study completion, completion of the first session of both test and retest phases resulted in a $4 payment, each subsequent session was paid $2, with the exception of session 6 and 11, which were paid $4 for each. Additional bonuses of $20 were paid for completion of the test phase and $30 for full study completion. Together, successful completion of the entire protocol resulted in a payment of $122.

A set of 5 sessions were posted at the beginning of each week through MTurk, and were also sent through emails to the subjects. Two reminder emails were also typically sent during the week to remind subjects of the completion deadline for the set (by the end of the week). If subjects failed to complete the weeks' sessions by the designated deadline, they were not invited back to participate in subsequent sessions. For each completed session, subjects would enter in a completion code and the experimenter would review each session results for completion and approve the payment within a week through TurkPrime. If subjects dropped from the study, they still received prorated payment for all sessions completed.

For each completed session, the experimenter checked for overall accuracy and completion of each task and questionnaire to make sure that subjects were complying with instructions and maintaining sufficient attention to the task. A criterion of 60% accuracy and response rate was used to determine whether the data would be included, and the subject invited to remain in the study. For each task or questionnaire that did not meet the criterion, the experimenter attempted to communicate with the subject first to determine if they had trouble understanding the instructions or had technical difficulties. If so, the subject was given a second chance to complete the task before a designated deadline. Within each of the test and retest phases, sessions were conducted in a fixed order for all subjects.

7.3 Task Paradigms

7.3.1 Stroop

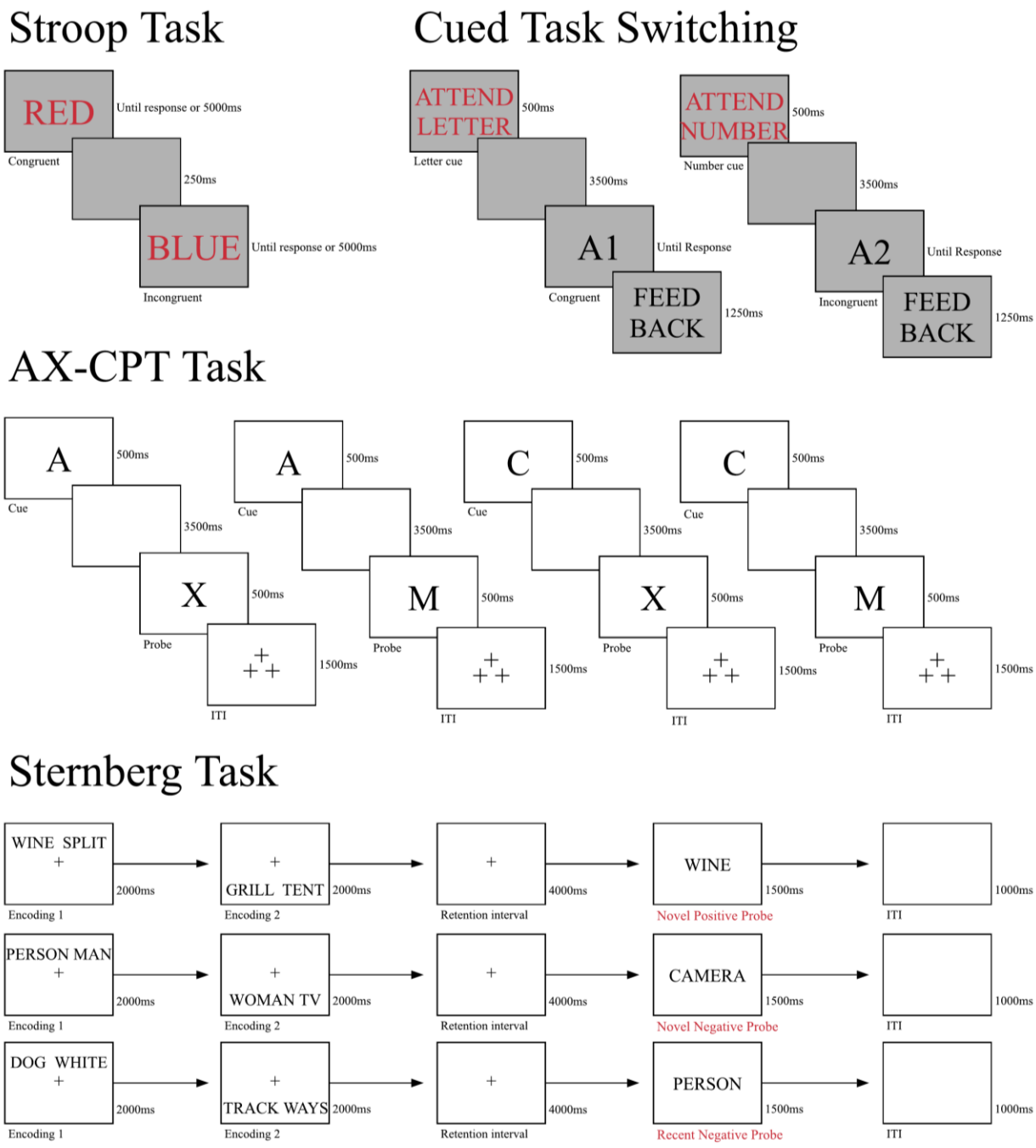
In this vocal Stroop task (see Figure 2), color words are presented in colored font and subjects name the font color out loud. For each trial, vocal response latencies were detected, and accuracy recorded using the computer's built in voice recognition software. Subjects were given standard instructions to respond as quickly as possible (in a normal voice) while retaining

accuracy. Adequacy of the automated voice recognition was validated in previous pilot testing, and individually for each subject based on their first testing session, which contained a practice block of 25 standard Stroop trials. If responses could not be detected for most of the trials, the subject was not asked to continue with further testing.

The current variants of the Stroop were based on the design of previously published work (Gonthier et al., 2016; Gourley et al., 2016) and constructed using two different sets of four colors, in which the relative proportion of congruent and incongruent trials were manipulated in different ways (details below). One set (black, green, pink, yellow) was *unbiased*, in that the proportion of congruent to incongruent stimuli was 50:50 (this set was termed PC-50). The other set of four colors (red, blue, purple, white) was *biased* in the proportion of congruent and incongruent trials, either mostly congruent or mostly incongruent, varied across conditions. The two sets of stimuli were nonoverlapping, such that on incongruent trials, the word name was one of the three remaining colors from that set (e.g., green font with “black”, “pink” or “yellow”; red font with “blue”, “purple” or “white”). All trials consisted of the following stimulus parameters: items were presented centrally on a gray screen for 5000 msec duration or until a response was detected, followed by a 250 msec inter-trial interval during which a blank screen was presented.

Figure 2

Schematic representation of tasks used and their conditions



Note. The trial proceedings pictured represent those of a typical baseline session. Slight variations from the baseline session might have been presented during the proactive or reactive sessions, but these are detailed in the methods section. The AX-CPT task is a simplified depiction for brevity. The font-to-screen ratio shown in this figure is not to scale. ITI = intra-trial interval.

Manipulation Rationale. A commonly used approach to manipulating cognitive control demands in the Stroop task is to vary *list-wide proportion congruence* (PC; Jacoby et al., 2003; Logan & Zbrodoff, 1979). Under high list-wide PC conditions, congruent trials are frequent and incongruent trials are rare within a block, such that control demands are on average low and intermittent. In contrast, under low list-wide PC conditions (rare congruent trials, frequent incongruent), there is a high probability of interference within a block, increasing anticipatory control demands.

In the *proactive condition*, PC is decreased in a list-wide manner; we and others have hypothesized that the tendency to utilize proactive control will increase in low PC conditions (Bugg, 2014; Bugg & Chanani, 2011; Gonthier et al., 2016; Hutchison et al., 2013). In this case, proactive control is theoretically associated with sustained maintenance of the task goal to attend to the ink color and ignore the word, which should be present in a consistent (i.e., present on all trials) and preparatory manner (i.e., engaged even prior to stimulus onset). Thus, the key prediction is that the Stroop effect (average slowing or increase in errors on incongruent relative to congruent trials) should be reduced on all trials, relative to a baseline, high list-wide PC condition, reflecting improved performance on incongruent trials (see Gonthier et al., 2016).

In the *reactive condition*, PC is also manipulated but in an item-specific, rather than list-wide fashion. In this case, specific colors will occur with low PC (e.g., items appearing in green font will frequently be incongruent), while others may occur with high PC (e.g., items appearing in red font will frequently be congruent), and these items are randomly intermixed such that subjects cannot predict whether a low PC or high PC item will appear on a given trial. This type of item-specific PC manipulation is theoretically predicted to enhance the utilization of reactive control for low PC items (Bugg et al., 2011; Bugg & Dey, 2018; Bugg & Hutchison, 2013). For

these items, strong associations develop between a critical feature (a specific ink color) and increased control demands (i.e., high interference), leading to more effective goal retrieval and utilization upon presentation of a stimulus that includes this feature (e.g., a word printed in a green font). The engagement of reactive control is expected to be transient, present only after stimulus onset, and only engaged by low PC incongruent items.

Baseline Session. In the baseline session the trials were manipulated in a list-wide, mostly congruent (LW-MC) manner. Subjects completed a total of 288 trials during the baseline session, in which there were 96 PC-50 trials (48 congruent, 48 incongruent), and 192 biased trials. The biased set had 75% congruent (144 trials) and 25% incongruent (48 trials) trials. Consequently, the list-wide proportion congruency for the baseline session was 66%. The session was divided into two blocks of 144 trials each, between which subjects were instructed to rest for one minute.

Proactive Session. In the proactive session, the trials were manipulated in a list-wide, mostly incongruent (LW-MI) manner. Subjects completed a total of 288 trials during the proactive session, in which there were 96 trials PC-50 (48 congruent, 48 incongruent), and 192 biased trials. The biased set had 25% congruent (48 trials) and 75% incongruent (144 trials) trials. Consequently, the list-wide proportion congruency for the proactive session was 33%. The session was divided into two blocks of 144 trials each, between which subjects were instructed to rest for one minute.

Reactive Session. In the reactive session the proportion congruency manipulation was at the item-level - item-specific proportion congruency (IS-PC). Specifically, blue and red color-font items were manipulated to be PC-100 (i.e., these font-color words were only presented on congruent trials; 192 trials). Purple and white color-font items were manipulated to be PC-25 (i.e., 25% congruent, 48 trials; 75% incongruent, 144 trials). Finally, as in the baseline and

proactive conditions, the remaining 96 trials were PC-50 (i.e., equal amount of congruent and incongruent trials). Thus, subjects completed a total of 480 trials during the reactive session. The session was divided into three blocks of 160 trials each, between which subjects were instructed to rest for one minute.

Cognitive Control Measures. Average reaction times (RTs) on correct trials and error rates were calculated for both congruent and incongruent trials for each subject in each session. The Stroop interference effect (incongruent – congruent) in both RT and also error rate was calculated separately for biased items and PC-50 items.

7.3.2 AX-CPT

The AX-CPT (see Figure 2) has become increasingly utilized as a task of context processing and cognitive control, given its simplicity, flexibility, and applicability in a wide-range of populations. In these variants of the AX-CPT, subjects make button press responses to visually presented cue-probe pairs. A target key press ("/") is made to the probe on AX trials; a non-target key press (".") is made to the probe on the other non-target (AY, BX, BY) trials, and to the cue on all trials. In addition to the four primary trial types, the task also includes no-go trials which require withholding response to the probe and are indicated by a digit (1-9) rather than letter probe. The task comprised 216 trials total, and included 72 AX trials, 72 BY trials, 18 AY trials, 18 BX trials and 36 no-go trials (18 following an A-cue, 18 following a B-cue). All trial types and no-go trials were presented in random order. The task was performed in three 72 trial blocks, between which subjects were instructed to take a minimum of 1-minute rest break. All trials consisted of the following parameters. The cue was presented centrally on a white screen for 500 msec duration. After a 4000 msec blank cue-probe interval, the target (in same size font) was presented for 500 msec but immediately preceded by a 250 msec period during which a bounding box was presented.

A 1500 msec inter-trial interval ended the trial (indicated by a central triangle of fixation crosses). In all of the current AX-CPT variants tested in this battery, the task structure, trial types and frequencies are identical, except for the specific manipulations for proactive and reactive conditions described in the next section.

Manipulation Rationale. First, all three variants include no-go trials, in which the probe is a digit rather than letter. Because of the increase in response uncertainty (i.e., three types of probe response are possible: target, non-target, no-go), the addition of no-go trials decreases the overall predictive utility of cue information for responding, and as a consequence was found to reduce overall proactive control bias typically observed in healthy young adults. As such the no-go trials result in a "low control" baseline, which can be contrasted and used to observe variant-related changes in control mode (Gonthier et al., 2016).

The *proactive condition* replicates prior work using context strategy training to increase predictive preparation of responses following contextual cue information (Gonthier et al., 2016). Specifically, subjects are provided with explicit information regarding the ratios of these cue-probe associations, and receive training and practice in utilizing them to prepare the dominant responses. In addition, during inter-trial intervals, subjects are provided with visual instructions to "remember to use the strategy". The key prediction is that the increased utilization of contextual cue information will lead to a bias to prepare a target response following an A-cue (analyzed in terms of both AX and AY trials) and a non-target response following a B-cue, leading to reduced interference on BX trials, but a side effect of which will be increased errors and response interference on AY trials, which occur when the A-cue is *not* followed by an X-probe.

The *reactive condition* involved a new manipulation which has not previously been examined. Specifically, the reactive condition utilizes item-specific probe cueing; for high control demand trials (i.e., AY, BX, no-go) the probe item appears in a distinct spatial location, and with a distinct border color surrounding it (presented briefly before the onset of the probe). Critically, because these featural associations only form at the time of probe onset, they were not hypothesized to modulate the utilization of proactive control strategies. Likewise, the probe features could not drive direct stimulus-response learning, since they did not directly indicate the appropriate response to be made (i.e., either a non-target or no-go response could be required). In contrast, the probe features can serve as cues signaling high control demand, and thus prompt more rapid and effective retrieval of contextual information to resolve the conflict. Because information about high-conflict probe features is not provided explicitly to subjects (in contrast to the proactive condition), it has to be learned implicitly through experience. The key prediction is that utilization of probe features should reduce the tendency to make BX errors, but could increase BX reaction time interference (due to the tendency to utilize the probe to drive context retrieval).

Baseline Session. The baseline session identically followed the description above. After receiving task instructions, subjects performed a 12-trial practice block before beginning the actual task.

Proactive Session. In the proactive condition, subjects received strategy training before completing the AX-CPT. The strategy training occurred during a practice block of 6 trials, during which an audio clip was played, which instructed subjects which button to prepare following the cue. After this first series of practice trials, subjects performed a second practice set (6 trials), during which they were asked to type which button they were preparing to press in

response to the second item. Subjects typed out “left” or “right” and the program told subjects if they were correct or not. If they were not correct, they were reminded what letter the first item was and asked to try again. This procedure was implemented to accommodate the on-line testing format, and deviated slightly from in-person versions, in which subjects responded verbally regarding the button they were preparing to press. Additionally, during the test phase, in the inter-trial interval periods, subjects were given the visual message to “Use the strategy!”. Otherwise, task structure was identical to the baseline session.

Reactive Session. The occurrence of high conflict trials (AY, BX, no-go) was implicitly signaled by presenting the probe in a distinct spatial location and preceded by a distinct border color. Specifically, while cues were always presented centrally (as in the baseline and proactive conditions) the probe stimuli were either presented in the upper half (AX, BY) or lower half (AY, BX, no-go) of the visual display. Furthermore, probe stimuli were immediately preceded (250 msec before probe onset) by either a white border (AX, BY) or red border (AY, BX, no-go). Otherwise, the task structure and trial proportions were identical to baseline and proactive sessions.

Cognitive Control Measures. Average reaction times (RTs) on correct trials and error rates were calculated for each of the 4 primary trial types (AX, AY, BX, BY) for each subject in each session. Average error rates for no-go trials were calculated as well. Additional derived indices were also computed: A-cue bias, *d'*-context, the Proactive Behavioral Index (PBI), and BX probe Interference (Gonthier et al., 2016). The first two indices, A-cue bias, and d’-context are based on signal detection theory, (Stanislaw & Todorov, 1999) and reflect the use of proactive control. The A-cue bias measure was calculated by computing a *c* criterion from hits on AX trials and false alarms on AY trials as 1/2*(Z[H] + Z[F]), with H representing hits on AX trials and F

representing false alarms on AY trials (Richmond et al., 2015). The *d'*-context index was calculated by computing a *d'* index from hits on AX trials and false alarms on BX trials as Z(H) – Z(F), with H representing hits on AX trials, F representing false alarms on BX trials, and Z representing the z-transform of a value. The third index was the PBI, calculated as (AY – BX)/(AY + BX) (Braver et al., 2009). This index reflects the relative balance of interference between AY and BX trials; a positive PBI reflects higher interference on AY trials, indicating proactive control, whereas a negative PBI reflects higher interference on BX trials, indicating reactive control. The PBI was computed separately for error rates (based on average error rates on AY and BX trials) and for RTs (based on average RTs on AY and BX trials). The fourth index was BX probe interference, calculated as (BX – BY) on both error rates and RTs, including a standardized RT computation. This index allows for examination of the interference that occurs when an "X" probe follows a non-target cue "A" and a target trial response must be inhibited. In order to correct for error rates that were equal to 0, a log-linear correction was applied to all error rate data prior to computing the *d'*-context, the A-cue bias, PBI, and BX interference (Braver et al., 2009; Hautus, 1995). This correction was applied as

$$error + 0.5/N.obs. + 1$$

7.3.3 Cued Task Switching

In the current Cued-TS paradigm (see Figure 2), we used the letter-digit task, which involves bivalent target stimuli consisting of a letter and a digit (e.g., E3). On each trial the subject is cued to perform either a letter task – consonant/vowel discrimination – or a digit task – odd/even discrimination. For the letter task, consonants required right key press ("L") and vowels required a left key press ("A"). For the digit task, even numbers required a right ("L") key press and odd numbers required a left ("A") key press. At the start of every trial the task is

cued by an on-screen message that indicates either “ATTEND LETTER” or “ATTEND NUMBER”, indicating whether attention and responding should be based on the letter or digit, respectively. Critically, because of the response mappings, certain stimuli are congruent, in that they require the same key press irrespective of the relevant task rule (e.g., H6, E3), while other stimuli are incongruent, in that the two tasks were associated with different required responses to the same target (e.g., I6, D4).

The target stimuli were constructed in terms of two distinct stimulus sets. One set of stimuli (A1, A2, B1, B2, 1A, 2A, 1B, 2B) were kept mostly congruent (80% congruent; 20% incongruent). The second set of stimuli (D4, E3, H5, I6, 4D, 3E, 5H, I6) were unbiased (50% congruent, 50% congruent). Trials randomly alternated between an equal number of “ATTEND LETTER” and “ATTEND NUMBER” trials. Due to the random presentation order of the cues, switch and repeat trials were on average equivalent, but deviated slightly in number across conditions and subjects. Each session consisted of 192 total trials, 96 mostly congruent (80 congruent, 16 incongruent) and 96 unbiased (48 congruent, 48 incongruent) and also equally split between the two tasks (i.e., 96 letter, 96 digit). Trials were separated into three 64 trial blocks, between which subjects were required to take a minimum of 1-minute rest break. Prior to starting each session subjects learned (or refreshed their memory) of these response mappings through a set of 16 practice trials. All trials consisted of the following stimulus parameters: trial initiation with a 300 msec alerting cue (flashing cross), followed by the task cue presented on a gray screen for 500 msec duration. After a 3500 msec blank cue-target interval, the target was presented until a response was made. The response was followed by a 1250 msec feedback period, then a 1000 msec inter-trial interval (indicated by a central triangle of fixation crosses).

Manipulation Rationale. An important metric of cognitive control in task-switching paradigms is the task rule congruency effect (TRCE), which refers to the increased interference (both errors and reaction time) when the target response required for the current task is incongruent with the response that would be required to the same target stimulus if the alternative task had been cued. In the *baseline condition,* target stimuli are list-wide mostly congruent (67%), as prior work has found that mostly congruent conditions result in a large and robust TRCE (Bugg & Braver, 2016). The *proactive condition* follows Bugg and Braver (2016) in keeping the same list-wide mostly congruent structure as the baseline condition, but adding reward incentives on a subset of trials. Specifically, on 33% of trials, reward cues are presented simultaneously with advance task cues (i.e., by presenting the task cue in green font), and indicate the opportunity to earn monetary bonuses if performance is accurate and fast (relative to baseline performance) on that trial. By only presenting reward cues on a subset of trials, the remaining subset of non-incentivized trials and target stimuli can be directly compared across the proactive and baseline conditions. The key prediction is that enhanced *proactive control* will lead to a global improvement of performance (i.e., faster RTs while maintaining accuracy).

The *reactive condition* utilizes a new manipulation which has not previously been examined in prior work. Specifically, the reactive condition includes punishment (rather than reward) incentives, again on the same 33% subset of trials that were incentivized in the proactive condition. However, in the reactive condition the incentive cue is presented at the time of the target stimulus, rather than with the task cue, which prevents the use of incentive motivation in a preparatory fashion. Subjects are instructed that they will lose a component of their potential monetary bonus if they make an error on these incentivized trials. Critically, the incentivized trials occur preferentially (75%) with incongruent target stimuli. This manipulation is intended to

associate punishment-related motivation with these high-conflict items, potentially leading to increased response monitoring and caution when incongruence is detected. As such, the key prediction is that enhanced reactive control should reduce the TRCE, even on the non-incentivized trials, when compared to baseline and proactive conditions.

Baseline Session. In this condition, no manipulations were made to the unbiased stimuli. However, to maintain consistency with the proactive and reactive sessions described below, for these stimuli task cues and target stimuli could appear in either red or green font. However, this distinction was irrelevant with regard to the instructions given subjects.

Proactive Session. The proactive variant of Cued-TS was identical to the baseline variant except for the addition of a reward-based motivational incentive. This motivational incentive provides subjects with a reward cue indicated during presentation of the task cue. When subjects responded to incentive trials faster than the baseline session's median RT while maintaining accuracy (this information was stored in a look-up table database, and accessed at the beginning of each session), they received a monetary bonus for that trial added to their compensation amount. Before the start of the proactive sessions, subjects were given the following instructions: "from here to end, you can obtain more payment on top of regular compensation by responding faster than before and maintaining accuracy. A green cue will let you know if you are performing a trial where you can obtain a larger reward." Non-incentive trials indicated by the task cue appearing in red font, while incentive trials were indicated by the task cue appearing in green font. Only the unbiased set of stimuli were incentivized (66% of unbiased, 33% of total, 64 trials) and presentation order was random with respect to the task cue and target stimuli pre-determined pairs. Subjects received feedback on all trials. The word "Reward!" appeared on the screen for 1250 ms if the subject earned the reward. If subjects were too slow or made an

incorrect response, the words “Too Slow!” or “Incorrect!”, respectively, appeared on the screen. The non-incentive trials also included feedback, showing “Correct” or “Incorrect” after each trial.

Reactive Session. The reactive variant of Cued-TS was identical to the baseline variant except for the addition of a punishment-based motivational incentive. This motivational incentive provides subjects with a punishment cue indicated during presentation of the target. When subjects made errors on incentive trials, they received a monetary penalty for that trial that was subtracted from their compensation amount. Before the start of the reactive sessions, subjects were given the following instructions: “from here to end, you can obtain lose money from your regular compensation by making errors. A green cue will let you know if you are performing a trial where you might receive a penalty.” Non-incentive trials indicated by the target stimulus appearing in red font, while incentive trials were indicated by the target stimulus appearing in green font. Only the unbiased set of stimuli were incentivized, and these were applied in an item-specific manner such that all of the incongruent stimuli (H5, 6I, 5H, 6I; 48 trials) were incentivized while only 33% of the congruent stimuli were associated with incentives (D4, E3, 3E, 4D; 16 trials). The sentence “Loss of 25 cents!” appeared on the screen for 1250 ms if the subject made an incorrect response. If subjects were correct or were too slow, or the words “Correct” or “Too Slow!” respectively, appeared on the screen. The non-incentive trials also included feedback, showing “Correct” or “Incorrect” after each trial.

Cognitive Control Measures. Average reaction times (RTs) on correct trials and error rates were calculated separated by congruent/incongruent for the biased items, for each subject in each session. Additionally, these measures were also calculated for the unbiased/incentivized items.

TRCE (Task Rule Congruency Effect) is calculated as a difference score between incongruent and congruent trials and was computed for biased and incentive items separately.

7.3.4 Sternberg

In the current Sternberg item-recognition task (SIRT; see Figure 2), subjects are presented with a list of words on each trial that serves as a memory set (e.g., “WINE”, “SPLIT”, “GRILL”, “INTENT”). After an encoding period and a retention interval delay, a probe item is presented, which requires a judgment as to whether it was part of the current trial’s memory set (i.e., a positive probe) or not (i.e., a negative probe). Specifically, the probe could be: (a) a novel positive word (NP), (b) a novel negative word (NN), or (c) a recent negative word (RN). Where the *novel* condition indicates an “until-then” unpresented word and the *recent* condition an word that was presented in a previous set. The current variants of the SIRT were constructed using two distinct sets of memory items: critical items had a constant memory set of 5 words; and a variable-load set which consisted of either low-load items (memory sets of 2-4 words) or high-load items (memory sets of 6-8 words).

Each session consisted of 120 total trials, broken down into 48 critical items, and 72 variable-load items. Trials were separated into three 40 trial blocks, between which subjects were required to take a minimum of 1-minute rest break. Prior to starting each session subjects learned (or refreshed their memory) of the task through a set of 10 practice trials. All trials consisted of the following stimulus parameters: visual presentation of the memory set across two encoding screens each of 2000 msec duration; in the first screen, were presented above a central fixation cross, and in the second screen, below the cross. Following memory set presentation, a retention interval of 4000 msec was presented (during which the fixation cross remained on screen), followed by 1500 msec presentation of the probe item, and then a 1000 msec inter-trial interval.

Manipulation Rationale. The Sternberg item-recognition task has been one of the most popular experimental paradigms used to assess short-term/working memory for over 50 years (Sternberg, 1966) but more recently has been adapted particularly for the study of cognitive control with the "recent probes" version (Jonides & Nee, 2006). In recent probes versions, the key manipulation is that the probe item can also be a part of the memory set of the previous trial, but not the current trial, which is termed a "recent negative" (RN) probe. On these RN trials, the probe is associated with high familiarity, which can increase response interference and errors, unless cognitive control is utilized to successfully determine that the familiarity is a misleading cue regarding probe status (target or non-target). The current variants of the Sternberg WM included in the battery are adapted from Burgess & Braver (2010) in using manipulations of WM load expectancy and RN frequency.

Specifically, in the *baseline condition*, most trials have high WM load (6-8 items; 60%) and RN frequency is low (20% of non-target probes), which should reduce tendencies to engage either proactive or reactive control strategies. However, in the *proactive condition*, most trials have low WM load (2-4 items; 60%), leading to the expectancy that active maintenance-focused and proactive attentional strategies will be effective, while RN frequency remains low (matched at 20% non-target probes), such that the utility of reactive control should be unchanged. The 5-item set size occurs equivalently in all conditions (40% of trials), and thus can be used to compare performance across different control mode conditions. The key prediction is that use of proactive control strategies, will improve both RT and accuracy, primarily for the target probe items (termed novel positive, or NP, since they never overlap across trials).

In the *reactive condition*, WM loads are identical to the baseline condition (i.e., high-load), while the frequency of RN trials is increased (80% of non-target probes). Thus, in the reactive

condition, it is familiarity-based interference expectancy that increases, rather than WM load expectancy. Based on the increased interference-expectancy, the theoretical hypothesis is that subjects will not rely on familiarity as a cue for responding, and will rather evaluate the match of the probe to items stored in WM. Consequently, the key prediction is that performance on RN (or rather the RN effect, the difference in performance between RN and NN trials) will be significantly improved relative to baseline.

Baseline Session. The baseline session involved high-load variable-items and a low proportion of RN trials (20% of negative probes, 10% of total trials). Specifically, the variable-load set consisted of a mixture of high-load memory sets (12 6-item, 24 7-item, 36 8-item) and very few RN trials (4 RN, 32 NN, 36 NP). For the critical 5-item set, the proportion was slightly adjusted, to increase the number of RN trials for analysis (8 RN, 16 NN, 24 NP).

Proactive Session. In the proactive session, the variable-load items were instead a mixture of low-load memory sets (36 2-item, 24 3-item, 12 4-item). The proportion of RN, NN, and NP trials was identical to the baseline session for both variable-load (4 RN, 32 NN, 36 NP) and critical item sets (8 RN, 16 NN, 24 NP).

Reactive Session. In the reactive session, the variable-load set used the identical mixture of high-load memory set items as the baseline session (12 6-item, 24 7-item, 36 8-item). However, the relative proportion of RN to NN trials was increased in both the variable-load (32 RN, 4 NN, 36 NP) and critical items (16 RN, 8 NN, 24 NP).

Cognitive Control Measures. Separate analyses were conducted for critical items (N = 5) and other variable-load items collapsing across load level. Average reaction times (RTs) on correct trials and error rates were calculated per trial type (i.e., NN, NP, RN trials) for critical items and

non-critical items. One additional index, the recency effect, was also calculated for both RTs and error rates as a difference score on negative trials as RN trials – NN trials.

7.4 Data Pre-Processing

To facilitate comparison of results across task paradigms, subjects who failed to complete all six sessions were not included in the analyses reported here; data from 128 subjects entered the pre-processing stage. The remaining data were conservatively pre-processed in two steps: (1) removal of extreme outliers, and (2) winsorization of remaining outliers. In step 1, all 128 subjects were screened for abnormalities such as extremely slow RTs or high error rates. RT plots were examined and cutoff decisions were made for each task separately. Trials with RTs slower than the cutoff threshold were discarded. The threshold for Stroop was 4000 ms; no RTs on correct trials surpassed the threshold. The threshold for AX-CPT was 2000 ms; no RTs on correct trials surpassed the threshold. The threshold for Cued Task-Switching was 5000 ms and resulted in 0.3% of the task's data discarded. The threshold for Sternberg was 3000 ms; no RTs on correct trials surpassed the threshold. After discarding trials with these RT outliers, the number of trials per condition remained sufficient for analyses. Finally, all subjects in all sessions had a subject-level error rate below 40%; this cutoff is based on Gonthier et al. (2016). No subjects were discarded based on this first step.

In step 2, a winsorization procedure was conducted on RT data at the trial level (i.e., data split by phase, session, trial type, and subject). The winsorization parameters for RTs were as follows: RTs lower than 200 ms were replaced by RTs of 200 ms and RTs above the mean plus 3 standard deviations were replaced by RTs of the mean plus 3 standard deviations. Across the four tasks 1.9% of RT observations were adjusted by the procedure. The adjustments did not vary considerably across tasks, sessions, or trial types. For error rate, the winzoration procedure

was conducted at the level of trial type (data split by phase, session, subject, and trial type), instead of at the subject level, which was examined in the first step of pre-processing. Following the cutoff used by Gonthier et al. (2016), error rates above 40% were replaced with error rates of 40%. This resulted in nearly 5% of error rates being adjusted for the AX-CPT and Sternberg tasks (i.e., 4.78%, 4.69%, respectively). The Stroop and Cued Task-Switching adjustments were much lower at .07% and 1.69%, respectively. Examining this more carefully revealed repeated subpar performance for some subjects (e.g., consistently greater than 80% error rate, large proportion of observations without responses) which inflated the winsorization adjustment rates. Those subjects were excluded from the final sample. We retained 126 subjects for Stroop, 121 for AX-CPT, 128 for Cued Task-Switching, and 126 for Sternberg. Subjects for the between-task correlations were selected pairwise and depending on the task pairing resulted in either a sample size of 120 or 122.

8 Results

8.1 Traditional Analyses

Broadly, the goal of the first set of analyses is to examine whether task manipulations based on a theory or framework that explains individual-level variability improves individual differences reliability. The DMC task battery was created to this end and here we report the individual differences reliability of the measures taken from its four tasks with its variants (e.g., baseline, proactive, reactive). Importantly, the following results only include the critical conditions of the tasks (i.e., Stroop biased condition, task-switching biased condition, Sternberg list-length 5 condition). The critical conditions were designed specifically to allow for comparison across tasks. Descriptive statistics and experimental results by session, task, and trial type are reported in Tang et al. (2021).

8.1.1 Reliability Estimates

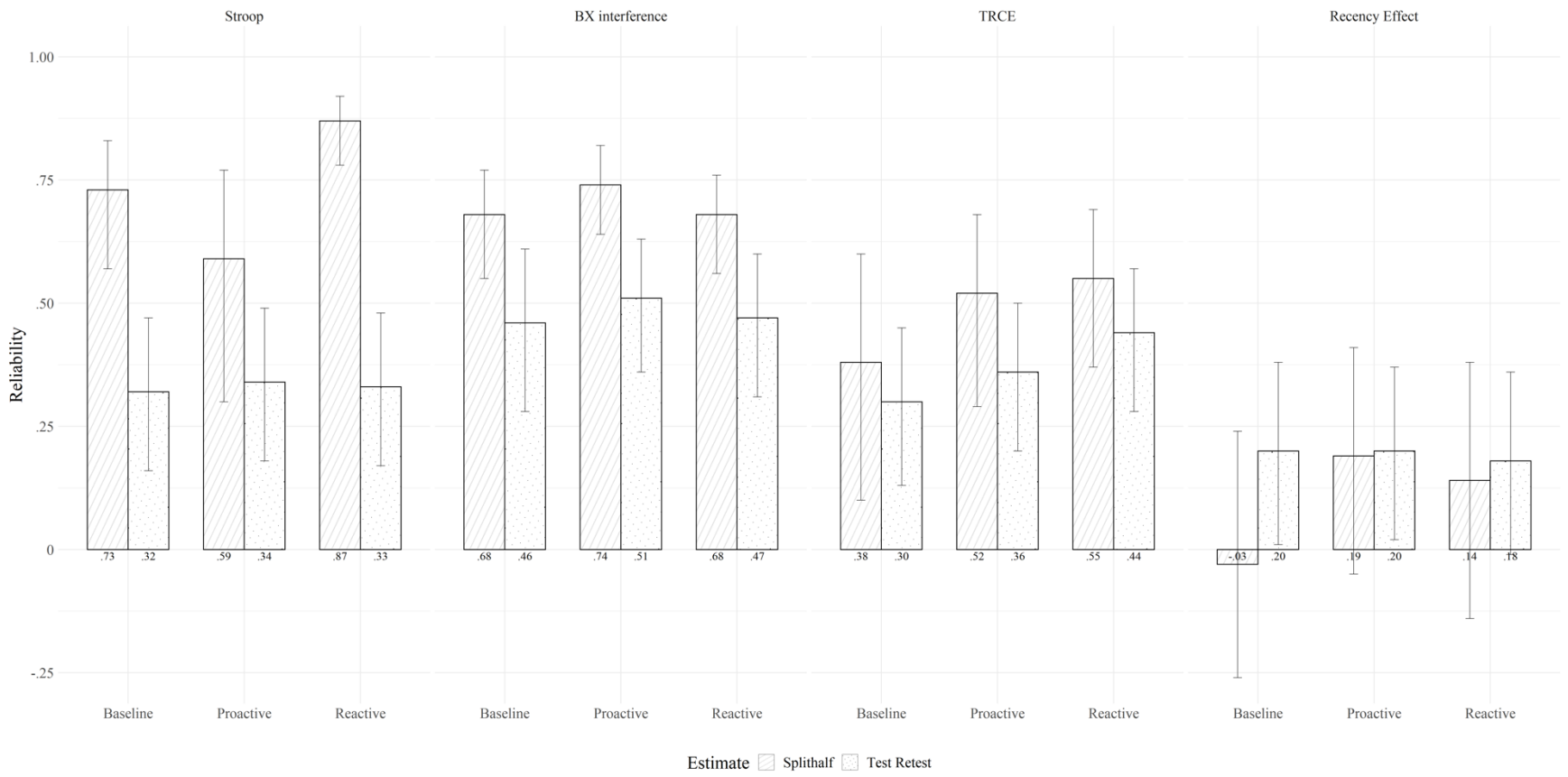
Internal consistency estimates were calculated as permutation-based split-half correlations. The data were repeatedly (5000 permutations) and randomly split into halves, which were then correlated and a Spearman-Brown correction was applied. The estimates reported here are an average of those 5000 corrected correlations. Test-retest reliabilities were calculated as intraclass correlation coefficients (ICC) using a two-way random-effects model of the single-rater type and absolute agreement (i.e., ICC2,1, Shrout and Fleiss (1979)). Because practice effects to are expected to occur from session to session and from test to retest, the ICC relationship parameter was set to absolute agreement. This form is sensitive to changes in the mean between repeated measures. For ease of interpretation, estimates of reliability below .50 are considered poor; between .50 and .75 are considered moderate; between .75 and .90 are considered good; and above .90 are considered excellent (Koo & Li, 2016). However, these

thresholds are somewhat arbitrary; they are offered here as a guide. Of course, the qualitative description of reliability is not a substitute for understanding the numerical estimate in its context.

Difference Score Estimates. Due to the large number of measures, all reliability estimates are presented in Appendix A (Tables A1-A6). There, a full report includes internal consistency and test-retest reliabilities for the aggregate measures (mean RT, error rate) for all trial types, across all tasks and sessions. The main focus here is whether the DMC battery, by introducing theoretically motivated task manipulations, yields reliable measures of cognitive control difference scores. Although the aggregate measures are briefly discussed, only the difference score results are presented here. Figure 3 shows both the split-half and test-retest reliability estimates across sessions (baseline, proactive, reactive) for each task paradigm (2x3x4 = 24 estimates) for RT. The corresponding 24 estimates based on error rate are shown in Figure 4.

Figure 3

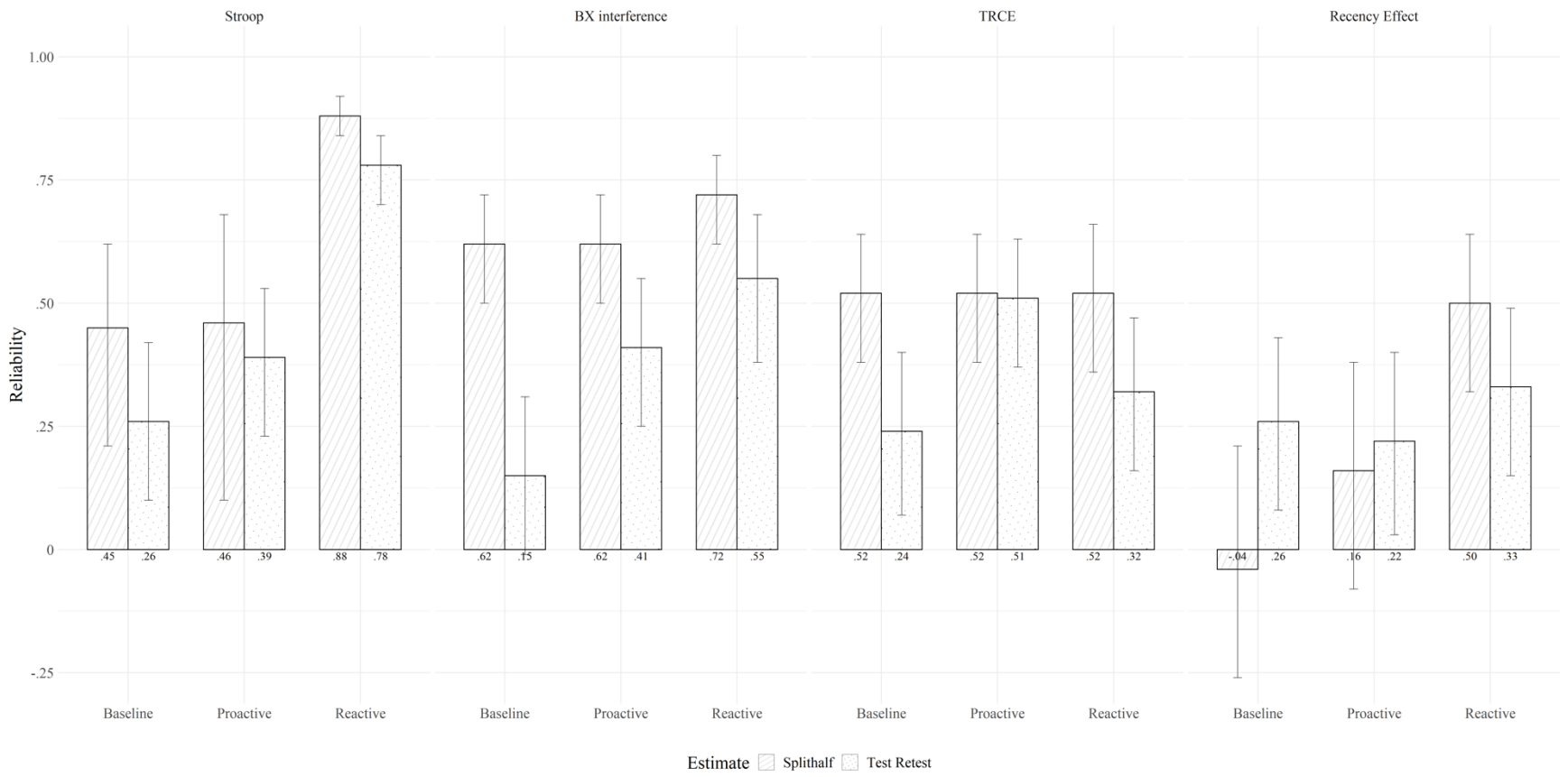
Reaction Time Split-half and Test-Retest Reliability Estimates of DMC Task Battery Difference Scores



Note. Split-half estimates are permutation-based split-half correlations, test-retest estimates are intraclass correlation coefficients (ICC(2,1)). Error bars are 95% confidence intervals.

Figure 4

Error Rate Split-half and Test-Retest Reliability Estimates of DMC Task Battery Difference Scores



Note. Split-half estimates are permutation-based split-half correlations, test-retest estimates are intraclass correlation coefficients (ICC(2,1)). Error bars are 95% confidence intervals.

As expected, the reliabilities of difference score measures are weaker than the reliabilities of aggregate measures. For example, the split-half reliability for Stroop incongruent RT is .99, Stroop congruent RT is .99, but the reliability of the Stroop RT effect is .55. The same general pattern is observed for the test-retest reliability estimates; test-retest for Stroop incongruent RT is .90, Stroop congruent RT is .93, but the reliability of the Stroop RT effect is .32. This pattern is observed across all tasks, for both split-half and test-retest reliability estimates. Because most indices of cognitive control are based on difference scores, this is of serious concern.

Furthermore, the Sternberg recency effect measure is unreliable across the board, for both RT and error rate. The poor reliability and high variability of the Sternberg estimates may stem from research design (i.e., low number of observations available to calculate a difference score). To induce proactive control, recent negative (RN) trials were presented infrequently in the baseline and proactive sessions, with only 8 RN trials per subject. Calculating a difference score from the current Sternberg paradigm for the use of individual differences research is therefore not advised.

The results for Stroop, AX-CPT, and task-switching are mixed. The split-half estimates indicate moderate to good reliability, for both RT and error rate (.52 – .88; with the exception of 3 weaker values). However, the test-retest estimates indicate poor reliability (.15 – .55; with a single .78 outlier). Unfortunately, the session level manipulations (proactive/reactive) did not produce demonstrative improvements in reliability. Although reliability was generally highest in the reactive session, the overlapping confidence intervals across sessions suggests that this is not a robust effect.

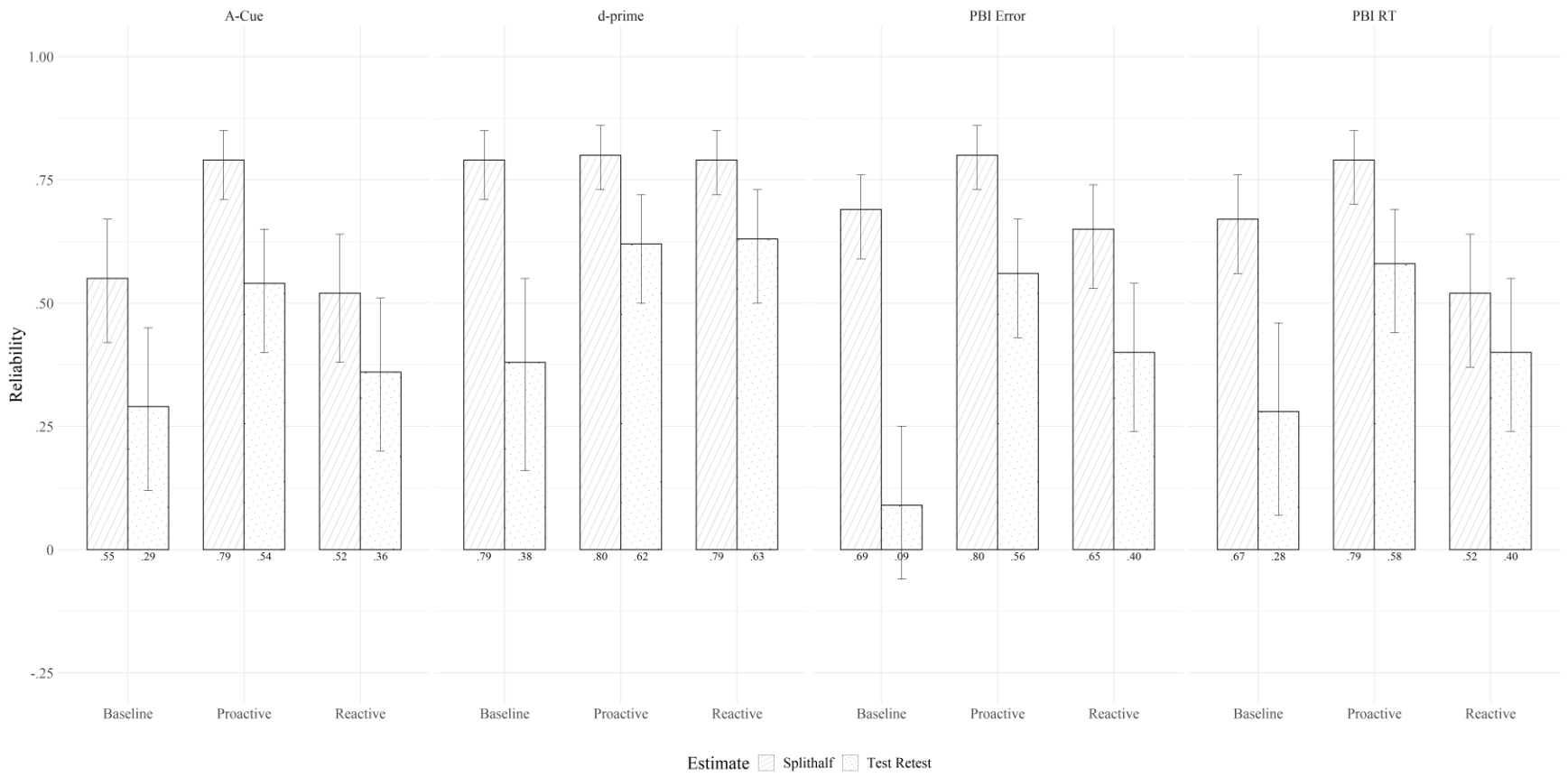
Overall, the reliability results are somewhat disappointing. The DMC task battery was designed to produce reliable and robust effects for both experimental and correlational research.

The battery was largely a success with respect to the experimental manipulations; proactive and reactive shifts in cognitive control were observed and replicated in each task paradigm (Tang et al., 2021; see also Gonthier et al., 2016). However, the current results suggest that the DMC task battery is not as successful when it comes to reliable measurement of individual differences in cognitive control. That said, the difference between split-half and test-retest estimates of reliability is intriguing and may provide some insight into the measurement of cognitive control; we discuss this finding in more detail in the discussion section.

AX-CPT Derived Indices Estimates. Four additional indices were derived for the AX-CPT: the signal detection indices *d′* and A-cue bias, and RT and error rate Proactive Behavioral Index (PBI). The reliability of these derived indices reveals a similar pattern as the difference score measures; the split-half reliability estimates are stronger than test-retest estimates (see Figure 5). In contrast, two novel and interesting patterns emerged. First, all four proactive session derived indices are internally consistent, with split-half estimates ranging from .78–.81. Second, split-half estimates for *d′* exceeded .75 in all sessions and thus is considered to be internally consistent as well. This suggest that the reliability of the *d′* and the proactive indices will not pose a bottleneck when used in between-task correlations.

Figure 5

Split-half and Test-Retest Reliability Estimates of AX-CPT Derived Indices



Note. Split-half estimates are permutation-based split-half correlations, test-retest estimates are intraclass correlation coefficients (ICC(2,1)). Error bars are 95% confidence intervals.

Between-Task Correlations

An important follow-up is the examination of how reliability affects correlations between the tasks. Theoretical conclusions are made from the magnitude of correlations, not from reliabilities. Of course, one should consider the latter when interpreting the former. For between-task correlations it is important to choose the correct type of correlation coefficient. Pearson's correlation coefficient has essentially become the default. However, for some designs other types of correlation coefficients are preferable.

Pearson's correlation coefficient, or Pearson's *r*, assumes that the relationship between two variables is both monotonic and linear (among other assumptions). The relationship between RT and error rate indices of cognitive-behavioral tasks is often monotonic, but not necessarily linear (Hedge, Powell, Bompas, 2018). Hence, Spearman's rho (ρ) is a good non-parametric substitute for the parametric Pearson's r. In addition, Spearman developed the disattenuated correlation coefficient (ρ_{dis}; Spearman, 1904). This provides an estimation of the maximal attainable correlation by correcting for measurement noise. Hence, by comparing ρ and ρ_{dis} one can examine the influence of reliability on correlations.

In total, we observed 198 between-task correlations. Most of the correlations are weak (ρ < .40; Dancey & Reidy, 2004), but some approached moderate strength (ρ => .40). In Table 1 we highlight 22 between-task Spearman's rho correlations exceeding a magnitude of .20. For these analyses, the test and retest data were combined to address the low number of observations in some conditions, with the added benefit that it maximizes power. Complete correlation tables per session are available in Appendix B (Tables B1–B3). For comparison; the corresponding Pearson's *r* correlation tables are provided in appendix B (Tables B4–B6).

Table 1

Selected Between-Task Spearman Correlations with Magnitude Larger than .20.

Index 1	Index 2	Session	ρ	ρ_{dis}	95% CI (ρ)	*n*
A-cue Bias	Stroop Error	Baseline	-.27	**-.54***	[-.430, -.098]	120
	Stroop Error	Proactive	**-.32***	**-.53****	[-.481, -.160]	120
	Stroop Error	Reactive	-.24	-.35†	[-.403, -.065]	120
	TRCE Error	Baseline	.24	.44†	[.061, .398]	122
	TRCE Error	Reactive	.23	.44†	[.056, .393]	122
BXI Error	Stroop Error	Reactive	**-.32***	**-.40****	[-.468, -.144]	120
	Stroop RT	Proactive	**-.31***	**-.51****	[-.468, -.144]	120
	Stroop RT	Reactive	-.26	**-.33***	[-.421, -.087]	120
	TRCE Error	Reactive	.23	.38†	[.053, .390]	122
BXI RT	TRCE Error	Reactive	.20	.34	[.029, .370]	122
d′	Stroop Error	Baseline	-.26	**-.44***	[-.430, -.098]	120
	Stroop Error	Proactive	**-.33***	**-.54****	[-.480, -.160]	120
	Stroop Error	Reactive	**-.37****	**-.44*****	[-.511, -.199]	120
	Stroop RT	Proactive	-.28†	**-.41***	[-.433, -.100]	120
	Stroop RT	Reactive	-.28	**-.34***	[-.432, -.099]	120
	TRCE Error	Baseline	.23	.36†	[.060, .396]	122
	TRCE Error	Reactive	**.30***	**.47****	[.136, .459]	122
PBI Error	Stroop Error	Reactive	**.34****	**.45****	[.167, .486]	120
	Stroop RT	Baseline	.20	.28	[.020, .364]	120
	Stroop RT	Reactive	.23	.31†	[.058, .398]	120
Stroop Error	TRCE RT	Baseline	-.23	-.55†	[-.393, -.053]	120
Stroop RT	TRCE RT	Proactive	.20	.32	[.019, .364]	120

Note. CI = confidence interval; ρ = Spearman's rank correlation coefficient; ρ_{dis} = Spearman's rank disattenuated correlation coefficient; BXI = BX Interference; PBI = Proactive Behavioral Index; TRCE = Task Rule Congruency Effect; Recency = recency effect. Test and retest phase combined.
*** p < .001; ** p < .01; * p < .05; † p < .10

Out of the 22 moderate correlations, seven were statistically significant at the $\alpha = .05$ level, with magnitudes ranging between .30 – .37. Six of these were between indices of the AX-CPT and the Stroop effect, with the remaining significant correlation being between AX-CPT *d′* and TRCE error rate (ρ = .30, p = .04). All of the seven significant correlations came from the

manipulated sessions (i.e., proactive and reactive sessions) but the low number (7 out of 198 total correlations) prevents us from drawing any meaningful conclusions from this result. Including the non-significant correlations, half of the highlighted correlations are between the Stroop effect and a second index (e.g., A-cue bias, BX interference, d').

Given that only seven out of 198 correlations exceeded .30, and given that these are all well- established tasks, it is understandable that some researchers have concluded that cognitive control is simply not a coherent psychometric construct (Rey-Mermet et al., 2018). In fact, the median correlation (r = .13) is on par with the so-called "crud factor" in differential psychology, which refers to the idea that correlations with magnitudes between 0 and .20 should be interpreted as nothing but noise (Lykken, 1968; Meehl, 1986; but see Orben & Lakens, 2020 for a recent critique).

The current study focuses on the importance of reliability when interpreting these correlations. As noted previously, a correlation between two measures is attenuated by the reliability of those two measures. Correlations are negatively affected by measurement error, which distorts the signal with noise. Equation 2 shows Spearman's correction for attenuation (1904), which uses each measurement's reliability (r_{xx}, r_{yy}) as an index of that noise which allows for an estimation of a maximum attainable or "true" correlation ($r_{x'y'}$) by dis-attenuating the measurement's correlation (r_{xy}). For comparison with ρ, Spearman's dis-attenuated rho correlations (ρ_{dis}) are presented in Table 1 as well. It is important to note that these estimates should not be used to make inferences to the tasks or as evidence of their supposed underlying construct (Muchinsky, 1996; Winne & Belfry, 1982). Rather, we present both estimates to examine the role of reliability on between-task correlations, and more importantly, the theoretical implications of the differing magnitudes.

$$r_{x'y'} = \frac{r_{xy}}{\sqrt{r_{xx}r_{yy}}} \quad (2)$$

The standard rho correlations in Table 1 are all considered of weak magnitude. Alternatively, the majority of disattenuated rho correlations reach moderate magnitude (median = .41). Importantly, all correlations now exceed the crud factor threshold. Furthermore, whereas 7 out of 22 rho estimates were statistically significant, disattenuated rho revealed 12 significant correlations. One would be hard-pressed to interpret the magnitude of the rho correlations as evidence for a general underlying construct. However, the disattenuated rho correlations suggest that cognitive control might be a coherent psychometric construct albeit a difficult to measure one.

8.1.2 Interim Discussion

Because studies continue to use popular tasks and estimates optimized for producing experimental effects, individual differences research in cognitive control progresses only slowly. This perpetuates measurement issues as described here, and by others (see Hedge, Powell, & Sumner, 2018). To address these issues, we designed a new task-battery with variants of classic cognitive control tasks. These variants and task manipulations were based on the DMC framework and were hypothesized to create new sources of between-subject variance to improve individual differences. Seemingly, the task battery did not improve reliability or between-task correlations above and beyond previous studies. Yet, from the results of the current set of analyses we can draw some important conclusions.

First, even though difference scores work well to show experimental (within-subject) effects (Tang et al., 2021), they are unsuitable to be used as individual differences predictors. Our findings, which are aligned with those from other studies, indicate that even with optimal methodology and theoretically based task manipulations, the reliability of cognitive control tasks

is a challenge to the field. As mentioned in the introduction, this is not a new phenomenon (e.g., Cronbach, 1957). Difference scores are consistently less reliable than its components (Appendix A), and nowhere near as robust as the experimental effects.

Second, only 11.1% of the between-task correlations reported here surpassed the crud threshold (r = |.20|). More importantly, we show that correlations disattenuated from their reliability bottleneck *do* reach acceptable levels. Hence, conclusions, or even suggestions, based on correlational results stemming from unreliable indices are themselves unreliable and should be treated as such.

Third, reliability is an important metric that needs to be estimated thoughtfully and reported routinely. There are many ways to estimate reliability, but only a few that are appropriate when taking task design and statistical assumptions in consideration. Our results show that split-half reliabilities are stronger than its respective test-retest reliabilities. This may indicate that cognitive-behavioral tasks are more sensitive to state-variability then is generally assumed.

8.2 Hierarchical Bayesian Analyses

The goal of the second set of analyses is to further examine the test-retest reliability (TRR) of cognitive control tasks from a different statistical perspective. As shown in the first set of analyses, we were not able to extract reliable individual differences from experimental task difference score measures. Some research suggests that an alternative statistical approach is needed to address a large flaw with traditional approaches (e.g., those used in the first set of analyses), namely, that they do not model trial-to-trial variability but use mean point-estimates (MPE) as a representative indicator of performance (Haines et al., 2020; Lee & Webb, 2005; Rouder & Haaf, 2019; Rouder & Lu, 2005). In these studies, hierarchical modeling is proposed as an alternative to traditional methods (see also von Bastian et al., 2020). Hierarchical modeling is a statistical framework for modeling data that have a natural hierarchical structure. For example, cognitive-behavioral data that has trials within subjects and subjects within in groups (Gelman et al., 2013). Furthermore, some research (e.g., Rouder & Haaf, 2019; von Bastian et al., 2020) have presented evidence that aggregating performance across trials attenuates reliability, which can be resolved by implementing hierarchical methods that allow for the modeling of trial-to-trial variability (i.e., individual-level standard deviation), in addition to the traditional averaged group-level performance.

There are other negative implications of the traditional MPE approach in analyzing cognitive-behavioral data. For example, Rouder and Haaf (2019) stress that the *portability* of a measurement instrument is "dramatically" overestimated in cognitive-behavioral tasks when using MPE. Portability indicates that measurement properties (e.g., reliability, effect size) of a testing instrument do not change drastically when measuring across varying samples sizes (e.g., number of subjects, number of trials). In other words, portability assumes that an instrument can

obtain an underlying population value that invariably emerges, regardless of experimental design and sample sizes. For example, one lab runs an experiment with 50 subjects each completing 50 trials in *x* conditions. Another lab runs the "same" experiment with 100 subjects each completing 25 trials in *x* conditions. If measurement properties (e.g., reliability, effect size) belong to the experiment, then we expect the measurement indices to be the same across labs (i.e., a portable experiment). If the measurement properties belong to the sample, then we expect measurement indices to differ.

Rouder and Haaf (2019), and Haines et al., (2020), show through simulations that reliability and effect size have a positive relationship with both number of trials presented and sample size in three common cognitive-behavioral tasks (i.e., Stroop, flanker, implicit association task). Furthermore, it is common that researchers alter an existing experiment to fit their testing needs, such as decreasing number of trials presented per subject. Given that common cognitive-behavioral tasks are importable, the MPE approach falsely assumes portability of those tasks where it should not.

Haines et al., (2020) focus on other important implications of using MPE. In the social sciences it is common practice to specify a behavioral model that tests a verbal or conceptual theory. For example, Stroop (1935) theorized that when two properties of a stimulus (e.g., physical (color ink) and semantic (color word)) are incongruent, there is a penalty in both reaction time and accuracy performance when compared to a stimulus with congruent properties. Equation 3 shows the behavioral model of the reaction time Stroop effect, indexed by *i* indicating the effect is calculated for each participant. The behavioral model is then tested through statistical inference with data from the task paradigm. Traditionally, inference involves two stages: (1) calculating MPE (e.g., average reaction times, average correct responses) as

components of an effect (e.g., a difference score), and (2) statistical models (e.g., multiple regression, t-test) are fitted using these averaged effects. This two-stage approach is generally sufficient for detecting reliable experimental effects. However, the lack of information about individual-level response patterns by aggregating trial-level data into MPE, renders such an approach unsuitable for individual differences for two main reasons.

$$\text{Stroop}_i = \overline{\text{RT}}_{i,incongruent} - \overline{\text{RT}}_{i,congruent} \quad (3)$$

First, the behavioral model in equation 3, which is among the most common, ignores individuals' trial-to-trial variability. This variability reflects important behavioral information; a negative relationship has been found between reaction time variability and working memory capacity, and mind-wandering frequency influences this variability as well (McVay & Kane, 2011). Second, the behavioral model in equation 3 does not consider what type of distribution this mean parameter belongs to. Researchers using this behavioral model implicitly assume a normal distribution, where the MPE is the most probable, and hence treated as a representative estimate of the underlying mental process (as a visual aid, see the normal distribution in Figure 6). However, response distributions based on reaction time are rarely normal, but rather heavily right-skewed (Hockley & Corballis, 1982), or modeled as exponential Gaussian (Moscoso del Prado, 2008). Figure 6 shows five different possible distributions that generate an MPE value of 3. Each distribution could imply a different behavioral mechanism generating the data, yet all produce the same MPE. A practical example of this using the Stroop: Heathcote and colleagues (1991) provide evidence that ignoring distribution shape in analyses can obscure behavioral mechanisms. In their study, modeling Stroop reaction time data using MPE-based difference scores revealed the Stroop interference effect, as expected. However, modeling the data with an

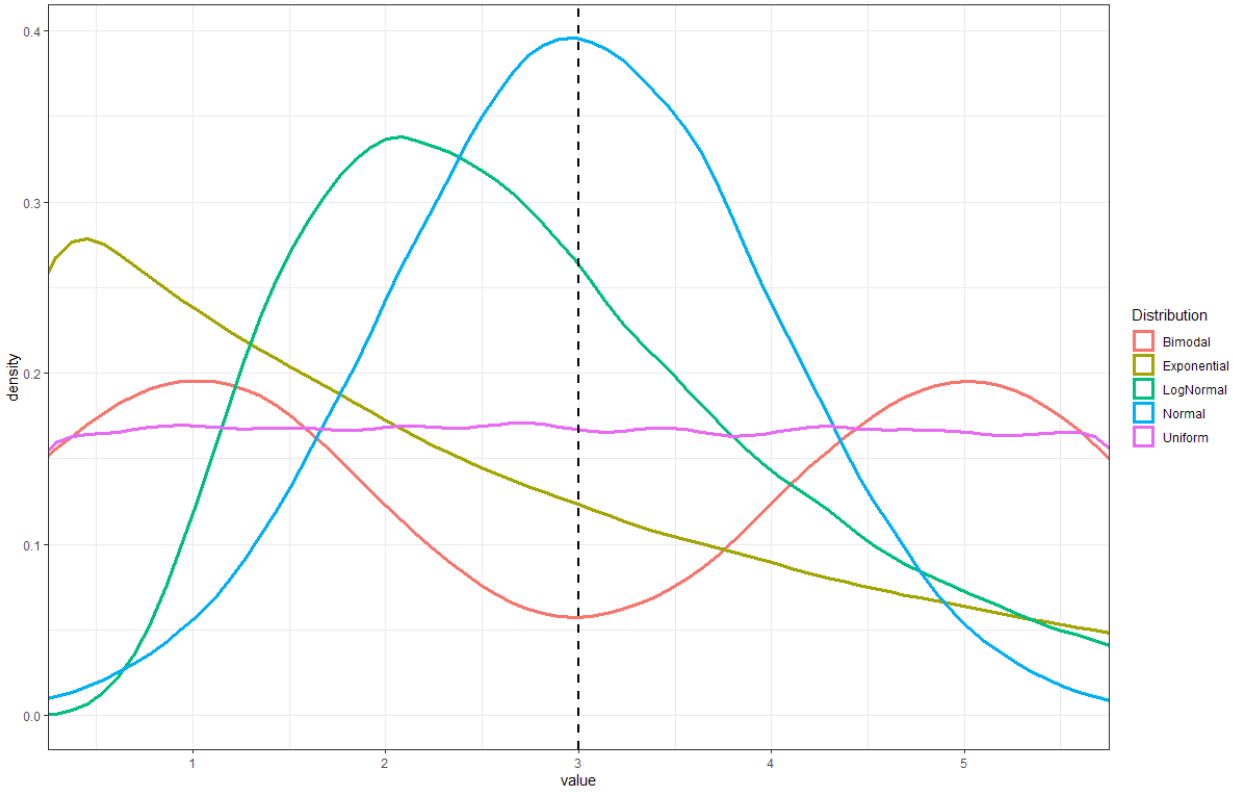
exponential Gaussian distribution revealed a facilitation effect on congruent trials *in addition* to the Stroop interference effect.

To summarize, models solely described by a MPE limit possible inferences about processes underlying behavior that vary between individuals. Theoretically important aspects of behavior can be inferred from parameters like variance (Johnson & Busemeyer, 2005), bimodality distribution shapes (Kvam, 2019), or skewness of the distribution (Kvam & Busemeyer, 2020; Leth-Steensen et al., 2000). It logically follows that models should reflect not only the mechanisms behind intra-individual processes (i.e., modeling mean difference; equation 3), but inter-individual differences in those processes as well (i.e., modeling individual-level variability).

Among others, Haines et al., (2020) suggest that models should ultimately "simulate data consistent with true behavioral observations *at the level of individual participants*", which they refer to as *generative models*. Hierarchical modeling (multilevel modeling, mixed effects modeling) is one framework that allows for such generative modeling. By restructuring a model hierarchically, it considers all subjects in two contexts; as an individual and as a contributing member of a group. This increases the number of available parameters from one (i.e., MPE) to many. The model can now distribute uncertainty that exist in the data (e.g., measurement error) over those multiple parameters, which results in more precise estimates at both the individual and group levels (Kupitz, 2020).

Figure 6

A Non-exhaustive Collection of Distributions with a Mean Value of 3



Note. The black dashed line indicates the mean of each distribution.

Hierarchical *Bayesian* modeling (HBM) has two main advantages over its frequentist alternatives. One, a hierarchical Bayesian model is specified in a single model that *jointly* captures individual- and group-level uncertainty. While subjects perform a limited number of trials and provide data confounded with measurement error, HBM can provide reasonable estimates of performance based on infinite trials (Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). In turn, this solves issues of non-portability in cognitive-behavioral tasks (Rouder & Haaf, 2019). Two, it allows for the specification of distributions and its parameters, which best fits a generative approach. This is not necessarily true for more traditional methods like structural

equation modeling or classical attenuation corrections (Kurdi et al., 2019; Westfall & Yarkoni, 2016).

8.2.1 Hierarchical Bayesian Model

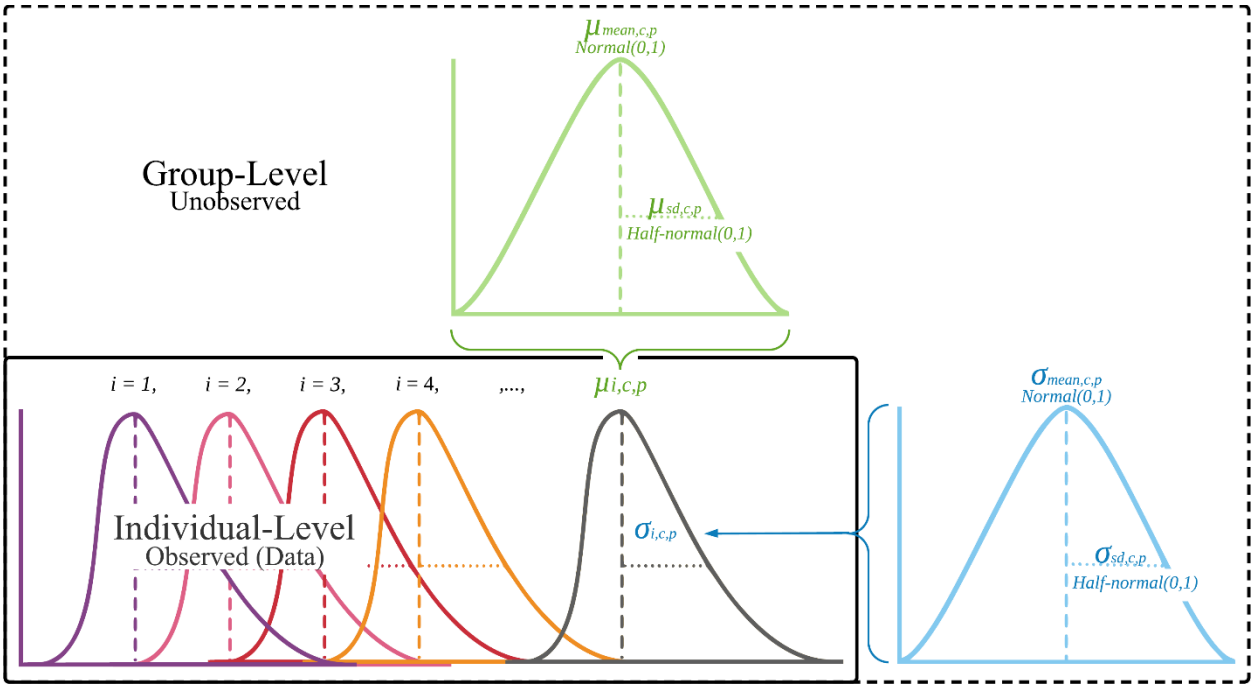
In the second set of analyses, HBM is used to generatively model the four reaction time difference score effects from the Dual Mechanisms of Cognitive Control (DMC) task battery. Specifically, the Stroop effect, the BX interference effect from the AX-CPT, Task Rule Congruency Effect (TRCE) from the cued task switching task, and the recency effect from the Sternberg task. Using these modeled estimates, the TRR and between-task correlations are examined. To facilitate the comparison across these different estimates, we limit the examination in the second set of analyses to a reaction time model only. This approach has the added benefit that a single model can be fit for all sessions within all tasks. A generative model can specified to encapsulate the shared assumptions among the tasks. First, reaction time cannot be negative. Second, reaction time responses vary around some central tendency (this is ignored with MPE). Third, the central tendency varies per subject. Fourth, individual-level variability varies per subject. And fifth, reaction time distributions from cognitive-behavioral tasks tend to be right-skewed (Wagenmakers & Brown, 2007). Although the HBM approach works for accuracy measures as well, it would require a significantly different model, which is outside of the scope of the current project.

As established in the previous paragraphs, it is key that the model considers trial variability at the individual-level, hence, the individual-level distribution is defined first, followed by the group-level distribution. Finally, this section will conclude with the specification of the test-retest reliability, model priors, and estimation procedure. All R scripts and the Stan

model file are available on https://osf.io/79jgs/. A graphical representation of the model is included as well (see Figure 7).

Figure 7

A Structured Schematic Representation of the Hierarchical Model



Note. i = subject; c = condition; p = phase; sd = standard deviation; μ_i = individual-level mean parameter; σ_i = individual-level variability parameter.

Individual-Level Parameter Model. Subject's reaction time response distributions are here conceptualized as coming from a lognormal distribution, satisfying our skewed distribution assumption (assumption 5). The distribution is further shaped by mean and standard deviation parameters, which *both* vary per subject and between each condition (satisfying assumptions 2, 3, and, 4). Theoretically, we do not expect the distribution parameters to vary much between the test and retest phase. However, for test retest reliability purposes, the model assumes unique distributions for each phase as well.

$$\mathrm{RT}_{i,c,p} \sim \mathrm{Lognormal}(\mu_{i,c,p}, \exp(\sigma_{i,c,p})) \quad (4)$$

Formally, in equation 4, $\mathrm{RT}_{i,c,p}$ is the observed reaction time data for subject $i = \{1,\ldots, N\}$, in condition $c = \{$control, interference$\}$[1], during phase $p = \{$test, retest$\}$. $\sim \mathrm{Lognormal}(\mu_{i,c,p}, \exp(\sigma_{i,c,p}))$ signifies that the data are drawn from a generative process producing a skewed distribution, shaped by a mean and standard deviation parameter for each subject, condition, and phase combination. A lognormal distribution has an asymmetrical spread; more variability is found on the right-side (i.e., slow reaction times) of the central tendency than the left-side (i.e., fast reaction time). Importantly, the lognormal distribution has a property that determines how the mean and standard deviation interact, allowing the model to fit the many different shapes of reaction time distributions produced by the ~ 120 subjects. Wagenmakers and Brown (2007) show that this property adheres to a *law of [reaction] time*, which states that in reaction time performance, the standard deviation increases linearly with the mean. In other words, the slower a subject's mean reaction time, the more individual-level variability they show. Additionally, to ensure that the individual-level standard deviation parameters are greater than 0, they were exponentially transformed.

[1] Control corresponds to non-interference trial types (e.g., Stroop congruent, Sternberg novel negative). Interference corresponds to interference trial types (e.g., AX-CPT AY and BX, task-switching incongruent).

Group-Level Parameter Model. In a hierarchical model, individual-level parameters are informed by group-level parameters, and vice versa. Here, the hierarchy of the model is constructed so that the individual-level distribution parameters from Equation 4, denoted by $\mu_{i,c,p}$ and $\sigma_{i,c,p}$, are drawn from group-level normal distributions (i.e., prior models), with unobserved (i.e., unknown) means and standard deviations (sd):

$$\begin{aligned} \mu_{i,c,p} &\sim \text{Normal}\left(\mu_{\text{mean},c,p}, \mu_{\text{sd},c,p}\right) \\ \sigma_{i,c,p} &\sim \text{Normal}\left(\sigma_{\text{mean},c,p}, \sigma_{\text{sd},c,p}\right) \end{aligned} \tag{5}$$

By defining these prior models, the group-level distribution allows for the pooling of information across subjects. Each individual-level parameters, $\mu_{i,c,p}$ and $\sigma_{i,c,p}$, inform the group-level means and standard deviations, $\mu_{\text{mean},c,p}, \mu_{\text{sd},c,p}$ and $\sigma_{\text{mean},c,p}, \sigma_{\text{sd},c,p}$, which in turn inform all other individual-level parameters. This mutual interaction creates *hierarchical pooling*, regressing the individual-level parameters towards a group mean (also called *shrinkage* or *regularization*), and increases the precision of Bayesian estimation (Gelman et al., 2013). Bayesian modeling allows for such a "joint model" specification, in which the individual-level and group-level parameters are estimated simultaneously. This embodies the generative perspective (Haines et al., 2020).

Keen observers might remark that the group-level distributions are both modeled as normal. Recall that the individual-level standard deviation parameter (Equation 4; $\exp(\sigma_{i,c,p})$) was exponentially transformed to force it to assume positive values only. Mathematically, when *y* has a normal distribution then the exponential function of *y* has a lognormal distribution. It follows then, that the group-level distribution modeled on the individual-level standard deviation parameter ($\exp(\sigma_{i,c,p})$) corresponds to a lognormal distribution.

Model Priors. One strength of Bayesian modeling is that we can define a prior probability distribution which expresses one's prior belief about an underlying distribution of interest. In the sections above the reasoning for the priors have been explained. In the current project, the parameter estimation is rather robust to prior models, because the priors are rather diffuse and the sample sizes of observed data are relatively large. More about the influence of the priors on parameter estimation can be found in the "WAMBS" section below.

The prior model for the group-level mean parameters were as specified as normal.

$$\begin{aligned} \mu_{mean,c,p} &\sim \text{Normal}(0, 1) \\ \sigma_{mean,c,p} &\sim \text{Normal}(0, 1) \end{aligned} \tag{6}$$

The prior model for the group-level standard deviations parameters were as specified as half-normal (e.g., if *y* is a normal distribution, then | *y* | is a half-normal distribution, folded along the mean with the purpose of consisting of only positive values). Because the individual-level standard deviation parameter is exponentially transformed, the group-level distribution assumes only positive values.

$$\begin{aligned} \mu_{sd,c,p} &\sim \text{Half} - \text{Normal}(0, 1) \\ \sigma_{sd,c,p} &\sim \text{Half} - \text{Normal}(0, 1) \end{aligned} \tag{7}$$

Parameter Estimation. All model parameters were estimated with Stan (Stan Development Team, 2020c) through an interface in R, called RStan (Stan Development Team, 2020b). Stan is a probabilistic programming language that includes inference algorithms for fitting models and making predictions. Bayesian inference for continuous variable models is achieved by Stan's implementation of a more efficient and robust variant of a Markov chain Monte Carlo (MCMC) algorithm, the Hamiltonian Monte Carlo (HMC) (Carpenter et al., 2017). All models were fitted with 3 chains of 3000 iterations after 1000 warm-up iterations.

As is common, the levels within hierarchical models are strongly correlated by design because the group-level distributions are generated from individual-level parameters. This can lead to inefficient exploring of the distribution parameter space by the HMC sampler. The sampler still explores the entire parameter space, but extremely slowly, resulting in requiring many iterations for proper convergence of the models. A commonly used practice to counter this, is to use non-centered parameterizations (Betancourt & Girolami, 2013). Critically, these parameterizations do not change the model, its interpretation, nor the resulting parameter estimates. For computational efficiency, we followed Haines et al., (2020) in offsetting the individual-level parameters. For an overview of non-centered parameterization see Betancourt and Girolami, (2013) and Papaspiliopoulus et al., (2007); also see chapter 22.7 of the Stan User's Guide (Stan Development Team, 2020a).

Extracted Parameters. For each of the four tasks in the task battery, the model was fit three times (e.g., once for each session), resulting in 12 model fits. From the model fits we extracted three families of parameters; delta, mu, and sigma. After the parameters are estimated and extracted, it is straight forward to generate a difference score estimate, which shall be referred to as delta (i.e., Δ).

$$\begin{aligned} \Delta_{i,test} &= \mu_{i,interference,test} - \mu_{i,control,test} \\ \Delta_{i,retest} &= \mu_{i,interference,retest} - \mu_{i,control,retest} \end{aligned} \tag{8}$$

Furthermore, the individual-level means (i.e., $\mu_{i,c,p}$; referred to as mu) and standard deviations (i.e., $\sigma_{i,c,p}$; referred to as sigma) were extracted for each condition and phase. No group-level means or standard deviations were extracted due to the individual differences nature of this project. The extracted delta, mu, and sigma parameters for each task and session combination are available on https://osf.io/79jgs/.

8.2.2 "WAMBS"

Bayesian statistical approaches are becoming a popular method across different disciplines. The advantages of Bayesian statistics can be attractive, but naively applying Bayesian methods can be dangerous for interpretation of the results. Because of the additional complexity of the method, its programming, and the freedom of distribution and parameter specification, there is a list of considerations that must be made before interpreting the results. Fortunately, these considerations are collected in the "*W*hen to worry and how to *A*void the *M*isuse of *B*ayesian *S*tatistics" checklist (WAMBS; Depaoli & van de Schoot, 2017). WAMBS describes potential issues that can come up before and after estimating the model, which are collected in a checklist format. The current project heeded the relevant warnings of the WAMBS checklist, which are succinctly reported next.

There are two relevant items regarding the priors on the WAMBS checklist: *do you understand the priors* and *is there a notable effect of the prior when compared with non-informative priors*. The choices for the priors are explained in the previous sections. In preparation for the parameter estimation, models with different theoretical plausible priors, and no priors at all, were run. When no priors are defined, Stan defaults all prior distribution models to a uniform distribution (i.e., a non-informative prior). However, because the effect of delta (i.e., the difference score) is small (see the Results section), and the relatively large sample size, the priors had a negligible impact on the results. The Stan model with no prior is included in the WAMBS folder on https://osf.io/79jgs/.

A large section of the WAMBS concerns itself with the convergence of the model and parameter estimation. WAMBS contains two items that can be answered straight-forwardly: *does convergence remain after doubling the number of iterations* and *does the posterior distribution*

make substantive sense. Yes, the models were run with different numbers of iterations and burn-in phases, including twice the number of the currently used number of iterations; no convergence issues were found in these test runs. The posterior distribution makes substantive sense, as can be seen in the Results. Other convergence items are more easily answered with convergence plots; *does the trace-plot exhibit convergence*, *does the histogram have enough information*, and *do the chains exhibit a strong degree of autocorrelation*. All relevant convergence statistics have been extracted and are visually presented in the WAMBS folder on https://osf.io/79jgs/.

8.2.3 Reliability Estimates

Test-retest reliability (TRR) estimates for the delta parameter (i.e., difference score; $\Delta_{i,test}, \Delta_{i,retest}$) were calculated for each task and session combination and shown in Table 2. For a comparison between the traditional and HBM approach, corresponding mean point-estimates TRRs (r_{MPE}) are provided as well in Table 2. Importantly, TRR is calculated as a Pearson r correlation between the test and retest phase estimates $r(\Delta_1, \Delta_2)$. Pearson r is chosen over an Intraclass Correlation Coefficient, because much of the variance has been modeled out by the sigma parameter.

Difference Score Estimates. As expected, the MPE approach indicates a poor to moderate test-retest reliability ($\bar{x}$ = .39; see the first set of analyses as well), which is consistent with previous studies (e.g., Hedge, Powell, and Sumner, (2018): $\bar{x}$ = .55; Chen et al., (2021): $\bar{x}$ = .49). In contrast, the TRR for the HBM delta parameters can be classified as good to excellent, with the reactive session recency effect being an exception (r = .52). These results are consistent with Haines et al., (2020), and Rouder and Haaf (2019). The delta TRR estimates indicate that HBM can indeed provide reliable individual differences from cognitive control tasks, even though the index is a difference score. An additional interesting pattern emerges when comparing

TRR across session; the TRR is highest for the proactive session, as was the case with the intraclass correlation coefficients in the first set of analyses.

Table 2

Reaction Time Test-Retest Correlations of the Delta Parameter from the DMC Task Battery.

Session	Task	Index	$r(\Delta_1, \Delta_2)$	r_{MPE}	n
Baseline	Stroop	Stroop Effect	.92	.54	122
Proactive			.98	.59	119
Reactive			.88	.55	122
Baseline	AX-CPT	BX Interference	.79	.50	112
Proactive			.93	.51	116
Reactive			.86	.49	113
Baseline	Cued TS	TRCE	.81	.22	116
Proactive			.94	.28	112
Reactive			.90	.39	122
Baseline	Sternberg	Recency Effect	.77	.16	120
Proactive			.89	.20	106
Reactive			.52	.20	127

Note. $r(\Delta_1, \Delta_2)$ = Pearson correlation coefficient of delta estimates obtained by Hierarchical Bayesian Modeling; r_{MPE} = Pearson correlation coefficient obtained from traditional Mean Point Estimates approach; TS = task-switching; TRCE = Task Rule Congruency Effect; different *n* sample sizes due to additional multivariate outlier removal.

8.2.4 Between-Task Correlations

It has been suggested that the weak between-task correlations of cognitive control paradigms stem from the poor reliability of the measures (Hedge, Powell, & Sumner, 2018). Here this suggestion is tested. In Table 3 the between-task correlations of the delta parameter are presented. For each estimate pair, in each session, a correlation is calculated for the test phase and retest phase. Additionally, the test and retest phase were also combined to follow the first set of analyses.

Table 3

Reaction Time Between-Task Correlations of the Delta Parameter from the DMC Task Battery.

Session	Index 1	Index 2	r_{test}	r_{retest}	$r_{combined}$	*n*
Baseline	Stroop Effect	BX Interference	.05	.17	.10	90
Baseline		TRCE	.02	.02	.02	90
Baseline		Recency Effect	-.01	-.02	-.02	90
Baseline	BX Interference	TRCE	-.01	.05	.03	90
Baseline		Recency Effect	-.12	-.12	-.13	90
Baseline	TRCE	Recency Effect	.11	-.04	.00	90
Proactive	Stroop Effect	BX Interference	.01	.01	.01	76
Proactive		TRCE	.00	.02	.01	76
Proactive		Recency Effect	-.06	-.07	-.07	76
Proactive	BX Interference	TRCE	-.09	-.16	-.13	76
Proactive		Recency Effect	-.03	-.04	-.04	76
Proactive	TRCE	Recency Effect	-.12	-.13	-.13	76
Reactive	Stroop Effect	BX Interference	.12	.01	.08	107
Reactive		TRCE	-.09	-.10	-.09	107
Reactive		Recency Effect	.06	-.10	-.05	107
Reactive	BX Interference	TRCE	-.04	-.01	-.02	107
Reactive		Recency Effect	.23	.15	.22	107
Reactive	TRCE	Recency Effect	-.10	.00	-.04	107

Note. r_{test} = Pearson r correlation at test phase; r_{retest} = Pearson r correlation at retest phase; $r_{combined}$ = Pearson r correlation of combined phases; TRCE = Task Rule Congruency Effect. Variability in sample sizes due to between-task differences in acceptable performance.

Despite the strong reliability of the delta parameter, the between-task correlations of the cognitive control measure are, yet again, on par with the so-called "crud factor" (i.e., correlations between -.20 and .20; see the first set of analyses). This is evidence against the suggestion that the weak between-task correlations of cognitive control are caused by poor reliability of its measures. The "strongest" between-task correlation, and the *only* that is not considered "crud", contains the reactive session recency effect (r = .23). Interestingly, this estimate had the lowest reliability (r = .55) and was also the *only* moderately reliable estimate among good and

excellently reliable estimates. It is important to note that the delta parameter is calculated as a difference score. As elaborated in the first set of analyses, difference scores are notoriously problematic from a psychometric perspective (Caruso, 2004; Cronbach & Furby, 1970; Lord, 1956). The finding that hierarchical Bayesian modeling cannot improve between-task correlations of difference score estimates is corroborated by Rouder and Haaf (2019).

8.2.5 Sigma

In psychology, measuring a subject's level of ability is often done by an estimate that represents their *average* ability. Such a mean point-estimate (MPE) is simply derived from the average performance on a task (e.g., *mean* reaction time, *mean* accuracy), or the mean parameter (e.g., μ_i) from a modeled distribution. However, some studies have focused on the *variability* of responses rather than a response average as index of cognitive ability (e.g., Der & Deary, 2006; Dykiert et al., 2012; Hultsch et al., 2008; MacDonald et al., 2006; Salthouse, 2007; Stuss et al., 2003). Each of these studies show that intra-individual variability (IIV) can reveal an important aspect of task performance and its underlying mental processes, namely, its consistency. Two subjects can have identical average performances (e.g., 700ms on incongruent Stroop trials), but might differ in how much they deviate from this average on a given trial (Unsworth, 2015).

Some research has suggested that increases in IIV are related to fluctuations in cognitive control which can lead to lapses of attention (Duchek et al., 2009; Jackson et al., 2012; Unsworth et al., 2010; West et al., 2002). Such lapses of attention would manifest in the data as sporadic slow reaction times due to task-unrelated thoughts or very fast errors due to failure to inhibit a prepotent response (Unsworth et al., 2004), in other words; an increase the IIV. This view is supported by studies that show that subjects with low cognitive ability (i.e., working memory capacity, fluid intelligence) demonstrate a larger number of very slow responses when compared

to subjects with high cognitive ability (McVay & Kane, 2011; Schmiedek et al., 2007; Unsworth et al., 2010, 2012). More importantly, this suggests IIV may be an important source of individual differences in cognitive control.

We examine whether IIV is a general trait that manifests itself similarly across different tests of cognitive control. An estimate of individual-level standard deviation is a common dependent variable of IIV. The hierarchical Bayesian model in the current project includes such an estimate: $\sigma_{i,c,p}$ (sigma). The between-task correlations of the sigma parameter for the DMC task battery interference (Table 4) and non-interference (Table 5) trial types are presented. Overall, the between-task correlations of the sigma parameter indicate that it is a better suited estimate of individual differences in cognitive control than the delta parameter. There are some intricate differences when examining the correlations across task, trial type (e.g., interference, non-interference), and session. To facilitate interpretation of these differences, the results are split between the trial type and session.

Between-task correlations of the baseline session, interference trial type, sigma parameters (Table 4) are generally moderate to good (median = .49), with the exception of correlations that include the Stroop incongruent trial type; those are generally weak (median = .17). A similar pattern can be found for the proactive session sigma parameters; a median of .39, and .18, respectively. However, for the reactive session no such difference is found; a median of .25 for both correlations including and excluding the Stroop incongruent trial type. This may indicate that the Stroop task is too simple to pick up on individual differences in IIV; or that it does not share a process that is required by the other tasks. Furthermore, there are some unexpected discrepancies when comparing the same correlations from the test phase to the retest

phase. However, no discernable pattern can be found; in some cases the correlations are near identical, in others they differ 15 points.

Table 4

Reaction Time Between-Task Correlations of the Interference Trial Sigma Parameter from the DMC Task Battery.

Session	Index 1	Index 2	r_{test}	r_{retest}	$r_{combined}$	*n*
Baseline	Stroop Incon.	AX-CPT BX	.17	.22	.20	90
Baseline		TS Incon.	.09	.23	.15	90
Baseline		Sternberg RN	.11	.12	.20	90
Baseline	AX-CPT BX	TS Incon.	.49	.48	.54	90
Baseline		Sternberg RN	.42	.33	.45	90
Baseline	TS Incon.	Sternberg RN	.34	.46	.48	90
Proactive	Stroop Incon.	AX-CPT BX	.12	.22	.18	76
Proactive		TS Incon.	.26	.32	.34	76
Proactive		Sternberg RN	-.01	.15	.10	76
Proactive	AX-CPT BX	TS Incon.	.39	.43	.48	76
Proactive		Sternberg RN	.33	.40	.45	76
Proactive	TS Incon.	Sternberg RN	.36	.18	.25	76
Reactive	Stroop Incon.	AX-CPT BX	.25	.24	.28	107
Reactive		TS Incon.	.29	.28	.35	107
Reactive		Sternberg RN	.01	.06	.05	107
Reactive	AX-CPT BX	TS Incon.	.31	.38	.40	107
Reactive		Sternberg RN	.26	.22	.24	107
Reactive	TS Incon.	Sternberg RN	.14	.25	.24	107

Note. r_{test} = Pearson r correlation at test phase; r_{retest} = Pearson r correlation at retest phase; $r_{combined}$ = Pearson r correlation of combined phases; Stroop Incon. = Stroop incongruent trial type; Sternberg RN = Sternberg recent negative trial type; TS Incon.= Task-Switching incongruent trial type. Variability in sample sizes due to between-task differences in acceptable performance.

The results of the non-interference trial type sigma parameters (Table 5) also reveal a difference in the between-task correlations of those including the Stroop, and those not including the Stroop estimates, but only in the baseline session; median = .28, and .40, respectively. For the proactive session (median = .31, and .31), and the reactive session (median = .29, and .31),

the correlations are more homogeneous across all tasks. Importantly, across all tasks, sessions, and phases, the non-interference sigma estimate has a median between-task correlation of .30. This is a meaningful difference when compared to the median between-task correlation of the delta estimate: -.01.

Table 5

Reaction Time Between-Task Correlations of the Non-Interference Trial Sigma Parameter from the DMC Task Battery.

Session	Index 1	Index 2	r_{test}	r_{retest}	$r_{combined}$	*n*
Baseline	Stroop Con.	AX-CPT BY	.21	.23	.30	90
Baseline		TS Con.	.13	.28	.22	90
Baseline		Sternberg NN	.41	.37	.44	90
Baseline	AX-CPT BY	TS Con.	.32	.39	.44	90
Baseline		Sternberg NN	.36	.26	.40	90
Baseline	TS Con.	Sternberg NN	.45	.44	.47	90
Proactive	Stroop Con.	AX-CPT BY	.15	.11	.14	76
Proactive		TS Con.	.32	.34	.39	76
Proactive		Sternberg NN	.20	.36	.31	76
Proactive	AX-CPT BY	TS Con.	.31	.31	.34	76
Proactive		Sternberg NN	.38	.34	.38	76
Proactive	TS Con.	Sternberg NN	.29	.25	.29	76
Reactive	Stroop Con.	AX-CPT BY	.29	.07	.18	107
Reactive		TS Con.	.28	.30	.34	107
Reactive		Sternberg NN	.21	.30	.34	107
Reactive	AX-CPT BY	TS Con.	.27	.34	.36	107
Reactive		Sternberg NN	.29	.22	.31	107
Reactive	TS Con.	Sternberg NN	.22	.36	.33	107

Note. r_{test} = Pearson r correlation at test phase; r_{retest} = Pearson r correlation at retest phase; $r_{combined}$ = Pearson r correlation of combined phases; Stroop Con.= Stroop congruent trial type; TS Con.= Task-Switching congruent trial type; Sternberg NN = Sternberg novel negative trial type. Variability in sample sizes due to between-task differences in acceptable performance.

9 Discussion

The second set of analyses replicates previous findings (Chen et al., 2021; Haines et al., 2020; Rouder & Haaf, 2019) showing that hierarchical Bayesian methods can produce reliable cognitive control indices, including difference score indices (Table 2). Our main research question was: *can existing experimental tasks provide reliable individual differences estimates of cognitive control?* We found that test-retest reliability (TRR) estimates for the delta parameters were found to be good or excellent. Compare this to the weak and moderate intraclass correlation coefficients (ICC) in the first set of analyses. This suggests, that accounting for individual-level variability and the type and shape of the distribution "rescues" the reliability estimation as formulated by Rouder and Haaf (2019). Additionally, in both sets of analyses, reliability estimates were highest in the proactive session tasks, indicating that theoretical motivated task manipulations may improve reliability as well.

In contrast, the stellar increase in difference score reliability was *not* matched by the between-task correlations. It has been suggested that the weak between-task correlations of classic cognitive control measures stem from a bottleneck of its reliability (Hedge, Powell, & Sumner, 2018). Our findings provide clear-cut evidence to the contrary; although the delta parameter can be estimated reliably, there is no relationship across delta parameters drawn from different tasks of cognitive control. This could indicate that cognitive control, as currently conceptualized, is not unified. In other words, classic cognitive control tasks may measure task-specific processes, rather than a single process (this follows Rey-Mermet et al., (2018)).

Although with decreasing likeliness, a measurement explanation behind the weak between-task correlations can still not be counted out completely. Our findings suggest that a difference score of average performances (i.e., the Stroop effect) does not capture individual

differences in cognitive control. In addition to the delta parameter (i.e., a difference score of *mean* performance), we examined the sigma parameter extracted from the HBM. It seems quite plausible that individual differences in cognitive control are not captured by *average* performance, but rather by *consistent* performance as indexed by individual-level standard deviation and as suggested by Unsworth (2015) and others. Such view is corroborated by our findings (Tables C and D). The sigma parameter explains as much as 25% of the covariance within our task battery. Perhaps consistently controlling cognition is an important trait with substantial differences among individuals. Interestingly, the between-task correlations of the sigma parameter show that the Stroop incongruent trial variability does not correlate much with the other interference sigma parameters. Stroop is a mainstay of cognitive control, however, the task paradigm cannot seem to pick up on individual differences. Our results suggest that the Stroop does not belong in research that measures individual differences in cognitive control.

9.1 Limitations and Future Directions

One of the potential limitations of the current study design is the fully online format of data collection. However, at the time of writing, the worldwide pandemic has shifted *most* research to an online format. Yet, it is possible that potential distractions could occur while subjects complete these tasks at home or other non-laboratory settings, subsequently influencing the results of the study. However, the nature of this multi-session study made frequent laboratory visits less optimal and more time-consuming for data collection of a large sample size.

Another limitation in the current study concerns the strategy training for the proactive condition of the AX-CPT task. As a refresher; subjects are provided with explicit information regarding the ratios of cue-probe associations, as well as receive training and practice in utilizing them to prepare the dominant responses. The main problem arises in conjunction with the test-

retest design; the explicit strategy training carries over to the retest baseline session. In other words, the retest baseline session correlates more strongly with the test proactive session, than it does with the test baseline session. This indicates that the strategy training effects task performance in the other sessions as well. Ideally, the task manipulation in the proactive condition for the AX-CPT in the DMC task battery should be reconsidered.

An important future direction is the examination of the intra-individual variability (IIV) as an individual differences index of cognitive control. For example, the relationship between IIV and important cognitive constructs could reveal more about the underlying nature the variability. At this point it is not clear whether IIV (or sigma, or standard deviation) indeed emerges as a result of some underlying cognitive processes shared between different cognitive control tasks. One specific avenue is to examine the relationship between cognitive control sigma parameters and indices of mind wandering. Theoretically, mind wandering could serve as one of many possible criterions for trial variability in cognitive control tasks (see also Unsworth, 2015). Another interesting avenue could be examination of ways to reduce this variability, through perhaps training.

9.2 Conclusion

By implementing theory-based task manipulations, we examined whether existing experimental tasks of cognitive control are a viable tool for measuring individual differences. The experimental effects and shifts toward proactive and reactive control were as hypothesized, but the current results indicate that it remains a challenge for experimental tasks to produce robust individual differences. It required hierarchical Bayesian modeling to bring out reliable estimates; traditional approaches did not render robust estimates as expected. However, the highly reliable difference score estimates did not correlate with the other cognitive control

estimates in the task battery. This indicates that reliability is not the cause of weak between-task correlations in the DMC task battery.

We suggest that there are two methodological takeaways from the current study: One, needless to say, reliability is of great importance in scientific research. A field-breaking result is nothing but a statistical artifact if it cannot be reliably reproduced. Reporting the reliability of measures ought to be a standard procedure. Here we provide two approaches to calculating individual differences reliability that are suitable to be used on cognitive behavioral tasks data. These take in consideration some psychometric pitfalls that are often ignored in common statistical software packages. And two, statistical approaches should be used thoughtfully, taking in consideration the research design and assumptions of the method used. For example, some methods assume that the distribution of the underlying data is normal, while this is not always the case. Ignoring properties of the data (e.g., distribution shape, standard deviation) can lead to unreliable and inaccurate estimates of cognitive processes. Here, we make a case for hierarchical Bayesian analyses, which allowed us to model important properties of the data resulting in more precise parameter estimation.

Through the implementation of different methods, we found that classic indices of cognitive control tasks can be estimated reliably. However, even though we addressed some prominent methodological issues, the underlying psychometric structure of cognitive control remains evasive. Finally, our results do suggest that intra-individual variability, rather than average performance difference scores, provides an exciting avenue of future research.

10 References

Alloway, T., & Alloway, R. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology*, *106*, 20–29. https://doi.org/10.1016/j.jecp.2009.11.003

Barch, D. M., Carter, C. S., Braver, T. S., Sabb, F. W., MacDonald, A., Noll, D. C., & Cohen, J. D. (2001). Selective Deficits in Prefrontal Cortex Function in Medication-Naive Patients With Schizophrenia. *Archives of General Psychiatry*, *58*(3), 280. https://doi.org/10.1001/archpsyc.58.3.280

Betancourt, M. J., & Girolami, M. (2013). Hamiltonian Monte Carlo for Hierarchical Models. *ArXiv:1312.0906 [Stat]*. http://arxiv.org/abs/1312.0906

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

Braver, T. S., Barch, D. M., Keys, B. A., Carter, C. S., Cohen, J. D., Kaye, J. A., Janowsky, J. S., Taylor, S. F., Yesavage, J. A., Mumenthaler, M. S., Jagust, W. J., & Reed, B. R. (2001). Context processing in older adults: Evidence for a theory relating cognitive control to neurobiology in healthy aging. *Journal of Experimental Psychology: General*, *130*(4), 746–763. https://doi.org/10.1037/0096-3445.130.4.746

Braver, T. S., Gray, J. R., & Burgess, G. C. (2007). Explaining the many varieties of working memory variation: Dual mechanisms of cognitive control. In *Variation in working memory* (pp. 76–106). Oxford University Press.

Braver, T. S., Paxton, J. L., Locke, H. S., & Barch, D. M. (2009). Flexible neural mechanisms of cognitive control within human prefrontal cortex. *Proceedings of the National Academy*

of Sciences of the United States of America, *106*(18), 7351–7356. https://doi.org/10.1073/pnas.0808187106

Braver, T. S., Satpute, A. B., Rush, B. K., Racine, C. A., & Barch, D. M. (2005). Context Processing and Context Maintenance in Healthy Aging and Early Stage Dementia of the Alzheimer's Type. *Psychology and Aging*, *20*(1), 33–46. https://doi.org/10.1037/0882-7974.20.1.33

Bugg, J. M. (2014). Evidence for the sparing of reactive cognitive control with age. *Psychology and Aging*, *29*(1), 115–127. https://doi.org/10.1037/a0035270

Bugg, J. M., & Braver, T. S. (2016). Proactive control of irrelevant task rules during cued task switching. *Psychological Research*, *80*(5), 860–876. https://doi.org/10.1007/s00426-015-0686-5

Bugg, J. M., & Chanani, S. (2011). List-wide control is not entirely elusive: Evidence from picture-word Stroop. *Psychonomic Bulletin & Review*, *18*(5), 930–936. https://doi.org/10.3758/s13423-011-0112-y

Bugg, J. M., & Dey, A. (2018). When stimulus-driven control settings compete: On the dominance of categories as cues for control. *Journal of Experimental Psychology. Human Perception and Performance*, *44*(12), 1905–1932. https://doi.org/10.1037/xhp0000580

Bugg, J. M., & Hutchison, K. A. (2013). Converging evidence for control of color–word Stroop interference at the item level. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(2), 433–449. https://doi.org/10.1037/a0029145

Bugg, J. M., Jacoby, L. L., & Chanani, S. (2011). Why it is too early to lose control in accounts of item-specific proportion congruency effects. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(3), 844–859. https://doi.org/10.1037/a0019957

Burgess. (1997). Theory and Methodology in Executive Function Research. *Burgess, P.W. (1997) Theory and Methodology in Executive Function Research. In: Rabbitt, P., (Ed.) Theory and Methodology of Frontal and Executive Function. Psychology Press, East Sussex, UK, Pp.81 - 116 . ISBN 9780863774857*.

Burgess, G. C., & Braver, T. S. (2010). Neural mechanisms of interference control in working memory: Effects of interference expectancy and fluid intelligence. *PloS One, 5*(9), e12861. https://doi.org/10.1371/journal.pone.0012861

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software, 76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Caruso, J. C. (2004). A Comparison of the Reliabilities of Four Types of Difference Scores for Five Cognitive Assessment Batteries. *European Journal of Psychological Assessment, 20*(3), 166–171. https://doi.org/10.1027/1015-5759.20.3.166

Chen, G., Pine, D., Brotman, M., & Smith, A. (2021). *Beyond the intraclass correlation: A hierarchical modeling approach to test-retest assessment | bioRxiv*. https://www.biorxiv.org/content/10.1101/2021.01.04.425305v1

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. https://doi.org/10.1007/BF02310555

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*(11), 671–684. https://doi.org/10.1037/h0043943

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin, 74*(1), 68–80. https://doi.org/10.1037/h0029382

Dancey, C. P., & Reidy, J. (2004). *Statistics Without Maths for Psychology*. Pearson Education.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450–466. https://doi.org/10.1016/S0022-5371(80)90312-6

Davis-Stober, C. P., Park, S., Brown, N., & Regenwetter, M. (2016). Reported violations of rationality may be aggregation artifacts. *Proceedings of the National Academy of Sciences*, *113*(33), E4761–E4763. https://doi.org/10.1073/pnas.1606997113

De Pisapia, N., & Braver, T. S. (2006). A model of dual control mechanisms through anterior cingulate and prefrontal cortex interactions. *Neurocomputing*, *69*(10), 1322–1326. https://doi.org/10.1016/j.neucom.2005.12.100

Dempster, F. N. (1993). Resistance to Interference: Developmental Changes in a Basic Processing Mechanism. In M. L. Howe & R. Pasnak (Eds.), *Emerging Themes in Cognitive Development: Volume I: Foundations* (pp. 3–27). Springer. https://doi.org/10.1007/978-1-4613-9220-0_1

Depaoli, S., & van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological Methods*, *22*(2), 240–261. https://doi.org/10.1037/met0000065

Der, G., & Deary, I. J. (2006). Age and sex differences in reaction time in adulthood: Results from the United Kingdom Health and Lifestyle Survey. *Psychology and Aging*, *21*(1), 62–73. https://doi.org/10.1037/0882-7974.21.1.62

Desimone, R., & Duncan, J. (1995). Neural Mechanisms of Selective Visual Attention. *Annual Review of Neuroscience*, *18*, 193–222. https://doi.org/10.1146/annurev.ne.18.030195.001205

Diamond, A. (2013). Executive Functions. *Annual Review of Psychology*, *64*(1), 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Draheim, C., Tsukahara, J. S., Martin, J. D., Mashburn, C. A., & Engle, R. W. (2020). A toolbox approach to improving the measurement of attention control. *Journal of Experimental Psychology. General*. https://doi.org/10.1037/xge0000783

Duchek, J. M., Balota, D. A., Tse, C.-S., Holtzman, D. M., Fagan, A. M., & Goate, A. M. (2009). The utility of intraindividual variability in selective attention tasks as an early marker for Alzheimer's disease. *Neuropsychology*, *23*(6), 746–758. https://doi.org/10.1037/a0016583

Duncan, J., Emslie, H., Williams, P., Johnson, R., & Freer, C. (1996). Intelligence and the Frontal Lobe: The Organization of Goal-Directed Behavior. *Cognitive Psychology*, *30*(3), 257–303. https://doi.org/10.1006/cogp.1996.0008

Dykiert, D., Der, G., Starr, J. M., & Deary, I. J. (2012). Age differences in intra-individual variability in simple and choice reaction time: Systematic review and meta-analysis. *PLoS ONE*, *7*(10). https://doi.org/10.1371/journal.pone.0045759

Ecker, U., Lewandowsky, S., Oberauer, K., & Chee, A. (2010). The Components of Working Memory Updating: An Experimental Decomposition and Individual Differences. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *36*, 170–189. https://doi.org/10.1037/a0017891

Edwards, J. R. (2001). Ten Difference Score Myths. *Organizational Research Methods*, *4*(3), 265–287. https://doi.org/10.1177/109442810143005

Engle, R. (2002). Working Memory Capacity as Executive Attention. *Current Directions in Psychological Science*, *11*(1), 19. https://doi.org/10.1111/1467-8721.00160

Engle, R., Conway, A., Tuholski, S., & Shisler Marshall, R. (1995). A Resource Account of Inhibition. *Psychological Science - PSYCHOL SCI*, *6*, 122–125. https://doi.org/10.1111/j.1467-9280.1995.tb00318.x

Engle, R., & Kane, M. (2004). Executive Attention, Working Memory Capacity, and a Two-Factor Theory of Cognitive Control. In *The psychology of learning and motivation: Advances in research and theory, Vol. 44* (pp. 145–199). Elsevier Science.

Engle, R. W. (2018). Working Memory and Executive Attention: A Revisit. *Perspectives on Psychological Science*, *13*(2), 190–193. https://doi.org/10.1177/1745691617720478

Enock, P. M., Hofmann, S. G., & McNally, R. J. (2014). Attention Bias Modification Training Via Smartphone to Reduce Social Anxiety: A Randomized, Controlled Multi-Session Experiment. *Cognitive Therapy and Research*, *38*(2), 200–216. https://doi.org/10.1007/s10608-014-9606-z

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*(2), 134–140. https://doi.org/10.1037/h0045156

Friedman, N. P., & Miyake, A. (2004). The Relations Among Inhibition and Interference Control Functions: A Latent-Variable Analysis. *Journal of Experimental Psychology: General*, *133*(1), 101–135. https://doi.org/10.1037/0096-3445.133.1.101

Friedman, N. P., & Miyake, A. (2017). Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *86*, 186–204. https://doi.org/10.1016/j.cortex.2016.04.023

Friedman, N. P., Miyake, A., Altamirano, L. J., Corley, R. P., Young, S. E., Rhea, S. A., & Hewitt, J. K. (2016). Stability and change in executive function abilities from late

adolescence to early adulthood: A longitudinal twin study. *Developmental Psychology*, *52*(2), 326–340. https://doi.org/10.1037/dev0000075

Friedman, N. P., Miyake, A., Young, S. E., DeFries, J. C., Corley, R. P., & Hewitt, J. K. (2008). Individual differences in executive functions are almost entirely genetic in origin. *Journal of Experimental Psychology. General*, *137*(2), 201–225. https://doi.org/10.1037/0096-3445.137.2.201

Gathercole, S. E., Pickering, S. J., Knight, C., & Stegmann, Z. (2003). Working memory skills and educational attainment: Evidence from national curriculum assessments at 7 and 14 years of age. *Applied Cognitive Psychology*, *18*(1), 1–16. https://doi.org/10.1002/acp.934

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., & Vehtari, A. (2013). *Gelman, A: Bayesian Data Analysis* (3rd edition). Taylor & Francis Ltd.

Gonthier, C., Macnamara, B. N., Chow, M., Conway, A. R. A., & Braver, T. S. (2016). Inducing Proactive Control Shifts in the AX-CPT. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01822

Gourley, E. M., Braver, T. S., & Bugg, J. M. (2016). *Dissociating proactive and reactive control: A replication and extension using color-word Stroop.* 57th annual meeting of the Psychonomics Society, Boston, MA.

Gratton, G., Coles, M. G., & Donchin, E. (1992). Optimizing the use of information: Strategic control of activation of responses. *Journal of Experimental Psychology. General*, *121*(4), 480–506. https://doi.org/10.1037//0096-3445.121.4.480

Gustavson, D. E., Panizzon, M. S., Elman, J. A., Franz, C. E., Reynolds, C. A., Jacobson, K. C., Friedman, N. P., Xian, H., Toomey, R., Lyons, M. J., & Kremen, W. S. (2018). Stability

of genetic and environmental influences on executive functions in midlife. *Psychology and Aging*, *33*(2), 219–231. https://doi.org/10.1037/pag0000230

Haines, N., Kvam, P. D., Irving, L. H., Smith, C., Beauchaine, T. P., Pitt, M. A., Ahn, W.-Y., & Turner, B. (2020). *Learning from the Reliability Paradox: How Theoretically Informed Generative Models Can Advance the Social, Behavioral, and Brain Sciences*. PsyArXiv. https://doi.org/10.31234/osf.io/xr7y3

Hallett, P. E. (1978). Primary and secondary saccades to goals defined by instructions. *Vision Research*, *18*(10), 1279–1296. https://doi.org/10.1016/0042-6989(78)90218-3

Hasher, L., Lustig, C., & Zacks, R. (2007). Inhibitory mechanisms and the control of attention. In *Variation in working memory* (pp. 227–249). Oxford University Press.

Hasher, L., Stoltzfus, E. R., Zacks, R. T., & Rypma, B. (1991). Age and inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(1), 163–169. https://doi.org/10.1037/0278-7393.17.1.163

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of*d′*. *Behavior Research Methods, Instruments, & Computers*, *27*(1), 46–51. https://doi.org/10.3758/BF03203619

Heathcote, A., Brown, S., & Mewhort, D. (2000). *Repealing the power law: The case for an exponential law of practice*.

Heathcote, A., Popiel, S., & Mewhort, D. (1991). Analysis of response time distributions: An example using the Stroop Task. *Psychological Bulletin*, *109*, 340–347. https://doi.org/10.1037/0033-2909.109.2.340

Hedge, C., Powell, G., Bompas, A., Vivian-Griffiths, S., & Sumner, P. (2018). Low and variable correlation between reaction time costs and accuracy costs explained by accumulation

models: Meta-analysis and simulations. *Psychological Bulletin*, *144*(11), 1200. https://doi.org/10.1037/bul0000164

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

Hester, R., & Garavan, H. (2004). Executive dysfunction in cocaine addiction: Evidence for discordant frontal, cingulate, and cerebellar activity. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *24*(49), 11017–11022. https://doi.org/10.1523/JNEUROSCI.3321-04.2004

Hockley, W. E., & Corballis, M. C. (1982). Tests of serial scanning in item recognition. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *36*(2), 189–212. https://doi.org/10.1037/h0080637

Hultsch, D. F., Strauss, E., Hunter, M. A., & MacDonald, S. W. S. (2008). Intraindividual variability, cognition, and aging. In *The handbook of aging and cognition*, *3rd ed* (pp. 491–556). Psychology Press.

Hutchison, K. A., Smith, J. L., & Ferris, A. (2013). Goals Can Be Threatened to Extinction: Using the Stroop Task to Clarify Working Memory Depletion Under Stereotype Threat. *Social Psychological and Personality Science*, *4*(1), 74–81. https://doi.org/10.1177/1948550612440734

Jackson, J. D., Balota, D. A., Duchek, J. M., & Head, D. (2012). White matter integrity and reaction time intraindividual variability in healthy aging and early-stage Alzheimer disease. *Neuropsychologia*, *50*(3), 357–366. https://doi.org/10.1016/j.neuropsychologia.2011.11.024

Jacoby, L. L., Lindsay, D. S., & Hessels, S. (2003). Item-specific control of automatic processes: Stroop process dissociations. *Psychonomic Bulletin & Review*, *10*(3), 638–644. https://doi.org/10.3758/BF03196526

Johnson, J. G., & Busemeyer, J. R. (2005). A Dynamic, Stochastic, Computational Model of Preference Reversal Phenomena. *Psychological Review*, *112*(4), 841–861. https://doi.org/10.1037/0033-295X.112.4.841

Jonides, J., & Nee, D. E. (2006). Brain mechanisms of proactive interference in working memory. *Neuroscience*, *139*(1), 181–193. https://doi.org/10.1016/j.neuroscience.2005.06.042

Kane, M. J., Conway, A. R. A., Miura, T. K., & Colflesh, G. J. H. (2007). Working memory, attention control, and the n-back task: A question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 615–622. https://doi.org/10.1037/0278-7393.33.3.615

Karr, J. E., Areshenkoff, C. N., Rast, P., Hofer, S. M., Iverson, G. L., & Garcia-Barrera, M. A. (2018). The unity and diversity of executive functions: A systematic review and re-analysis of latent variable studies. *Psychological Bulletin*, *144*(11), 1147–1185. https://doi.org/10.1037/bul0000160

Keye, D., Wilhelm, O., Oberauer, K., & van Ravenzwaaij, D. (2009). Individual differences in conflict-monitoring: Testing means and covariance hypothesis about the Simon and the Eriksen Flanker task. *Psychological Research*, *74*. https://doi.org/10.1007/s00426-009-0257-8

Kimberg, D. Y., & Farah, M. J. (1993). A unified account of cognitive impairments following frontal lobe damage: The role of working memory in complex, organized behavior.

Journal of Experimental Psychology: General, *122*(4), 411–428. https://doi.org/10.1037/0096-3445.122.4.411

Klauer, K. C., Schmitz, F., Teige-Mocigemba, S., & Voss, A. (2010). Understanding the role of executive control in the Implicit Association Test: Why flexible people have small IAT effects. *Quarterly Journal of Experimental Psychology*, *63*(3), 595–619. https://doi.org/10.1080/17470210903076826

Koch, I., Poljac, E., Müller, H., & Kiesel, A. (2018). Cognitive structure, flexibility, and plasticity in human multitasking-An integrative review of dual-task and task-switching research. *Psychological Bulletin*, *144*(6), 557–583. https://doi.org/10.1037/bul0000144

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kovacs, K., & Conway, A. R. A. (2016). Process Overlap Theory: A Unified Account of the General Factor of Intelligence. *Psychological Inquiry*, *27*(3), 151–177. https://doi.org/10.1080/1047840X.2016.1153946

Krumm, S., Schmidt-Atzert, L., Buehner, M., Ziegler, M., Michalczyk, K., & Arrow, K. (2009). Storage and non-storage components of working memory predicting reasoning: A simultaneous examination of a wide range of ability factors. *Intelligence*, *37*(4), 347–364. https://doi.org/10.1016/j.intell.2009.02.003

Kupitz, C. N. (2020). *Applications of Hierarchical Bayesian Cognitive Modeling* [UC Irvine]. https://escholarship.org/uc/item/0zh727fz

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G., & Banaji, M. R. (2019). Relationship between the Implicit Association

Test and intergroup behavior: A meta-analysis. *American Psychologist*, *74*(5), 569–586. https://doi.org/10.1037/amp0000364

Kvam. (2019). Modeling Accuracy, Response Time, and Bias in Continuous Orientation Judgments. *Journal of Experimental Psychology Human Perception & Performance*, *45*, 301–318. https://doi.org/10.1037/xhp0000606

Kvam, & Busemeyer, J. R. (2020). A distributional and dynamic theory of pricing and preference. *Psychological Review*, *127*(6), 1053–1078. https://doi.org/10.1037/rev0000215

Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, *12*(4), 605–621. https://doi.org/10.3758/BF03196751

Leth-Steensen, C., Elbaz, Z. K., & Douglas, V. I. (2000). Mean response times, variability, and skew in the responding of ADHD children: A response time distributional approach. *Acta Psychologica*, *104*(2), 167–190. https://doi.org/10.1016/s0001-6918(00)00019-6

LHERMITTE, F., F, L., J, D., & JL, S. (1972). ANALYSE NEUROPSYCHOLOGIQUE DU SYNDROME FRONTAL. *ANALYSE NEUROPSYCHOLOGIQUE DU SYNDROME FRONTAL*.

Liew, S. X., Howe, P. D. L., & Little, D. R. (2016). The appropriacy of averaging in the study of context effects. *Psychonomic Bulletin & Review*, *23*(5), 1639–1646. https://doi.org/10.3758/s13423-016-1032-7

Logan, G. D., & Zbrodoff, N. J. (1979). When it helps to be misled: Facilitative effects of increasing the frequency of conflicting stimuli in a Stroop-like task. *Memory & Cognition*, *7*(3), 166–174. https://doi.org/10.3758/BF03197535

Lord, F. M. (1956). The Measurement of Growth: *Educational and Psychological Measurement*. https://doi.org/10.1177/001316445601600401

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*(3, Pt.1), 151–159. https://doi.org/10.1037/h0026141

MacDonald, S. W. S., Nyberg, L., & Bäckman, L. (2006). Intra-individual variability in behavior: Links to brain structure, neurotransmission and neuronal activity. *Trends in Neurosciences*, *29*(8), 474–480. https://doi.org/10.1016/j.tins.2006.06.011

McFie, J. (1960). Psychological testing in clinical neurology. *Journal of Nervous and Mental Disease*, *131*, 383–393. https://doi.org/10.1097/00005053-196011000-00002

Mcgraw, K., & Wong, S. P. (1996). Forming Inferences About Some Intraclass Correlation Coefficients. *Psychological Methods*, *1*, 30–46. https://doi.org/10.1037/1082-989X.1.1.30

McVay, J., & Kane, M. (2011). Drifting From Slow to "D'oh!": Working Memory Capacity and Mind Wandering Predict Extreme Reaction Times and Executive Control Errors. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *38*, 525–549. https://doi.org/10.1037/a0025896

Meehl, P. E. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment*, *50*(3), 370–375. https://doi.org/10.1207/s15327752jpa5003_6

Miller, E. K., & Cohen, J. D. (2001). An Integrative Theory of Prefrontal Cortex Function. *Annual Review of Neuroscience*, *24*(1), 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167

Miyake, A., & Friedman, N. P. (2012). The Nature and Organization of Individual Differences in Executive Functions: Four General Conclusions. *Current Directions in Psychological Science*, *21*(1), 8–14. https://doi.org/10.1177/0963721411429458

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The Unity and Diversity of Executive Functions and Their Contributions to Complex "Frontal Lobe" Tasks: A Latent Variable Analysis. *Cognitive Psychology*, *41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

Moscoso del Prado, F. (2008). *A Theory of Reaction Time Distributions*.

Muchinsky, P. M. (1996). The Correction for Attenuation. *Educational and Psychological Measurement*, *56*(1), 63–75. https://doi.org/10.1177/0013164496056001004

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology*, *7*(1), 44–64. https://doi.org/10.1016/0010-0285(75)90004-3

Norman, D., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In *Cognitive Neuroscience: A reader*.

Nunnally Jr., J. C. (1970). *Introduction to psychological measurement* (pp. xv, 572). McGraw-Hill.

Orben, A., & Lakens, D. (2020). Crud (Re)Defined. *Advances in Methods and Practices in Psychological Science*, *3*(2), 238–247. https://doi.org/10.1177/2515245920917961

Paap, K. R., & Greenberg, Z. I. (2013). There is no coherent evidence for a bilingual advantage in executive processing. *Cognitive Psychology*, *66*(2), 232–258. https://doi.org/10.1016/j.cogpsych.2012.12.002

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

Pagan, A. (1984). Econometric Issues in the Analysis of Regressions with Generated Regressors. *International Economic Review*, *25*(1), 221–247. https://doi.org/10.2307/2648877

Papaspiliopoulos, O., Roberts, G. O., & Sköld, M. (2007). A General Framework for the Parametrization of Hierarchical Models. *Statistical Science*, *22*(1), 59–73. https://doi.org/10.1214/088342307000000014

Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological Science Needs a Standard Practice of Reporting the Reliability of Cognitive-Behavioral Measurements. *Advances in Methods and Practices in Psychological Science*, *2*(4), 378–395. https://doi.org/10.1177/2515245919879695

Paxton, J. L., Barch, D. M., Racine, C. A., & Braver, T. S. (2008). Cognitive Control, Goal Maintenance, and Prefrontal Function in Healthy Aging. *Cerebral Cortex*, *18*(5), 1010–1028. https://doi.org/10.1093/cercor/bhm135

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

Rabbitt, P. M. (1966). Errors and error correction in choice-response tasks. *Journal of Experimental Psychology*, *71*(2), 264–272. https://doi.org/10.1037/h0022853

Ramirez, G., Chang, H., Maloney, E. A., Levine, S. C., & Beilock, S. L. (2016). On the relationship between math anxiety and math achievement in early elementary school: The role of problem solving strategies. *Journal of Experimental Child Psychology*, *141*, 83–100. https://doi.org/10.1016/j.jecp.2015.07.014

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd edition). SAGE Publications, Inc.

Redick, T. S., Shipstead, Z., Meier, M. E., Montroy, J. J., Hicks, K. L., Unsworth, N., Kane, M. J., Hambrick, D. Z., & Engle, R. W. (2016). Cognitive predictors of a common

multitasking ability: Contributions from working memory, attention control, and fluid intelligence. *Journal of Experimental Psychology: General*, *145*(11), 1473–1492. https://doi.org/10.1037/xge0000219

Rey-Mermet, A., Gade, M., & Oberauer, K. (2018). Should we stop thinking about inhibition? Searching for individual and age differences in inhibition ability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(4), 501–526. https://doi.org/10.1037/xlm0000450

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

Richmond, L. L., Redick, T. S., & Braver, T. S. (2015). Remembering to prepare: The benefits (and costs) of high working memory capacity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(6), 1764–1777. https://doi.org/10.1037/xlm0000122

Rodebaugh, T. L., Scullin, R. B., Langer, J. K., Dixon, D. J., Huppert, J. D., Bernstein, A., Zvielli, A., & Lenze, E. J. (2016). Unreliability as a threat to understanding psychopathology: The cautionary tale of attentional bias. *Journal of Abnormal Psychology*, *125*(6), 840–851. https://doi.org/10.1037/abn0000184

Rogosa, D. (1988). Myths about longitudinal research. In *Methodological issues in aging research* (pp. 171–209). Springer Publishing Company.

Rogosa, D. (1995). *Myths and methods: "Myths about longitudinal research" plus supplemental questions*. https://doi.org/null

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

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*(4), 573–604. https://doi.org/10.3758/BF03196750

Salthouse, T. A. (2007). Implications of within-person variability in cognitive and neuropsychological functioning for the interpretation of change. *Neuropsychology*, *21*(4), 401–411. https://doi.org/10.1037/0894-4105.21.4.401

Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H.-M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General*, *136*(3), 414–429. https://doi.org/10.1037/0096-3445.136.3.414

Shallice, T., Burgess, P., Robertson, I., Roberts, A. C., Robbins, T. W., & Weiskrantz, L. (1996). The domain of supervisory processes and temporal organization of behaviour. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *351*(1346), 1405–1412. https://doi.org/10.1098/rstb.1996.0124

Sharma, L., Markon, K. E., & Clark, L. A. (2014). Toward a theory of distinct types of "impulsive" behaviors: A meta-analysis of self-report and behavioral measures. *Psychological Bulletin*, *140*(2), 374–408. https://doi.org/10.1037/a0034418

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Singh, K. A., Gignac, G. E., Brydges, C. R., & Ecker, U. K. H. (2018). Working memory capacity mediates the relationship between removal and fluid intelligence. *Journal of Memory and Language*, *101*, 18–36. https://doi.org/10.1016/j.jml.2018.03.002

Snijders, T., & Bosker, R. (1999). Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling. *Http://Lst-Iiep.Iiep-Unesco.Org/Cgi-Bin/Wwwi32.Exe/[In=epidoc1.in]/?T2000=013777/(100)*.

Snyder, H. R., Miyake, A., & Hankin, B. L. (2015). Advancing understanding of executive function impairments and psychopathology: Bridging the gap between clinical and cognitive approaches. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00328

Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201–293. https://doi.org/10.2307/1412107

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

Stan Development Team. (2020a). *22.7 Reparameterization | Stan User's Guide*. https://mc-stan.org/docs/2_26/stan-users-guide/reparameterization-section.html

Stan Development Team. (2020b). *RStan: The R interface to Stan.* (2.21.2) [Computer software]. https://mc-stan.org

Stan Development Team. (2020c). *Stan Modeling Language Users Guide and Reference Manual, 2.26.* https://mc-stan.org

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149. https://doi.org/10.3758/BF03207704

Sternberg, S. (1966). High-Speed Scanning in Human Memory. *Science*, *153*(3736), 652–654. https://doi.org/10.1126/science.153.3736.652

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

Stuss, D. T., Murphy, K. J., Binns, M. A., & Alexander, M. P. (2003). Staying on the job: The frontal lobes control individual performance variability. *Brain: A Journal of Neurology*, *126*(Pt 11), 2363–2380. https://doi.org/10.1093/brain/awg237

Unsworth, N. (2015). Consistency of attentional control as an important cognitive trait: A latent variable analysis. *Intelligence*, *49*, 110–128. https://doi.org/10.1016/j.intell.2015.01.005

Unsworth, N., Redick, T. S., Lakey, C. E., & Young, D. L. (2010). Lapses in sustained attention and their relation to executive control and fluid abilities: An individual differences investigation. *Intelligence*, *38*(1), 111–122. https://doi.org/10.1016/j.intell.2009.08.002

Unsworth, N., Redick, T. S., Spillers, G. J., & Brewer, G. (2012). Variation in working memory capacity and cognitive control: Goal maintenance and microadjustments of control. *Quarterly Journal of Experimental Psychology*, *65*(2), 326–355. https://doi.org/10.1080/17470218.2011.597865

Unsworth, N., Schrock, J. C., & Engle, R. W. (2004). Working Memory Capacity and the Antisaccade Task: Individual Differences in Voluntary Saccade Control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(6), 1302–1321. https://doi.org/10.1037/0278-7393.30.6.1302

Van Der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, *113*(4), 842–861. https://doi.org/10.1037/0033-295X.113.4.842

van der Sluis, S., de Jong, P. F., & van der Leij, A. (2007). Executive functioning in children, and its relations with reasoning, reading, and arithmetic. *Intelligence*, *35*(5), 427–449. https://doi.org/10.1016/j.intell.2006.09.001

Vandekerckhove, J. (2014). A cognitive latent variable model for the simultaneous analysis of behavioral and personality data. *Journal of Mathematical Psychology*, *60*, 58–71. https://doi.org/10.1016/j.jmp.2014.06.004

von Bastian, C. C., Blais, C., Brewer, G., Gyurkovics, M., Hedge, C., Kałamała, P., Meier, M., Oberauer, K., Rey-Mermet, A., Rouder, J. N., Souza, A. S., Bartsch, L. M., Conway, A. R. A., Draheim, C., Engle, R. W., Friedman, N. P., Frischkorn, G. T., Gustavson, D. E., Koch, I., … Wiemers, E. (2020). *Advancing the understanding of individual differences in attentional control: Theoretical, methodological, and analytical considerations*. PsyArXiv. https://doi.org/10.31234/osf.io/x3b9k

von Bastian, C., & Druey, M. (2017). Shifting Between Mental Sets: An Individual Differences Approach to Commonalities and Differences of Task Switching Components. *Journal of Experimental Psychology: General*, *146*, 1266–1285. https://doi.org/10.1037/xge0000333

Wagenmakers, E.-J., & Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological Review*, *114*(3), 830–841. https://doi.org/10.1037/0033-295X.114.3.830

Walsh, K. W. (1978). *Neuropsychology: A clinical approach* (p. 371). Churchill Livingstone.

West, R., Murphy, K. J., Armilio, M. L., Craik, F. I. M., & Stuss, D. T. (2002). Lapses of Intention and Performance Variability Reveal Age-Related Increases in Fluctuations of Executive Control. *Brain and Cognition*, *49*(3), 402–419. https://doi.org/10.1006/brcg.2001.1507

Westfall, J., & Yarkoni, T. (2016). Statistically Controlling for Confounding Constructs Is Harder than You Think. *PLOS ONE*, *11*(3), e0152719. https://doi.org/10.1371/journal.pone.0152719

Wilhelm, O., Hildebrandt, A. H., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00433

Willett, J. B. (1988). Questions and Answers in the Measurement of Change. *Review of Research in Education*, *15*, 345–422. https://doi.org/10.2307/1167368

Winne, P. H., & Belfry, M. J. (1982). Interpretive problems when correcting for attenuation. *Journal of Educational Measurement*, *19*(2), 125–134. https://doi.org/10.1111/j.1745-3984.1982.tb00121.x

Zimmerman, D. W., & Williams, R. H. (1998). Reliability of gain scores under realistic assumptions about properties of pre-test and post-test scores. *British Journal of Mathematical and Statistical Psychology*, *51*(2), 343–351. https://doi.org/10.1111/j.2044-8317.1998.tb00685.x

Zumbo, B. (1999). The simple difference score as an inherently poor measure of change: Some reality, much mythology. In *Advances in Social Science Methodology* (Vol. 5, pp. 269–304).

11 Appendix A

Table A1

Stroop (Biased) Reliability across Sessions

Measure	Split-half (95% CI)	Test-Retest (95% CI)	*M*	Range
Baseline				
Reaction Time				
Congruent	1.00 (1.00–1.00)	.91 (.88–.94)	781 ms	431 – 2706 ms
Incongruent	.99 (.98–1.00)	.93 (.90–.96)	918 ms	477 – 2851 ms
Stroop Effect	.73 (.57–.83)	.32 (.16–.47)	137 ms	-267 – 385 ms
Error				
Congruent	.93 (.89–.96)	.16 (-.01–.32)	2.2 %	0 – 24 %
Incongruent	.80 (.72–.86)	.23 (.06–.38)	5.2 %	0 – 40 %
Stroop Effect	.45 (.22–.62)	.26 (.10–.42)	3.0 %	-5 – 26 %
Proactive				
Reaction Time				
Congruent	.99 (.99–1.00)	.85 (.80–.89)	798 ms	415 – 3387 ms
Incongruent	1.00 (1.00–1.00)	.87 (.82–.91)	880 ms	450 – 3596 ms
Stroop Effect	.59 (.31–.77)	.34 (.18–.49)	83 ms	-200 – 300 ms
Error				
Congruent	.81 (.68–.90)	.69 (.58–.77)	1.2 %	0 – 27 %
Incongruent	.91 (.87–.94)	.79 (.71–.82)	2.9 %	0 – 29 %
Stroop Effect	.46 (.10–.68)	.39 (.23–.53)	1.7 %	-4 – 18 %
Reactive				
Reaction Time				
Congruent	1.00 (1.00–1.00)	.91 (.87–.93)	790 ms	428 – 3787 ms
Incongruent	1.00 (1.00–1.00)	.88 (.83–.91)	882 ms	451 – 3763 ms
Stroop Effect	.87 (.78–.92)	.33 (.17–.48)	93 ms	-480 – 479 ms
Error				
Congruent	.98 (.98–1.00)	.82 (.76–.84)	1.6 %	0 – 40 %
Incongruent	.94 (.92–.96)	.53 (.39–.64)	3.9 %	0 – 42 %
Stroop Effect	.88 (.84–.92)	.78 (.70–.84)	2.3 %	-28 – 21 %

Note. N = 126. CI = confidence interval. Split-half is an average of the test and retest phase split-half reliabilities.

Table A2

Cued Task Switching (Non-Incentivized) Reliability across Sessions

Measure	Split-half (95% CI)	Test-Retest (95% CI)	*M*	Range
Baseline				
Reaction Time				
Congruent	.98 (.98–.99)	.60 (.31–.76)	906 ms	448 – 2370 ms
Incongruent	.89 (.84–.93)	.52 (.35–.65)	983 ms	458 – 2657 ms
TRCE	.39 (.10–.61)	.30 (.13–.45)	77 ms	-319 – 921 ms
Error				
Congruent	.88 (.84–.92)	.51 (.34–.64)	3.9 %	0 – 38 %
Incongruent	.66 (.54–.74)	.46 (.31–.58)	11 %	0 – 60 %
TRCE	.52 (.38–.64)	.33 (.17–.47)	7.1 %	-12 – 56 %
Proactive				
Reaction Time				
Congruent	.99 (.98–.99)	.79 (.67–.86)	718 ms	421 – 2203 ms
Incongruent	.90 (.86–.94)	.66 (.55–.75)	780 ms	425 – 2343 ms
TRCE	.52 (.28–.68)	.38 (.22–.52)	62 ms	-236 – 683 ms
Error				
Congruent	.84 (.77–.88)	.66 (.55–.75)	4.3 %	0 – 34 %
Incongruent	.57 (.45–.68)	.52 (.38–.64)	14.9 %	0 – 56 %
TRCE	.52 (.38–.64)	.51 (.37–.63)	10.7 %	-14 – 56 %
Reactive				
Reaction Time				
Congruent	.99 (.98–.99)	.66 (.42–.79)	1003 ms	501 – 2802 ms
Incongruent	.90 (.86–.94)	.60 (.39–.74)	1098 ms	510 – 3311 ms
TRCE	.55 (.38–.69)	.46 (.31–.59)	94 ms	-642 – 967 ms
Error				
Congruent	.84 (.76–.90)	.35 (.19–.49)	1.5 %	0 – 31 %
Incongruent	.59 (.44–.70)	.41 (.26–.55)	6.7 %	0 – 56 %
TRCE	.52 (.36–.66)	.35 (.19–.49)	5.1 %	-11 – 54 %

Note. $N = 128$. CI = confidence interval; TRCE = task rule congruency effect. Split-half is an average of the test and retest phase split-half reliabilities.

Table A3

AX-Continuous Performance Task Baseline Session Reliability

Measure	Split-half (95% CI)	Test-Retest (95% CI)	*M*	Range
Reaction Time				
AX trials	.98 (.96–.98)	.63 (.43–.76)	449 ms	295 – 827 ms
AY trials	.87 (.83–.90)	.69 (.58–.78)	540 ms	376 – 835 ms
BX trials	.88 (.84–.92)	.51 (.25–.68)	516 ms	267 – 1468 ms
BY trials	.98 (.97–.98)	.63 (.19–.81)	441 ms	273 – 788 ms
PBI	.66 (.55–.75)	.31 (.10–.48)	.03	-.40 - .24
BX Interference	.68 (.56–.77)	.36 (.20–.51)	75 ms	-109 – 872 ms
Error				
AX trials	.89 (.86–.92)	.27 (.10–.43)	6.6 %	0 – 80 %
AY trials	.44 (.27–.60)	.18 (.01–.34)	7 %	0 – 44 %
BX trials	.68 (.57–.76)	.20 (.02–.37)	13.8 %	0 – 80 %
BY trials	.64 (.48–.78)	.05 (-.12–.22)	1.1 %	0 – 19 %
A no-go trials	.65 (.54–.74)	.25 (.08–.40)	11.1 %	0 – 72 %
B no-go trials	.73 (.66–.80)	.43 (.28–.56)	22.3 %	0 – 80 %
PBI	.69 (.59–.77)	.16 (-.01–.32)	-.18	-.94 - .89
d' context	.78 (.70–.84)	.36 (.16–.52)	2.85	-.23 – 4.4
A-cue bias	.56 (.42–.67)	.18 (.01–.34)	.09	-1.14 - .87
BX Interference	.62 (.50–.72)	.15 (-.01–.31)	1.08	-.52 – 2.83

Note. N = 121. CI = confidence interval; PBI = proactive behavioral index. Split-half is an average of the test and retest phase split-half reliabilities.

Table A4

AX-Continuous Performance Task Proactive Session Reliability

Measure	Split-half (95% CI)	Test-Retest (95% CI)	*M*	Range
Reaction Time				
AX trials	.98 (.97–.99)	.70 (.60–.78)	415 ms	257 – 832 ms
AY trials	.86 (.80–.90)	.68 (.57–.77)	541 ms	378 – 871 ms
BX trials	.92 (.89–.94)	.73 (.63–.80)	460 ms	259 – 1010 ms
BY trials	.98 (.98–.99)	.80 (.73–.86)	410 ms	253 – 710 ms
PBI	.78 (.70–.84)	.61 (.49–.71)	.09	-.26 - .32
BX Interference	.74 (.65–.82)	.57 (.44–.69)	51 ms	-91 – 493 ms
Error				
AX trials	.92 (.88–.94)	.59 (.46–.69)	5.7 %	0 – 80 %
AY trials	.81 (.76–.86)	.60 (.47–.70)	18.6 %	0 – 80 %
BX trials	.67 (.56–.76)	.43 (.27–.57)	10.7 %	0 – 56 %
BY trials	.59 (.40–.73)	.35 (.18–.49)	1.1 %	0 – 15 %
A no-go trials	.83 (.78–.88)	.66 (.55–.75)	17 %	0 – 80 %
B no-go trials	.82 (.78–.87)	.70 (.59–.78)	32 %	0 – 80 %
PBI	.80 (.73–.86)	.54 (.40–.65)	.16	-.89 - .94
d′ context	.81 (.73–.86)	.55 (.41–.66)	3.09	-.92 – 4.40
A-cue bias	.79 (.71–.85)	.59 (.47–.70)	.37	-1.99 – 1.47
BX Interference	.62 (.50–.72)	.28 (.11–.44)	.93	-.5 – 2.47

Note. $N = 121$. CI = confidence interval; PBI = proactive behavioral index. Split-half is an average of the test and retest phase split-half reliabilities.

Table A5

AX-Continuous Performance Task Reactive Session Reliability

Measure	Split-half (95% CI)	Test-Retest (95% CI)	*M*	Range
Reaction Time				
AX trials	.98 (.98–.99)	.75 (.61–.84)	435 ms	259 – 923 ms
AY trials	.91 (.88–.93)	.69 (.53–.79)	558 ms	373 – 905 ms
BX trials	.88 (.84–.91)	.67 (.49–.78)	546 ms	336 – 993 ms
BY trials	.98 (.98–.99)	.76 (.55–.86)	420 ms	258 – 783 ms
PBI	.52 (.37–.64)	.44 (.29–.58)	.02	-.3 - .21
BX Interference	.67 (.56–.76)	.52 (.39–.64)	125 ms	-52 – 510 ms
Error				
AX trials	.84 (.78–.88)	.55 (.41–.66)	7.2 %	0 – 47 %
AY trials	.44 (.26–.58)	.28 (.11–.44)	7.0 %	0 – 33 %
BX trials	.75 (.66–.82)	.56 (.39–.68)	11.2 %	0 – 78 %
BY trials	.73 (.60–.82)	.19 (.01–.35)	1.2 %	0 – 29 %
A no-go trials	.45 (.29–.59)	.41 (.25–.55)	8.4 %	0 – 50 %
B no-go trials	.59 (.46–.70)	.46 (.30–.59)	12.8 %	0 – 56 %
PBI	.65 (.53–.74)	.23 (.06–.39)	-.09	-.93. - .86
d′ context	.79 (.72–.85)	.66 (.54–.75)	2.93	.58 – 4.4
A-cue bias	.52 (.38–.64)	.45 (.29–.58)	.06	-.8 - .82
BX Interference	.72 (.62–.80)	.39 (.20–.55)	.93	-.27 – 3.18

Note. $N = 121$. CI = confidence interval; PBI = proactive behavioral index. Split-half is an average of the test and retest phase split-half reliabilities.

Table A6

Sternberg (Critical) Reliability across Sessions

Measure	Split-half (95% CI)	Test-Retest (95% CI)	*M*	Range
Baseline				
Reaction Time				
NN	.94 (.91–.96)	.57 (.44–.68)	834 ms	466 – 1704 ms
NP	.92 (.90–.94)	.58 (.45–.68)	878 ms	444 – 1615 ms
RN	.76 (.69–.82)	.46 (.32–.59)	951 ms	492 – 1750 ms
Recency Effect	-.02 (-.26–.24)	.20 (.02–.36)	117 ms	-201 – 480 ms
Error				
NN	.73 (.62–.81)	.28 (.11–.43)	3.6 %	0 – 56 %
NP	.84 (.78–.88)	.58 (.45–.68)	13.2 %	0 – 58 %
RN	-.04 (-.26–.22)	.45 (.29–.58)	17.3 %	0 – 60 %
Recency Effect	.20 (-.02–.40)	.33 (.16–.47)	13.8 %	-12 – 60 %
Proactive				
Reaction Time				
NN	.92 (.88–.94)	.63 (.51–.73)	834 ms	445 – 1477 ms
NP	.92 (.88–.94)	.62 (.50–.72)	845 ms	420 – 1505 ms
RN	.76 (.70–.83)	.52 (.36–.64)	1003 ms	448 – 1958 ms
Recency Effect	.18 (-.05–.42)	.19 (.02–.34)	169 ms	-180 – 560 ms
Error				
NN	.68 (.55–.78)	.42 (.27–.55)	5 %	0 – 50 %
NP	.80 (.73–.86)	.47 (.32–.60)	12.4 %	0 – 60 %
RN	.16 (-.09–.38)	.52 (.38–.64)	25.6 %	0 – 60 %
Recency Effect	.32 (.11–.49)	.39 (.23–.53)	20.6 %	-25 – 60 %
Reactive				
Reaction Time				
NN	.84 (.77–.88)	.51 (.37–.63)	851 ms	460 – 1661 ms
NP	.92 (.88–.94)	.58 (.45–.69)	856 ms	482 – 1400 ms
RN	.88 (.84–.91)	.66 (.54–.75)	963 ms	491 – 1582 ms
Recency Effect	.12 (-.15–.38)	.21 (.05–.37)	85 ms	-176 – 350 ms
Error				
NN	.54 (.32–.70)	.34 (.18–.49)	4.3 %	0 – 50 %
NP	.78 (.72–.84)	.49 (.35–.61)	10.3 %	0 – 54 %
RN	.50 (.32–.65)	.62 (.50–.71)	12.7 %	0 – 56 %
Recency Effect	.78 (.72–.84)	.42 (.27–.55)	8.3 %	-25 – 50 %

Note. N = 126. CI = confidence interval; NN = novel negatives; NP = novel positives; RN = recent negatives. Split-half is an average of the test and retest phase split-half reliabilities.

12 Appendix B

Table B1

Spearman Rho Correlations of Between-Task Selected Measures, Baseline Session.

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11
1. A-Cue	3.16	0.68											
2. BXI Error	-0.13	0.12	.19*										
3. BXI RT	67.47	71.71	.22*	-.00									
4. *d′*	2.84	0.76	.57**	.79**	.00								
5. PBI Error	0.05	0.09	.16	-.86**	.14	-.63**							
6. PBI RT	0.04	0.07	-.25**	.18*	-.83**	.14	-.34**						
7. Recency Error	0.13	0.10	-.01	-.25**	-.06	-.12	.21*	.05					
8. Recency RT	116.60	81.08	.13	-.04	-.00	.03	.08	-.10	.01				
9. Stroop Error	0.03	0.04	-.27**	-.17	.01	-.27**	.07	-.03	-.01	.04			
10. Stroop RT	138.21	65.84	.09	-.18*	.10	-.12	.20*	-.09	-.02	.08	.10		
11. TRCE Error	-0.08	0.08	.24**	.18	.03	.24**	-.06	-.01	-.08	-.02	-.08	-.14	
12. TRCE RT	78.16	120.22	.15	.05	.12	.11	-.04	-.03	.01	.09	-.23*	.03	-.26**

Note. N = 120. *M* and *SD* are used to represent mean and standard deviation, respectively. BXI = BX Interference; *d′* = d prime; PBI = Proactive Behavioral Index; Recency = recency effect; TRCE = Task Rule Congruency Effect. Test and retest phase combined.
** p < .01; * p < .05

Table B2

Spearman Rho Correlations of Between-Task Selected Measures, Proactive Session.

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11
1. A-Cue	2.84	0.69											
2. BXI Error	-0.09	0.10	.21*										
3. BXI RT	48.35	64.98	.36**	-.22*									
4. *d′*	3.13	0.90	.53**	.82**	-.15								
5. PBI Error	-0.06	0.14	.34**	-.66**	.50**	-.54**							
6. PBI RT	0.09	0.09	-.34**	.37**	-.78**	.35**	-.72**						
7. Recency Error	0.18	0.11	-.06	-.08	-.04	-.10	.06	-.07					
8. Recency RT	165.66	100.36	.02	.17	.11	.19*	-.20*	.02	-.00				
9. Stroop Error	0.02	0.02	-.33**	-.15	-.02	-.33**	.02	-.02	.05	-.05			
10. Stroop RT	82.81	53.41	-.10	-.31**	.08	-.27**	.19*	-.17	-.11	.00	.29**		
11. TRCE Error	-0.13	0.10	.06	.11	-.04	.09	-.03	.03	-.11	.03	-.01	-.16	
12. TRCE RT	32.96	64.87	.05	-.15	.08	-.13	.18*	-.08	.08	-.03	-.12	.20*	-.37**

Note. N = 120. *M* and *SD* are used to represent mean and standard deviation, respectively. BXI = BX Interference; *d′* = d prime; PBI = Proactive Behavioral Index; Recency = recency effect; TRCE = Task Rule Congruency Effect. Test and retest phase combined.
** p < .01; * p < .05

Table B3

Spearman Rho Correlations of Between-Task Selected Measures, Reactive Session.

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11
1. A-Cue	3.08	0.66											
2. BXI Error	-0.10	0.12	.29**										
3. BXI RT	125.76	63.30	.22*	.07									
4. *d′*	2.94	0.85	.57**	.87**	.13								
5. PBI Error	0.03	0.08	.02	-.85**	.01	-.76**							
6. PBI RT	0.02	0.05	-.15	.30**	-.64**	.25**	-.40**						
7. Recency Error	0.08	0.09	-.05	-.24**	-.02	-.26**	.21*	-.11					
8. Recency RT	87.80	75.81	.10	.12	.15	.19*	-.10	-.04	-.10				
9. Stroop Error	0.02	0.05	-.24**	-.32**	-.09	-.37**	.34**	-.06	-.07	.08			
10. Stroop RT	91.29	64.23	-.07	-.26**	.10	-.27**	.24**	-.19*	.13	.00	.44**		
11. TRCE Error	-0.05	0.05	.23*	.23*	.21*	.31**	-.19*	-.10	-.08	.03	-.16	-.17	
12. TRCE RT	59.67	132.11	-.02	-.03	-.02	-.03	.06	-.01	.03	-.06	-.06	.06	-.16

Note. N = 120. *M* and *SD* are used to represent mean and standard deviation, respectively. BXI = BX Interference; *d′* = d prime; PBI = Proactive Behavioral Index; Recency = recency effect; TRCE = Task Rule Congruency Effect. Test and retest phase combined.
** p < .01; * p < .05

Table B4

Pearson r Correlations of Between-Task Selected Measures, Baseline Session.

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11
1. A-Cue	3.16	0.68											
2. BXI Error	-0.13	0.12	.16										
3. BXI RT	67.47	71.71	.14	-.02									
4. *d′*	2.84	0.76	.62**	.66**	-.04								
5. PBI Error	0.05	0.09	.11	-.92**	.11	-.53**							
6. PBI RT	0.04	0.07	-.18*	.19*	-.87**	.16	-.31**						
7. Recency Error	0.13	0.10	.02	-.23*	-.00	-.13	.21*	-.00					
8. Recency RT	116.60	81.08	.09	.02	.03	.03	.01	-.13	-.01				
9. Stroop Error	0.03	0.04	-.35**	-.27**	.10	-.29**	.19*	-.09	.02	.08			
10. Stroop RT	138.21	65.84	.08	-.13	.14	-.08	.15	-.15	-.01	.04	.11		
11. TRCE Error	-0.08	0.08	.24**	.13	-.00	.23*	-.05	-.00	-.08	-.01	-.12	-.09	
12. TRCE RT	78.16	120.22	.14	-.01	.06	.05	.05	-.03	-.02	.07	-.23*	.05	-.36**

Note. N = 120. *M* and *SD* are used to represent mean and standard deviation, respectively. BXI = BX Interference; *d′* = d prime; PBI = Proactive Behavioral Index; Recency = recency effect; TRCE = Task Rule Congruency Effect. Test and retest phase combined.
** p < .01; * p < .05

Table B5

Pearson's r Correlations of Between-Task Selected Measures, Proactive Session.

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11
1. A-Cue	2.84	0.69											
2. BXI Error	-0.09	0.10	.18*										
3. BXI RT	48.35	64.98	.26**	-.35**									
4. *d′*	3.13	0.90	.51**	.78**	-.22*								
5. PBI Error	-0.06	0.14	.37**	-.57**	.48**	-.48**							
6. PBI RT	0.09	0.09	-.28**	.45**	-.79**	.40**	-.74**						
7. Recency Error	0.18	0.11	-.00	-.04	-.07	-.11	.08	-.07					
8. Recency RT	165.66	100.36	.06	.19*	.04	.19*	-.17	.03	.00				
9. Stroop Error	0.02	0.02	-.24**	-.30**	.09	-.30**	.10	-.12	.07	-.12			
10. Stroop RT	82.81	53.41	-.05	-.32**	.15	-.24**	.18*	-.19*	-.10	.03	.19*		
11. TRCE Error	-0.13	0.10	.05	.09	-.04	.08	-.04	.05	-.12	.03	-.04	-.18*	
12. TRCE RT	32.96	64.87	.00	-.15	.01	-.18	.18*	-.05	.03	-.05	-.10	.19*	-.34**

Note. N = 120. *M* and *SD* are used to represent mean and standard deviation, respectively. BXI = BX Interference; *d′* = d prime; PBI = Proactive Behavioral Index; Recency = recency effect; TRCE = Task Rule Congruency Effect. Test and retest phase combined.
** p < .01; * p < .05

Table B6

Pearson's r Correlations of Between-Task Selected Measures, Reactive Session.

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11
1. A-Cue	3.08	0.66											
2. BXI Error	-0.10	0.12	.34**										
3. BXI RT	125.76	63.30	.22*	-.04									
4. *d′*	2.94	0.85	.64**	.74**	.08								
5. PBI Error	0.03	0.08	-.08	-.91**	.16	-.68**							
6. PBI RT	0.02	0.05	-.13	.32**	-.70**	.28**	-.44**						
7. Recency Error	0.08	0.09	-.06	-.11	-.01	-.18*	.09	-.10					
8. Recency RT	87.80	75.81	.11	.06	.13	.16	-.06	-.06	.02				
9. Stroop Error	0.02	0.05	-.13	-.24**	-.02	-.21*	.26**	-.08	-.04	.14			
10. Stroop RT	91.29	64.23	-.07	-.15	.08	-.19*	.15	-.14	.10	.02	.38**		
11. TRCE Error	-0.05	0.05	.23*	.23*	.16	.30**	-.22*	-.08	.00	.06	-.10	-.12	
12. TRCE RT	59.67	132.11	.04	-.08	-.03	-.01	.10	-.01	.04	-.01	-.03	.00	-.17

Note. N = 120. *M* and *SD* are used to represent mean and standard deviation, respectively. BXI = BX Interference; *d′* = d prime; PBI = Proactive Behavioral Index; Recency = recency effect; TRCE = Task Rule Congruency Effect. Test and retest phase combined.
** p < .01; * p < .05