On the Psychometric Evaluation of Cognitive Control Tasks: An Investigation with the Dual Mechanisms of Cognitive Control (DMCC) Battery

Jean-Paul Snijder[1], Rongxiang Tang[2], Julie M. Bugg[2], Andrew R. A. Conway[3], Todd S. Braver[2]

[1] Department of Psychology, Heidelberg University

[2] Department of Psychology, Washington University in St. Louis

[3] Division of Behavioral & Organizational Sciences, Claremont Graduate University

Author Note

This research was supported by National Institutes of Health grant R37 MH066078 and through funds provided by the McDonnell Center for Systems Neuroscience at Washington University in St. Louis. The authors declare that there is no conflict of interest. We would like to thank Erin Gourley for her contributions to task programming and data collection for the DMCC project.

Correspondence concerning this article should be addressed to Jean-Paul Snijder, Psychological Institute, Heidelberg University. Email: *jean-paul.snijder@psychologie.uni-heidelberg.de*

Abstract

The domain of cognitive control has been a major focus of experimental, neuroscience, and individual differences research. Currently, however, no theory of cognitive control successfully unifies both experimental and individual differences findings. Some perspectives deny that there even exists a unified psychometric cognitive control construct to be measured at all. These shortcomings of the current literature may reflect the fact that current cognitive control paradigms are optimized for the detection of within-subject experimental effects rather than individual differences. In the current study, we examine the psychometric properties of the Dual Mechanisms of Cognitive Control (DMCC) task battery, which was designed in accordance with a theoretical framework that postulates common sources of within-subject and individual differences variation. We evaluated both internal consistency and test-retest reliability, and for the latter, utilized both classical test theory measures (i.e., split-half methods, intraclass correlation) and newer hierarchical Bayesian estimation of generative models. Although traditional psychometric measures suggested poor reliability, the hierarchical Bayesian models indicated a different pattern, with good to excellent test-retest reliability in almost all tasks and conditions examined. Moreover, within-task, between-condition correlations were generally increased when using the Bayesian model derived estimates, and these higher correlations appeared to be directly linked to the higher reliability of the measures. In contrast, between-task correlations remained low regardless of theoretical manipulations or estimation approach. Together, these findings highlight the advantages of Bayesian estimation methods, while also pointing to the important role of reliability in the search for a unified theory of cognitive control.

Keywords: cognitive control, reliability, individual differences, Dual Mechanisms of Control, hierarchical Bayesian modeling

On the Psychometric Evaluation of Cognitive Control Tasks: An Investigation with the Dual Mechanisms of Cognitive Control (DMCC) Battery

Cognitive control refers to the set of processes involved in deliberate regulation of information processing to facilitate goal-directed behavior (Miller & Cohen, 2001). Nearly a half-century of research in cognitive psychology has been devoted to the development of experimental task paradigms designed to investigate the processes involved in cognitive control (Posner & Snyder, 1975). Well-known examples from this literature include the Stroop, Simon, Flanker, Stop-Signal, Cued Task-Switching, AX-CPT, and certain variants of the Sternberg item-recognition task. Most of the research in this literature has focused on detailed investigation of individual tasks and “benchmark findings” (e.g., the classic Stroop interference effect), as a means of testing theories and models regarding core mechanisms of cognitive control (Braem et al., 2019; Bugg, 2012; Kiesel et al., 2010; Verbruggen & Logan, 2009). However, more recent work has focused on the question of whether cognitive control can be considered a domain-general construct, with individuals varying systematically (i.e., in a trait-like fashion) in cognitive control functioning. This shift in the literature has prompted a focus on analyses and measurement of individual differences in cognitive control tasks and batteries (von Bastian et al., 2020).

The current study is situated relative to other recent attempts investigating the measurement of individual differences in cognitive control function (Friedman & Miyake, 2017; Frischkorn et al., 2019; Paap & Sawi, 2016; Rey-Mermet et al., 2018; Whitehead et al., 2019). Specifically, we focus on one of the key issues that has become of recent interest and controversy within this literature; namely, whether it is fundamentally problematic to utilize classic cognitive control tasks, which were developed within the tradition of experimental psychology, to assess individual differences in control functions (Cooper et al., 2017; Hedge, Powell, & Sumner, 2018; Tucker-Drob, 2011). The cognitive control tasks developed from the experimental tradition are popular because their effects replicate under a wide variety of research settings and task conditions. This success is largely attributable to a combination of low between-subject variance and high within-subject variance. Unfortunately, an individual differences approach requires the opposite, i.e., high between-subject and low within-subject variance. As a result, when these tasks are used in individual differences research, the measures have often been found to be inconsistent and unreliable, which has been recently termed “the reliability paradox” (Hedge, Powell, & Sumner, 2018; Kucina et al., 2022; Rey-Mermet et al., 2018; Rouder & Haaf, 2019).

The development of the Dual Mechanisms of Cognitive Control (DMCC) project and task battery (Braver et al., 2021; Tang, Bugg, et al., 2021) was in part motivated by this paradox. A key distinguishing feature of the DMCC battery is that the tasks included in the battery were specifically designed to test the Dual Mechanisms of Control theoretical framework. This framework postulates distinct proactive and reactive modes of control (Braver, 2012; Braver et al., 2007), that may reflect key dimensions of individual variation in control function. The Dual Mechanisms of Control account provides a theoretical framework that decomposes cognitive control into two qualitatively distinct mechanisms – proactive control and reactive control (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, and relies upon retrieval of task goals and the rapid mobilization of processing resources following the onset of a cognitively demanding event (Braver, 2012; Braver et al., 2007). In other words, proactive control is preparatory in nature, while reactive control operates in a just-in-time manner. The DMCC task battery includes conditions that are designed to experimentally and independently bias participants towards the use of proactive and reactive control modes. In the current paper, we utilize the DMCC battery as a vehicle from which to evaluate theoretically-designed cognitive control tasks from the perspective of psychometric reliability and individual differences.

In contrast to the selection of tasks used in prior work, we explicitly developed the DMCC task battery to more closely exemplify the integrated experimental/correlational approach first advocated by Cronbach (1957). As Cronbach (1957) articulated, experimental evidence is standardly utilized to inform normative models of the structure and function of cognitive abilities, while correlational/differential data is used to investigate individual differences in those abilities and their role in real-world behavior. Ideally, the experimental and differential 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.

The goal of the current study is to test whether a task battery designed in accordance with a unifying theoretical framework, can more successfully bridge the divide between experimental and differential approaches in cognitive control research. As the DMCC battery utilizes theoretically motivated task manipulations, a critical question is whether such manipulations exhibit increased sensitivity to variation in task performance that is due specifically to differences in cognitive control as specified by the theory. In classical test theory, this proportion of variability is referred to as "true score variance" (Novick, 1966). Tasks that have high true-score variance are also expected to exhibit stronger reliability and validity (Chapman & Chapman, 1978). In the sections that follow, we briefly review the literature on individual differences in cognitive control, the approaches used to assess such individual differences, and measurement issues associated with the evaluation of task reliability.

Measuring Individual Differences in Cognitive Control

Individual differences in cognitive control are 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 decline (Hasher et al., 1991). The ability to engage cognitive control is strongly linked to working memory capacity, 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., 2013), and multi-tasking (Redick et al., 2016). Cognitive control plays an important role in contemporary theories of intelligence. By some accounts, cognitive control is considered to be the primary source of variance in overall cognitive ability (Engle & Kane, 2004; Kovacs & Conway, 2016).

Despite these established findings, a major concern in the field is that the tasks used to measure cognitive control often show poor reliability and weak correlational results. Recently, several research groups reported low task reliabilities and/or weak between-task correlations, especially with respect to tasks thought to index aspects of inhibitory control (Hedge, Powell, and Sumner, 2018, Rey-Mermet et al., 2018, and Stahl et al., 2014). For example, in the Hedge, Powell, and Sumner (2018) study, the median *test-retest* reliability across 7 classic experimental effects (e.g., Stroop, flanker) was surprisingly low, with a median of .40[1]. Similarly, across

[1]

multiple studies, the correlation between flanker (Eriksen & Eriksen, 1974) and Stroop (Stroop, 1935) effects was below .20 (Draheim et al., 2020; Gärtner & Strobel, 2019; Hedge, Powell, & Sumner, 2018; Rey-Mermet et al., 2018). Based on these and other similar dismal correlational results, Rey-Mermet et al. (2018) concluded, "we should perhaps stop thinking about inhibition as a general cognitive construct".

One of the key concerns raised by these findings is whether classic experimental tasks are suitable for examining individual differences (Tucker-Drob, 2011). The current literature suggests three important issues that complicate the measurement of individual differences in cognitive control. One, experimental tasks are designed to minimize between-subject variance, which increases group-level sensitivity to experimental manipulations, but makes it difficult to tease apart individual variation and consistently rank individuals by performance; this factor may be a primary source of poor reliability when such tasks are used in individual differences research (Hedge, Powell, & Sumner, 2018). Two, the popular use of difference scores further accentuates the issue of low reliability, because it increases the ratio of measurement error to between-subject error (Cronbach & Furby, 1970; Hedge, Powell, & Sumner, 2018). And three, the correlation between two measures of cognitive control (e.g., Stroop effect and flanker effect) will be constrained by the reliability of each measure, so conclusions drawn from the correlational results are themselves inconsistent and unreliable (Nunnally Jr., 1970; Parsons et al., 2019; Spearman, 1904). Thus, based on these reliability issues, it could be argued that the examination of relationships between individual difference measures extracted from experimental tasks (i.e., between-task relationships) maybe highly problematic in a foundational way (Spearman, 1910).

The Measurement and Reporting of Reliability

Test-retest reliability magnitude below .50 is considered poor (Koo & Li, 2016).

In addition to the concerns regarding the measurement of individual differences in experimental tasks, there are numerous issues related to the measurement and reporting of reliability itself. One of the most important issues is that reliability is actually only infrequently reported in cognitive individual differences analyses (Parsons et al., 2019). As described above, part of the reason may be that experimental researchers often have less fluency and familiarity with psychometric issues, including a confusion regarding the technical meaning of reliability as it is utilized in psychometrics. A potential source of confusion may be that the term "reliable" has different meanings in experimental versus 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 consistently gives similar rankings for individuals. This lack of concern regarding psychometric reliability may be one of the reasons it has not been typically considered as a source of poor correlational results (Flake et al., 2017; Hussey & Hughes, 2020). Conversely, based on this confusion, some results may have been erroneously reported as replicable and generalizable, perhaps propagating false standards in the field (e.g., the replication crisis).

A second and more fundamental issue is that there is currently no gold-standard procedure for estimating reliability, particularly for experimental tasks (Parsons et al., 2019). Consequently, even when reliability is reported for these tasks, it is not always clearly communicated what estimation approach was utilized, which can lead to erroneous assumptions regarding the reliability of a particular experimental measure. Relatedly, although many statistical software packages supply functionality for computing reliability, these packages assume that the data conforms to analysis-specific assumptions which may not be valid for common experimental tasks and measures. An illuminating example can be seen in the case of Cronbach's alpha, a measure of internal consistency, which is probably the most common and well-known index of reliability. 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 items and the order of the items are 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 random. More concerning, the cognitive processes involved in task performance may change across trials, as a function of practice, fatigue, sequential effects, or strategy development/deployment. If these issues are ignored, which is typically the case, then reliability estimates may not be accurate or valid. Hence, Cronbach's alpha is unsuitable for tasks designed to measure individual differences in cognitive control.

There are other issues with the use of Cronbach's alpha as a measure of split-half reliability. Formally, if the assumptions above hold, Cronbach's alpha is identical to the average of all correlations between two halves of the data. However, split-half reliability is most commonly calculated in a sample by splitting the data – once – into the first and second half or even- and odd-numbered trials, and computing the correlation between these measures. However, it has been demonstrated that split-half reliabilities based on these kinds of simple split methods are unstable. Enock et al. (2012) showed that reliabilities vary depending on which trials were used in the partitioning. They recommend applying 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., 2012; Parsons et al., 2019). This permutation-based method for calculating split-half reliability approximates Cronbach's alpha (Cronbach, 1951), while simultaneously avoiding the pitfalls described above. However, another important issue is that splitting the number of observations in half leads to underestimation. The Spearman-Brown (prophecy) formula can be applied to correct for this underestimation (corrected reliability = [2*reliability] / [1+reliability]), yet this correction approach is not well-known or frequently utilized.

A third important issue is that internal consistency reliability is not the same as test-retest reliability. The measurement and utilization of test-retest reliability can be used when the same individuals are measured on the same test on two or more assessment occasions. Test-retest reliability indices estimate the degree to which the measure provides stable rankings of individuals across time. The most well-established index of test-retest reliability is the Intraclass Correlation Coefficient (ICC), which indicates how well the measurements consistently rank-order the subjects. However, one of the complexities of ICC, which has also created some confusion in its usage, is that there are 10 distinct forms available (Mcgraw & Wong, 1996). Yet only two forms are particularly pertinent for measures from cognitive experimental tasks (for a more in-depth discussion see Koo & Li (2016).

A critical distinction in the use of ICC estimates is whether reliability is based on either consistency or the 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. An absolute agreement relationship is one in which the two measurements are expected 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 (3,1) = 1.00), but the measurements would be far from absolute agreement (ICC (2,1) = .09). Thus, the type of relationship expected is a critical consideration when deciding which form of ICC to use when calculating test-retest reliability of samples from cognitive behavioral measures. If the researcher expects systematic differences between measurement occasions (e.g., practice effects), then the preferred form of ICC is the type termed ICC (3,1) in the standard terminological conventions developed by Shrout and Fleiss (1979). Conversely, if systematic differences between occasions should be considered to be problematic for the reliability of a measure, then the ICC (2,1) type should be selected. Importantly, it is necessary for the researcher to explicitly specify which type of ICC was used for calculation, and the rationale for selection, so that no ambiguity exists with regard to interpretation.

A final issue is that the traditional analytic approaches, such as ICC, may be sub-optimal, and actually even inappropriate, when calculating test-retest reliability in cognitive experimental tasks. Specifically, traditional approaches to test-retest reliability treat summary score measures (sometimes referred to as mean point-estimates; MPE) as representative indicators of performance; yet these measures do not consider trial-to-trial variability, which in itself could be an important source of individual differences (Haines et al., 2020; Lee & Webb, 2005; Rouder & Haaf, 2019; Rouder & Lu, 2005). Indeed, Rouder and Haaf (2019) have presented evidence that by ignoring trial-to-trial variability, test-retest reliability is “greatly” attenuated (see also von Bastian et al., 2020). As an alternative approach, newer analytic methods, involving hierarchical modeling (also termed multilevel or linear mixed effects modeling), have been introduced for measuring reliability, which simultaneously assess between- and within-subject (i.e., trial-by-trial) variation. Hierarchical modeling is a statistical framework for modeling data that have a natural hierarchical structure. For example, data from cognitive-behavioral tasks often have trials within subjects and subjects within groups (Gelman et al., 2013). By restructuring a model hierarchically, all individuals are considered in two contexts: in isolation, to determine how behavior varies across trials, and as a contributing member of a group, to determine how behavior varies across the group. This increases the number of available parameters from one (i.e., MPE) to multiple (e.g., mean, standard deviation). The model can now distribute uncertainty (e.g., measurement error) that exist in the data over those multiple parameters, which results in more precise estimates at both the individual and group levels (Kupitz, 2020). In particular, hierarchical models provide the means to appropriately correct for the attenuation of reliability that may occur when using more traditional methods.

Additionally, these recent efforts have also pointed to the advantages of Hierarchical *Bayesian* models (HBM), relative to classic “frequentist” approaches. A key advantage of the HBM approach is that it can be used to specify a single model that *jointly* captures the uncertainty at both the individual- and group-level. Even in a typical study that involves a modest number of participants, each performing a limited number of trials with the observed data confounded by measurement error, HBM can provide reasonable estimates of performance, by assuming that the data are generated from a population of infinite trials (Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). A second advantage of HBM is it enables explicit specification of distributions and associated parameters, which best fits a generative approach in which individual trial performance measures are thought to reflect samples drawn from these distributions. Among others, Haines et al. (2020) highlight the advantages of generative models, by suggesting that models more accurately “simulate data consistent with true behavioral observations *at the level of individual participants*”. In contrast to HBM, frequentist methods of accounting for hierarchical sources of variability, such as structural equation modeling or classical attenuation corrections, do not provide a natural framework for generative modeling (Kurdi et al., 2019; Westfall & Yarkoni, 2016).

This brief review of the current state of research on individual differences in cognitive control function suggests that a barrier to progress is the lack of knowledge on the part of researchers coming from the cognitive experimental tradition, regarding some of the psychometric complexities associated with individual difference measurement. A potential remedy is for researchers to be more explicit regarding assumptions that are being utilized regarding measurement method. Part of this explicitness relates to the reporting of measurement reliability and the analytic approach used for estimation. Moreover, when possible, estimates of both internal consistency (i.e., permutation-based split-half) and temporal stability (i.e., test-retest, ICC) forms of reliability should be assessed and reported. Finally, further investigation and comparison is needed between traditional frequentist and Bayesian approaches to estimation, since the use of Bayesian approaches in individual differences analyses is a relatively new development in the literature.

Current Study

Here we provide an evaluation focused on the utility of cognitive control measures for individual differences research purposes; specifically, we examine the psychometric issues described above within the context of the DMCC battery. In particular, a key objective associated with the development of the DMCC battery was to examine how experimental manipulation of cognitive control mode affects individual difference properties of classic cognitive control tasks (Stroop, AX-CPT, Cued Task-Switching, and Sternberg). Our primary hypothesis was that these task manipulations would generate new sources of between-subject variance, allowing for more reliable measurement of individual differences in cognitive control function (Cooper et al., 2017). More specifically, by employing task variants that induce the proactive and reactive control modes, respectively, a more mode-specific variance structure is hypothesized to emerge for each task variant (i.e., proactive and reactive variants). This should lower measurement error and, hence, improve both experimental robustness and individual differences reliability. In prior work, we demonstrated that the DMCC tasks generally exhibit robust effects of the experimental manipulations at the group level (Tang, Bugg, et al., 2021). However, it remains an open question as to whether the tasks also demonstrate strong psychometric reliability as individual difference measures of cognitive control ability. Consequently, we sought to assess task reliability in a systematic and comprehensive manner.

Another important focus of the paper was to compare traditional and the newer HBM approaches described above, for the assessment of psychometric reliability. The first set of analyses thus report reliability, both internal consistency and test-retest, employing traditional approaches based on summary score measures (MPEs) from each participant. In contrast, for the second set of analyses we implement hierarchical methods to incorporate modeling of trial-to-trial variability (i.e., individual-level standard deviation) (Rouder & Haaf, 2019). Specifically, we directly compare the traditionally derived test-retest reliability measures with those derived from the HBM approach. Our second hypothesis was that traditional approaches would substantially under-estimate the degree of reliability present in cognitive control tasks, replicating prior findings (Rouder & Haaf, 2019). As a final analysis, we examined correlations present in the DMCC, both within-task (between control mode conditions) and between-task (same condition, across task), using both analytic approaches. Our third hypothesis was that Bayesian parameter estimates, if more reliable, would also be more suitable for individual differences analyses that address the question of whether cognitive control can be considered a domain-general construct (i.e., with individuals varying in a consistent, trait-like manner).

Method

Subjects

Subjects were recruited via the Amazon Mechanical Turk (MTurk) online platform. 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 consent was given, the web-links for the first session of the study were made available on MTurk. Subjects were not restricted with regard to age range, and as such a wide range was included in the sample (N=128; 22-64, M=37.11, SD=9.90; 82 females, 46 males).

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 completed the sessions at a rate of 5 per week, i.e., taking 6 weeks to complete the full protocol. Baseline task variants were completed during the first and fourth week, the reactive task variants during the second and fifth week, and the proactive task variants were completed during the third and sixth week. 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, along with 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.

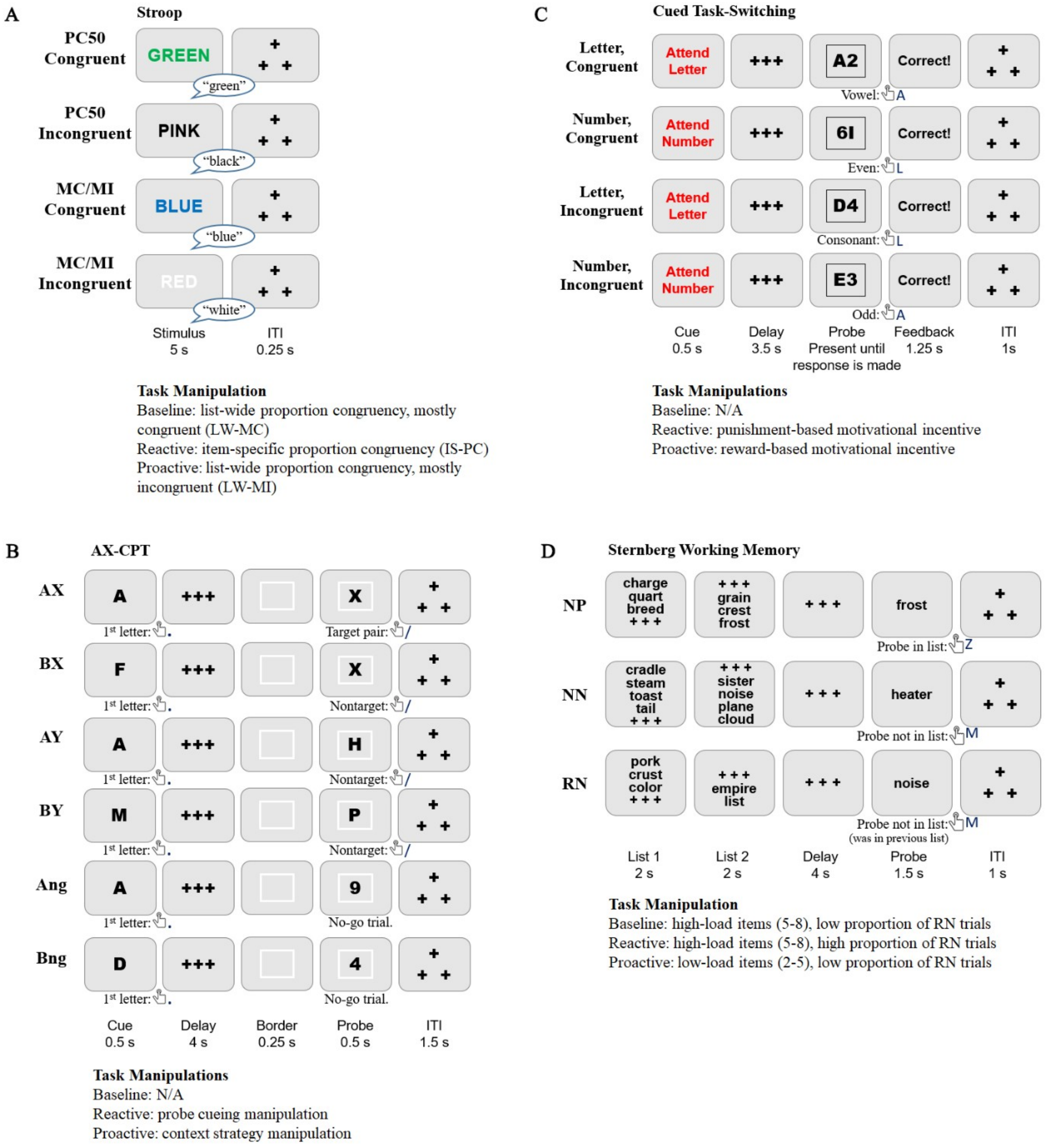
For each completed session, the experimenter checked for overall accuracy and completion of each task and questionnaire to make sure that subjects were complying with instructions and maintaining sufficient attention to the task. A criterion of 60% accuracy and response rate was used to determine whether the data would be included, and the subject invited to remain in the study. For each task 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.

Task Paradigms

Here we present a schematic representation of the tasks and their manipulations (see Figure 1). A more detailed task description is provided in Appendix C, so as not to increase the length of the primary text (see also Tang et al, 2021 for the theoretical rationale behind task manipulations).

Figure 1

DMCC Task Paradigms and Overview of Session Manipulations



Task Manipulation
Baseline: list-wide proportion congruency, mostly congruent (LW-MC)
Reactive: item-specific proportion congruency (IS-PC)
Proactive: list-wide proportion congruency, mostly incongruent (LW-MI)

Task Manipulations
Baseline: N/A
Reactive: punishment-based motivational incentive
Proactive: reward-based motivational incentive

Task Manipulations
Baseline: N/A
Reactive: probe cueing manipulation
Proactive: context strategy manipulation

Task Manipulation
Baseline: high-load items (5-8), low proportion of RN trials
Reactive: high-load items (5-8), high proportion of RN trials
Proactive: low-load items (2-5), low proportion of RN trials

Note. For a more detailed description, see Appendix C.

Data Pre-Processing

To facilitate comparison of results across task paradigms, subjects who failed to complete all 30 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. We suggest that cut-off decisions are made based on the sample, and not on the tasks in this battery. 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 winsorization 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.

Data Analyses

We assessed individual differences reliability (both split-half and test-retest) of the measures taken from the four DMCC tasks within each of three sessions (e.g., baseline, proactive, reactive). The analyses reported in the main text focused on the critical conditions of the tasks (e.g., Stroop biased condition, task-switching biased condition, Sternberg list-length 5 condition). The critical conditions were designed specifically to allow for comparison across tasks and analytic methods. Full descriptive statistics and experimental results by session, task, and trial type are reported in Tang et al. (2021). Additional reliability analyses (using traditional approaches only) of other non-critical conditions are reported in the Appendix. In addition to examining the reliability of each critical condition measure, we also examined the strength of correlation between measures, focusing on both within-task, between-condition correlations (e.g., Stroop baseline vs. Stroop proactive) and between-task, same-condition correlations (e.g., AX-CPT reactive vs. Sternberg reactive). If reliability serves as a bottleneck that limits the magnitude of between-measure correlations, then one might predict that measures showing higher reliability would also show stronger correlations.

Reliability Estimates: Traditional Approach

Both internal consistency and test-retest forms of reliability were calculated, based on traditional psychometric approaches. 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 are reported as intraclass correlation coefficients (ICC). Because practice effects are expected to occur from session to session and from test to retest phases, the ICC relationship parameter was examined as both absolute agreement (ICC(2,1)) and consistency (ICC(3,1)), as per the Shrout and Fleiss (1979) convention. The former is sensitive to changes in the mean between repeated measures, whereas the latter appropriately corrects for such changes. Here, we report both forms for comparison purposes.

Reliability Estimates: Hierarchical Bayesian Model

In addition to the traditional psychometric approach to test-retest reliability estimation, HBM was also used to generatively model the reaction time difference score effects from the four tasks in 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. To facilitate the comparison between traditionally and HBM derived estimates, we limit the examination to a reaction time model only. This approach has the added benefit that a single modeling approach can be utilized for all conditions in all tasks. Furthermore, specifying such a generative model encapsulates the shared assumptions among the tasks: (1) reaction time cannot be negative; (2) reaction time responses vary around some central tendency (this is ignored with MPE); (3) the central tendency varies per subject; (4) within-individual (i.e., trial-by-trial) variability varies per subject; and (5) 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 study. In the HBM approach, it is important that estimation of test-retest reliability considers trial variability at the individual-level; hence, the individual-level distribution is defined first, followed by the group-level distribution. Given the additional complexity and lower reader familiarity with the HBM approach, we next provide an elaborated description of these distributions and parameters.

Individual-level reaction time response distributions are here conceptualized as coming from a lognormal distribution, satisfying the 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, the distribution parameters are not expected 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.

$RT_{i,c,p}$ *Lognormal*¿
¿

Formally, in equation 4, $RT_{i,c,p}$is the observed reaction time data for subject *i* = {1,…, *N}*, in condition *c* = {control, interference}^{2}, during phase *p* = {test, retest}. *Lognormal*¿

[2]

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 are exponentially transformed.

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):

$$\mu_{i,c,p} \quad Normal\left(\mu_{mean,c,p}, \mu_{sd,c,p}\right)$$
$$\text{¿}\sigma_{i,c,p} \quad Normal\left(\sigma_{mean,c,p}, \sigma_{sd,c,p}\right)$$
$$(5)$$

By defining these prior models, the group-level distribution allows for the pooling of information across subjects. Each of the individual-level parameters, $\mu_{i,c,p}$ and $\sigma_{i,c,p}$, inform the

[2] 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 means and standard deviations, $\mu_{mean,c,p}, \mu_{sd,c,p}$ and $\sigma_{mean,c,p}, \sigma_{sd,c,p}$, 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 will notice that the group-level distributions are both modeled as normal, whereas the individual-level distributions are lognormal. 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 ¿¿) corresponds to a lognormal distribution.

Another key aspect of HBM is the definition of prior probability distribution, which expresses a prior belief about an underlying distribution of interest. Here, parameter estimation is rather robust to prior models, because the priors are rather diffuse and the sample sizes of observed data are relatively large. The prior model for the group-level mean parameters were specified as normal.

$$\mu_{mean,c,p} \quad Normal(0,1)$$
$$¿\sigma_{mean,c,p} \quad Normal(0,1)$$
$$(6)$$

The prior model for the group-level standard deviations parameters were specified as half-normal (i.e., 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.

$$\mu_{sd,c,p} \quad Half-Normal(0,1)$$
$$\text{¿}\sigma_{sd,c,p} \quad Half-Normal(0,1)$$
$$(7)$$

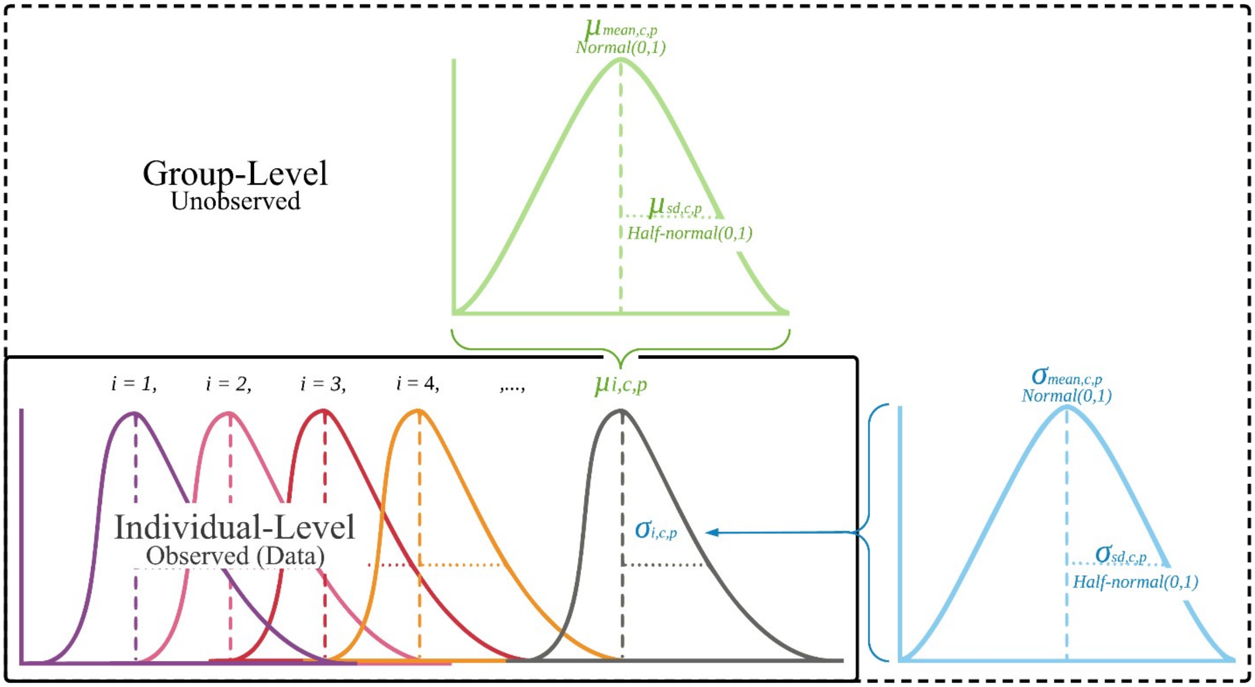
All model parameters were estimated with Stan (Stan Development Team, 2020b) through an interface in R, called RStan (Stan Development Team, 2020a). All models were fit with 3 chains of 3000 iterations after 1000 warm-up iterations. For each of the four tasks in the task battery, the model was fit three times (e.g., once for each task-variant), resulting in 12 model fits. From the model fits we extracted three families of parameters; mu, sigma and also delta. 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).

$$\Delta_{i,test} = \mu_{i,interference,test} - \mu_{i,control,test}$$
$$\text{¿}\Delta_{i,retest} = \mu_{i,interference,retest} - \mu_{i,control,retest}$$
$$(8)$$

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. All R scripts and the Stan model file are available on https://osf.io/pqvga/. A graphical representation of the model is included as well (see Figure 2). The extracted delta, mu, and sigma parameters for each task and session combination are available on https://osf.io/pqvga/. All relevant convergence statistics have been extracted and are visually presented on https://osf.io/pqvga/ as well.

Figure 2

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.

Between-Measure Correlations

For computation of the comprehensive between-task correlations that are reported in the Appendix, we utilized Spearman's rho (ρ). In particular, Spearman's rho (ρ) is a good non-parametric substitute for the parametric Pearson's r, since 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). Thus, Spearman's rho will likely provide a more robust alternative, since the Pearson's r assumptions are not likely to be met. However, for the between-task and within-task analyses discussed in the Results section below, the focus was on *reaction time* indices associated with common difference score measures (e.g., RT Stroop effect). Hence, with the linearity assumption met, we employed Pearson r correlations for the latter hierarchical Bayesian within-, and between-task, correlational analyses.

Results

Reliability Estimates: Traditional Approach

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. Although the aggregate measures are briefly discussed, only the difference score results are presented here due to their theoretical importance as measures of cognitive control, and within the DMCC battery (Tang, Bugg, et al., 2021). Table 1 presents both the split-half and test-retest reliability estimates for RT, computed separately for each control mode condition (baseline, reactive, proactive), for each task paradigm (3x3x4 = 36 estimates total). The corresponding 36 error rate estimates are shown in Table 2. In addition, for the AX-CPT task, four additional derived indices were also examined in addition to the difference scores (A-cue bias, d'-context, and Proactive Behavioral Index (PBI) for both RT and errors; see Table 3). These AX-CPT derived estimates have been commonly employed as theoretically-sensitive measures of cognitive control in this task, and have also been the focus of prior psychometric investigations (Boudewyn et al., 2015; Cohen et al., 1999; Richmond et al., 2015; Stawarczyk et al., 2014). Consequently, they were also of particular interest, to determine whether psychometric properties were improved within the context of the DMCC battery and experimental manipulations. For ease of interpretation, estimates of test-retest 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.

As expected, the reliabilities of difference score measures were weaker than the reliabilities of aggregate measures. For example, the split-half reliability for Stroop incongruent RT was on average .99 across sessions, Stroop congruent RT was on average .99 across sessions (see Appendix A), but the reliability of the RT Stroop effect was on average .73 across sessions. The same general pattern is observed for the test-retest reliability RT estimates: .89, .89, .22, respectively. This pattern is observed across all tasks, for both split-half and test-retest reliability estimates, for both RT and accuracy measures.

Table 1

Reaction Time Reliability across Sessions

Measure	Split-half (95% CI)	ICC2,1 (95% CI)	ICC3,1 (95% CI)	*M*	Range
Baseline					
Stroop Effect	.73 (.57–.83)	.32 (.16–.47)	.34 (.17–.48)	137 ms	-267 – 385 ms
BX Interference	.68 (.56–.77)	.36 (.20–.51)	.50 (.35–.62)	75 ms	-109 – 872 ms
TRCE	.39 (.10–.61)	.30 (.13–.45)	.30 (.13–.45)	77 ms	-319 – 921 ms
Recency Effect	-.02 (-.26–.24)	.20 (.02–.36)	.20 (.02–.36)	117 ms	-201 – 480 ms
Proactive					
Stroop Effect	.59 (.31–.77)	.34 (.18–.49)	.34 (.18–.49)	83 ms	-200 – 300 ms
BX Interference	.74 (.65–.82)	.57 (.44–.69)	.51 (.36–.63)	51 ms	-91 – 493 ms
TRCE	.52 (.28–.68)	.38 (.22–.52)	.36 (.20–.50)	62 ms	-236 – 683 ms
Recency Effect	.18 (-.05–.42)	.19 (.02–.34)	.20 (.02–.36)	169 ms	-180 – 560 ms
Reactive					
Stroop Effect	.87 (.78–.92)	.33 (.17–.48)	.33 (.17–.48)	93 ms	-480 – 479 ms
BX Interference	.67 (.56–.76)	.52 (.39–.64)	.47 (.32–.60)	125 ms	-52 – 510 ms
TRCE	.55 (.38–.69)	.46 (.31–.59)	.43 (.28–.57)	94 ms	-642 – 967 ms
Recency Effect	.12 (-.15–.38)	.21 (.05–.37)	.22 (.05–.38)	85 ms	-176 – 350 ms

Note. Split-half is an average of the test and retest phase split-half reliabilities. ICC2,1 is a two-way random effects, absolute agreement, single rater Intraclass Correlation Coefficient; a measure of test-retest reliability. ICC3,1 is a two-way mixed effects, consistency, single rater Intraclass Correlation Coefficient; a measure of test-retest reliability. CI = confidence interval. *M* = mean.

Table 2

Error Rate Reliability across Sessions

Measure	Split-half (95% CI)	ICC2,1 (95% CI)	ICC3,1 (95% CI)	*M*	Range
Baseline					
Stroop Effect	.45 (.22–.62)	.26 (.10–.42)	.27 (.10–.42)	3.0 %	-5 – 26 %
BX Interference	.62 (.50–.72)	.15 (-.01–.31)	.18 (.01–.35)	1.08	-.52 – 2.83
TRCE	.52 (.38–.64)	.33 (.17–.47)	.24 (.07–.40)	7.1 %	-12 – 56 %
Recency Effect	.20 (-.02–.40)	.33 (.16–.47)	.27 (.11–.43)	13.8 %	-12 – 60 %
Proactive					
Stroop Effect	.46 (.10–.68)	.39 (.23–.53)	.39 (.24–.53)	1.7 %	-4 – 18 %
BX Interference	.62 (.50–.72)	.28 (.11–.44)	.41 (.25–.54)	.93	-.50 – 2.47
TRCE	.52 (.38–.64)	.51 (.37–.63)	.52 (.38–.64)	10.7 %	-14 – 56 %
Recency Effect	.32 (.11–.49)	.39 (.23–.53)	.28 (.12–.44)	20.6 %	-25 – 60 %
Reactive					
Stroop Effect	.88 (.84–.92)	.78 (.70–.84)	.78 (.70–.84)	2.3 %	-28 – 21 %
BX Interference	.72 (.62–.80)	.39 (.20–.55)	.59 (.46–.69)	.93	-.27 – 3.18
TRCE	.52 (.36–.66)	.35 (.19–.49)	.33 (.16–.47)	5.1 %	-11 – 54 %
Recency Effect	.78 (.72–.84)	.42 (.27–.55)	.34 (.18–.49)	8.3 %	-25 – 50 %

Note. Split-half is an average of the test and retest phase split-half reliabilities. ICC2,1 is a two-way random effects, absolute agreement, single rater Intraclass Correlation Coefficient; a measure of test-retest reliability. ICC3,1 is a two-way mixed effects, consistency, single rater Intraclass Correlation Coefficient; a measure of test-retest reliability. CI = confidence interval. *M* = mean.

Table 3

AX-CPT Derived Indices Reliability across Sessions

Measure	Split-half (95% CI)	ICC2,1 (95% CI)	ICC3,1 (95% CI)	*M*	Range
Baseline					
A-cue bias	.56 (.42–.67)	.18 (.01–.34)	.29 (.12–.45)	.09	-1.14 - .87
d' context	.78 (.70–.84)	.36 (.16–.52)	.44 (.28–.57)	2.85	-.23 – 4.4
$\text{PBI}_{\text{error}}$	.69 (.59–.77)	.16 (-.01–.32)	.11 (-.07–.28)	-.18	-.94 - .89
PBI_{rt}	.66 (.55–.75)	.31 (.10–.48)	.34 (.17–.49)	.03	-.40 - .24
Proactive					
A-cue bias	.79 (.71–.85)	.59 (.47–.70)	.54 (.40–.65)	.37	-1.99 – 1.47
d' context	.81 (.73–.86)	.55 (.41–.66)	.62 (.50–.72)	3.09	-.92 – 4.40
$\text{PBI}_{\text{error}}$	.80 (.73–.86)	.54 (.40–.65)	.57 (.43–.68)	.16	-.89 - .94
PBI_{rt}	.78 (.70–.84)	.61 (.49–.71)	.58 (.44–.69)	.09	-.26 - .32
Reactive					
A-cue bias	.52 (.38–.64)	.45 (.29–.58)	.36 (.20–.51)	.06	-.80 - .82
d' context	.79 (.72–.85)	.66 (.54–.75)	.64 (.52–.74)	2.93	.58 – 4.4
$\text{PBI}_{\text{error}}$	.65 (.53–.74)	.23 (.06–.39)	.42 (.26–.56)	-.09	-.93. - .86
PBI_{rt}	.52 (.37–.64)	.44 (.29–.58)	.40 (.24–.55)	.02	-.30 - .21

Note. Split-half is an average of the test and retest phase split-half reliabilities. ICC2,1 is a two-way random effects, absolute agreement, single rater Intraclass Correlation Coefficient; a measure of test-retest reliability. ICC3,1 is a two-way mixed effects, consistency, single rater Intraclass Correlation Coefficient; a measure of test-retest reliability. CI = confidence interval. *M* = mean.

For the Stroop, cued task-switching, and AX-CPT summary score estimates the reliability results yield mixed conclusions. The split-half estimates indicate mostly moderate to good reliability, for both RT and error rate ($\bar{x}$ = .65, range = .39 – .88). However, the test-retest estimates indicate poor reliability, regardless of which ICC computation was used. ICC2,1: $\bar{x}$ = .40, range = .15 – .78; ICC3,1: $\bar{x}$ = .42, range = .11 – .78. Unfortunately, the session level manipulations (i.e., proactive and reactive variants) 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 was not a robust effect.

The reliability of the AX-CPT derived indices revealed a similar pattern as the difference score measures; the split-half reliability estimates were stronger than test-retest estimates. In contrast, two novel and interesting patterns emerged. First, all four proactive session derived indices were internally consistent, with split-half estimates ranging from .78–.81. Second, split-half estimates for *d'-context* exceeded .75 in all sessions and thus is considered to be internally consistent as well. This suggest that the reliability of the *d'-context* and the proactive indices will not pose a bottleneck when used to examine between-measure correlations.

In the Sternberg task, the recency effect measure was found to be generally unreliable, in both RT and error rate. The poor reliability and high variability of the Sternberg estimates may stem from the task 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. It is therefore not advised to calculate a traditional difference score from the current Sternberg paradigm for use in individual differences research.

Overall, the reliability analyses computed in the traditional manner did not support our primary hypothesis that theoretically-based task manipulations would improve reliability estimates. 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. We next examined whether the reliability analyses produce different results when computed using HBM approaches to estimation.

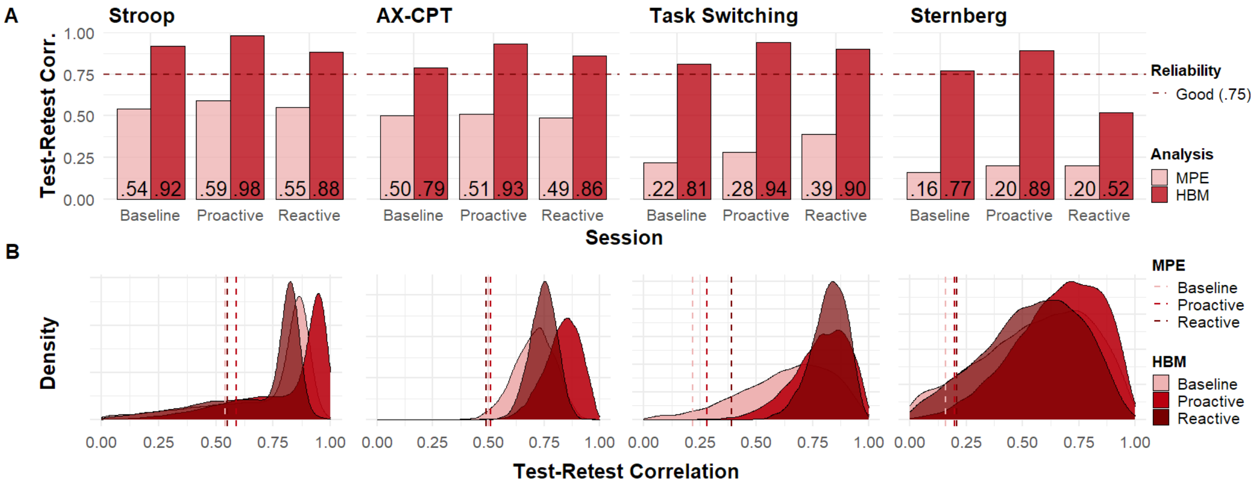
Reliability Estimates: Hierarchical Bayesian Modeling Approach

As shown in the first set of analyses, we were not able to extract reliable individual differences from experimental task difference score measures. The goal of the second set of analyses was to examine whether modeling trial-level variability improved reliability estimation of the data from the DMCC task battery. Test-retest reliability 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 Figure 3, indicated as HBM. Importantly, test-retest reliability 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 (ICC), because much of the variance has been modeled out by the sigma parameter, and to replicate the generative modeling approach of prior work similar to the current study (i.e., Haines et al., 2020; Rouder & Haaf, 2019). For a comparison between the traditional and HBM approach, the corresponding mean point-estimate of test-retest reliability (also using Pearson r to increase comparability) is provided as well in Figure 3. As guidelines for test-retest reliability, we again follow Koo and Li's (2016) thresholds (i.e., respectively, poor, moderate, good, excellent : < .50, .50 – .75, .75 – .90, > .90). Although those guidelines are for ICC, commonly accepted test-retest correlation guidelines based on Pearson's product-moment correlation coefficient do not exist to our knowledge.

In contrast to the traditional psychometric approach to estimating test-retest reliability (i.e., based on mean point estimates), which indicated poor-to-moderate test-retest reliability ($\bar{x}$ = .39), the HBM extracted estimates of test-retest reliability could be classified as good to excellent (all above .75, $\bar{x}$ = .85), with the only possible exception being the Sternberg recency effect in the reactive condition (*r* = .52). The strong reliability estimates obtained using the HBM approach are consistent with Haines et al., (2020), and Rouder and Haaf (2019). The test-retest estimates of the delta parameter indicate that HBM can indeed provide reliable individual differences from cognitive control tasks, even with a difference score index. An additional interesting pattern emerged when comparing test-retest reliability in the different control mode conditions. In particular, reliability was highest for the proactive session ($\bar{x}$ = .94; vs. $\bar{x}$ = .82 for baseline, and $\bar{x}$ = .79 for reactive), as was the case with ICC from the first set of analyses.

Figure 3

Test-Retest Reliability Estimates of Difference Score Parameters from the DMCC Task Battery



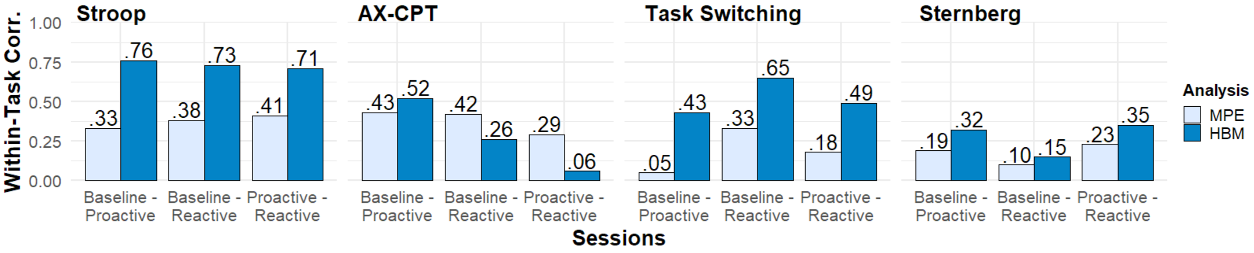
Note. HBM = Pearson correlation coefficient of delta estimates obtained by Hierarchical Bayesian Modeling; MPE = Pearson correlation coefficient obtained from traditional Mean Point Estimates approach; *n* ranges between 106 and 122; different *n* sample sizes due to additional multivariate outlier removal. Panel A: Distribution of observed reliability estimates, split by analysis type for comparison. Panel B: Density plot to visualize uncertainty of HBM delta estimate, dashed line of respective MPE estimates for comparison of reliability magnitude.

Between-Measure Correlations

Next, our analyses compared the traditional MPE estimates with the HBM derived ones, when examining correlations between conditions. We began by focusing on correlations within the same task, between sessions (see Figure 4). Because these are within-task correlations, we expected them to be consistently positive, and overall relatively high, since the experimental manipulations of cognitive control mode are quite subtle. Thus, they provided a potentially more useful testbed from which to examine the relationship between reliability of measures and their correlations. Interestingly, again a clear pattern emerged differentiating the two approaches. In particular, the within-task correlations derived with the traditional approach were moderate ($\bar{x}$ = .31), with a maximum correlation (between AX-CPT baseline and AX-CPT proactive) of *r* = .43, and 5 of the 12 correlations below *r* = .25. In contrast, for the HBM derived correlations, the values were substantially higher ($\bar{x}$ = .45), with a maximum correlation (between Stroop baseline and Stroop proactive) of *r* = .76, and only 2 of the 12 correlations below *r* = .25. Indeed, 10 of the 12 correlations were numerically greater when based on the HBM estimates; the only exceptions involved the AX-CPT reactive condition (baseline - reactive, proactive - reactive).

Figure 4

Within-Task Correlations of Difference Score Parameters from the DMCC Task Battery

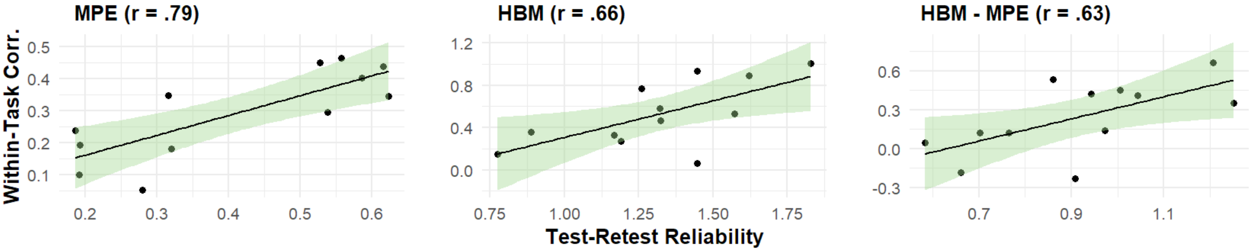


Note. HBM = Pearson correlation coefficient of delta estimates obtained by Hierarchical Bayesian Modeling; MPE = Pearson correlation coefficient obtained from traditional Mean Point Estimates approach; *n* = 104. Distribution of observed correlations within-task, split by analysis type for comparison.

To further examine the relationship between the test-retest reliabilities and within-task correlations, we also tested an even stronger hypothesis: that the increase in reliability observed in the HBM estimates would predict the increase in within-task correlation that we observed. To conduct these analyses, we first computed r-to-z transformations, to linearize the r-values. Indeed, the results were supportive of the hypothesis (see Figure 5). For the MPE-derived estimates, the average reliability was highly predictive of within-task correlation (r=.79), and a similar predictive relationship was found for the HBM-derived estimates (r=.66). Moreover, when looking at the increased reliability of the HBM-derived estimates relative to MPE, this was also strongly predictive of the increase in within-task correlation value (r=.63). Thus, when examining the within-task relationships, we find clear support for the hypothesis that the improvement in correlation magnitude that we observed in the HBM extracted values is related to the gain in reliability that was also observed.

Figure 5

Relationship Between Reliability and Within-Task Correlations



Note. MPE = Pearson correlation coefficient obtained from traditional Mean Point Estimates approach; HBM = Pearson correlation coefficient of delta estimates obtained by Hierarchical Bayesian Modeling; *n* = 104. Both reliability and within-task correlations were r-to-z transformed. Grey area is 95% confidence interval around linear regression line.

Lastly, we conducted a more comprehensive examination of between-task correlations present in the DMCC battery, first using the traditional MPE estimates. Because of the large number of tasks, conditions, and measures, we relegate full reporting of these correlations to Appendix B, and only provide a brief summary here. In total, we examined 198 between-task correlations, but of these only 12 were above *r* = .25, with the median correlation of *r* = .13. This value 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).

We then focused on between-task, same-condition correlations (e.g., correlation of Stroop baseline to AX-CPT baseline) and compared between traditional MPE and HBM approaches to test-retest reliability of key difference score measures (see Table 4). With both approaches, a similar pattern emerged, with all of the 18 correlations between -.25 and +.25. Moreover, there was no consistent difference between the correlations computed from the traditional MPE ($\bar{x}$ = .01) and HBM estimated values ($\bar{x}$ = -.01). Thus, the results of this analysis do not support our hypothesis that the increased test-retest reliabilities observed in the HBM parameters would also translate into higher between-task correlations.

Table 4

Between-Task Correlations of Reaction Time Difference Score Parameters.

Session	Index 1	Index 2	MPE	HBM	*n*
Baseline	Stroop Effect	BX Interference	.12	.10	90
Baseline		TRCE	-.10	.02	90
Baseline		Recency Effect	.16	-.02	90
Baseline	BX Interference	TRCE	.07	.03	90
Baseline		Recency Effect	.05	-.13	90
Baseline	TRCE	Recency Effect	-.04	.00	90
Proactive	Stroop Effect	BX Interference	-.03	.01	76
Proactive		TRCE	.20	.01	76
Proactive		Recency Effect	.01	-.07	76
Proactive	BX Interference	TRCE	-.10	-.13	76
Proactive		Recency Effect	.07	-.04	76
Proactive	TRCE	Recency Effect	-.16	-.13	76
Reactive	Stroop Effect	BX Interference	.11	.08	107
Reactive		TRCE	-.13	-.09	107
Reactive		Recency Effect	-.07	-.05	107
Reactive	BX Interference	TRCE	-.08	-.02	107
Reactive		Recency Effect	.15	.22	107
Reactive	TRCE	Recency Effect	-.09	-.04	107

Note. Indices are based on averaged test and retest phases. MPE = Pearson r correlation of Mean Point Estimated differences scores; HBM = Pearson r correlation of Hierarchical Bayesian Modeling estimated differences scores; TRCE = Task Rule Congruency Effect. Variability in sample sizes due to between-task differences in acceptable performance.

Discussion

The goal of the current study was to examine psychometric reliability in experimental tasks of cognitive control. To this end, we utilized the new DMCC task-battery, as it includes classic cognitive control tasks, but with theoretically-derived task variants and experimental manipulations based on the dual mechanisms of control framework (Braver, 2012; Braver et al., 2021). Our primary hypothesis was that the theoretically-based development and optimization of the tasks would create new sources of between-subject variance to improve individual differences. However, when using traditional statistical approaches (i.e., split-half, ICC), psychometric analyses suggested that our theoretically-optimized task battery did not improve reliability above and beyond that of existing tasks and batteries. Plainly stated, the reliability of the DMCC task battery measures, when computed with popular difference score indices, were moderate at best, which is quite consistent with prior psychometric reports using different task variants (von Bastian et al., 2020 see also; Hedge, Powell, & Sumner, 2018; Rouder & Haaf, 2019). In particular, when analyses were conducted with traditional psychometric methods, there was no evidence in support of our primary hypothesis that the theoretically-motivated task manipulations would improve reliability in cognitive control tasks.

One important finding was that the reliability estimates focused on internal consistency (i.e., split-half indices) were almost always higher than those focused on temporal stability (i.e., test-retest; i.e., ICC2,1 & ICC3,1). Given that split-half methods are calculated on a single timepoint measure, and test-retest on two (or more) timepoint measures, this finding is not surprising. But it does reaffirm that the two methods cannot be treated as interchangeable indices of reliability. When possible, an index of both internal consistency and temporal stability should be reported. Importantly, the observed discrepancy indicates that our measures of cognitive control have some internal consistency, but additional work needs to be conducted to determine why temporal stability appears to be lower than desirable. In our case, the "additional work" meant that we investigated whether traditional psychometric statistics might not be appropriate or well-aligned for the calculation of individual differences in experimental cognitive control tasks.

Towards this end, we conducted a second set of analyses that replicated recent studies in utilizing hierarchical Bayesian modeling (HBM) as an alternative approach that might be better suited for reliability estimation with cognitive experimental tasks (Chen et al., 2021; Haines et al., 2020; Rouder & Haaf, 2019). In particular, we found that with HBM methods highly reliable cognitive control indices can be obtained, even when using indices derived from difference scores. Specifically, our findings indicate that test-retest reliability estimates for the delta (difference score) parameters in our sample can be almost always classified as good, and sometimes even excellent. This finding is a striking one, in comparison with the weak and moderate intraclass correlation coefficients (ICC) observed in the traditional set of analyses. The HBM analyses clearly suggest that accounting for individual-level variability and the type and shape of the distribution, can "rescue" the reliability estimation, using the formulation of Rouder and Haaf (2019). Interestingly, it was also found that in both the traditional and HBM analyses, reliability estimates were highest for the proactive task variants, which also supports our hypothesis that theoretical motivated task manipulations may contribute to improved reliability.

One of the primary reasons for the enduring importance and need for attention to reliability measures is the view – which is well-accepted in the psychometric literature (Hedge, Powell, & Sumner, 2018; Parsons et al., 2019; Rouder et al., 2019; Spearman, 1904) – that reliability might serve as a bottle-neck or constraint on the ability to detect correlations between measures of individual differences. The key point is that, for measures with low reliability, there should be reduced sensitivity for the detection of between-measure correlations. Yet this assumption has been rarely experimentally tested (Cooper et al., 2017). In our analyses, we did in fact, find support for this contention, when examining correlations between DMCC task measures within tasks (i.e., between control modes; baseline, proactive, reactive). Specifically, we assumed that within-task correlations could be treated as “benchmarks” since we assumed ground-truth positive correlations, given that the same participants were performing subtle variants of the same task across sessions. Indeed, we found that not only was test-retest reliability increased with HBM estimates relative to the traditional ICC measures, but also so were the within-task correlations. In fact, the data also provided support for the assumption that the degree of improvement in within-task correlation was directly tied to the increase in reliability observed with HBM estimates. Thus, the results provide clear support for the psychometric perspective, in demonstrating the importance of reliability, as well as the improved potential to estimate individual differences in cognitive control with HBM-based approaches.

Unfortunately, the one dimension in which the increased reliability obtained with HBM-estimates *did not* translate into improved correlation strength, was in the correlations observed between DMCC tasks. Here, we observed on average quite low correlations (less than 0.2) that did not differ between measures derived from traditional difference scores and those from HBM-based estimates. Thus, at least in the case of the DMCC task battery, it cannot be claimed that the weak between-task correlations are due to the unreliability of the measures. Indeed, the contrast among the within-task and between-task correlations is striking. Moreover, it clearly points to the need for future research to understand the basis for the repeated findings of low between-task correlations among cognitive control measures (von Bastian et al., 2020), particularly given that our results argue against an interpretation in terms of low measurement reliability. We discuss this issue further below, along with other limitations of the current work and fruitful directions for further research.

Limitations and Future Directions

The current study design, though promising for validation of the dual mechanisms of control theoretical framework and newly developed DMCC task battery, does come with some limitations. First, it is important to acknowledge the fully online format of the design. This design has clear and significant advantages, the foremost of which is that the multi-session nature of the study would place a stronger burden on participants if frequent laboratory visits were required. Moreover, at the time of this writing, the worldwide pandemic has accelerated this shift of experimental research towards an online format. Finally, much work has validated online task administration as a viable format for cognitive tasks, with many important results replicated (Anwyl-Irvine et al., 2021; Bridges et al., 2020; Chaytor et al., 2021; Crump et al., 2013; Pronk et al., 2021). Nevertheless, the online format also has a number of drawbacks, which are well-known in the literature. These include reduced experimental control over the task environment, and an increased risk of potential distractions being present.

Another limitation of the design comes from the fact that not all of the tasks are optimized to be delivered in a test-retest format. In addition to standard concerns about practice effects impacting retest sessions, the DMCC battery also includes some tasks and conditions that are likely to be more impacted by prior experience than others. For example, in the Cued-TS proactive and reactive conditions, incentives are given based on performance, although these are not present in the baseline condition. During the initial baseline condition, participants are not told about the potential for incentives in the subsequent proactive and reactive sessions.

However, during the retest baseline session, they do have this knowledge, which could impact the cognitive strategies used in this session. Likewise, in the AX-CPT proactive condition, participants receive explicit strategy training for how to utilize the contextual cues. Again, in the preceding baseline test session, which is otherwise identical to proactive, they have not yet received this strategy training, but in the retest baseline sessions participants have already had much experience in following the strategy instructions, which could also impact their performance in this session. Thus, in future investigations of test-retest reliability with the DMCC battery, it would be useful to reconsider the manipulations used for the proactive and reactive sessions, to minimize the carry-over effects of prior practice.

The current study adds to a growing literature highlighting the promise and potential of HBM approaches for analyzing cognitive experimental tasks. Yet, currently these types of Bayesian analyses are still relatively rare in the literature; consequently, there is still a poor understanding of how they are different from traditional analyses, or how effects might diverge. Given the lack of widespread adoption of HBM methods, we opted for a more conservative approach, of first presenting results from traditional psychometric analyses of reliability, before comparing them with HBM estimates. We utilized Bayesian models that estimated effects for each task-variant separately, following current literature (Haines et al., 2020; Rouder & Haaf, 2019). However, the approach could be expanded to a single all-encompassing model. In particular, it should be possible to develop a generative model in which the different conditions and even different tasks are assumed to be additional level(s) of hierarchy from which the distributions arise (i.e., analogous to the way participants are drawn from a higher-level distribution). Our current model benefits from shared information across subjects and trial-types (i.e., congruent, incongruent), but only within one variant (i.e., baseline, proactive, reactive) of each task-paradigm. A complete generative model has the benefit of between-condition and between-task information sharing as well. However, building full generative models will increase the complexity of the modeling endeavor, so it is worthwhile to progress in a more incremental fashion. Nevertheless, the promise of the current approach suggests that further development of Bayesian statistical approaches to task parameter estimation may be a particularly worthwhile direction for the field (Gelman et al., 2013; Lee & Wagenmakers, 2014; McElreath, 2020).

As part of the limitations of the current study, we acknowledge recent work suggesting that analyses solely based on reaction time measures also pose a challenge in interpreting results. For example, Draheim and colleagues (2019, 2020) have argued that the use of reaction time difference scores "is the primary cause of null and conflicting results" when examining individual differences in attentional control. Their work suggests that measures based on accuracy rather than reaction time can improve reliability, intercorrelations among tasks, latent factor scores, and associations with measures of working memory and fluid intelligence. Although it was beyond the scope of the current study, it is of course possible to use HBM approaches with accuracy measures as well, which suggests another possible direction for future work (Lin et al., 2022). Other work by Hedge and colleagues (2021) suggests the importance of cognitive modeling to properly estimate latent processes (diffusion model for conflict tasks; Ulrich et al., 2015). In this work it was found that, when conflict processes were decomposed from non-conflict processes, only weak correlations ($r < .05$) were observed between conflict process across different cognitive control tasks. Contrarily, correlations between model parameters representing processing speed and strategy were consistently positive, with moderate to strong correlations. As a combined resolution, future work would do well to use cognitive models that account for the speed-accuracy tradeoff and the multiple latent processes that underlie observed measures.

The key unresolved question from the current study relates to the low between-task correlations observed, even among the theoretically-derived tasks that comprise the DMCC battery. These findings are not unprecedented; indeed, they are quite consistent with a number of prior studies that have examined correlations among cognitive control measures through task batteries and latent variable modeling (Draheim et al., 2020; Rey-Mermet et al., 2018; von Bastian et al., 2020). Nevertheless, the current results are quite discouraging, as they increase doubt on the domain-generality of cognitive control constructs. In some ways, however, the results are also discrepant from work that has been emerging from the neuroimaging literature, which has also become more attuned to questions of individual differences and domain-generality (Dubois & Adolphs, 2016; Elliott et al., 2020; Finn et al., 2017; Freund et al., 2021; Gratton et al., 2018).

Indeed, within the neuroimaging literature, an important emerging finding is that although lower-dimensional (e.g., “univariate”) descriptions may not be reliable for characterizing individual differences in brain activity, higher-dimensional (e.g., multivariate) descriptions can be quite discriminative. This can be seen most clearly in “fingerprinting” studies (Finn et al., 2015), in which pattern similarity techniques demonstrate that individuals show high test-retest reliability, such that their activation profile from a test scan can be easily discriminated from other individuals in a retest session (i.e., significantly higher test-retest similarity within-individuals than between). Moreover, our group has extended this approach into the domain of task fMRI and cognitive control, using twin-based study designs to demonstrate a remarkable degree of similarity among identical twin-pairs relative to unrelated pairs (or even fraternal pairs) in the fronto-parietal regions most strongly associated with cognitive control functions (Tang, Etzel, et al., 2021). Most strikingly, these effects were only observed when utilizing multivariate activation pattern similarity, rather than univariate measures (Etzel et al., 2020), and demonstrated clear domain-generality (i.e., cross-task effects; Tang, Etzel, et al., 2021). Together, this work suggests the possibility that utilizing multivariate rather than univariate descriptions of the individual might be a promising direction even for behavioral characterizations. Indeed, initial work in this direction, utilizing behavioral fingerprinting approaches, has begun (Han & Adolphs, 2020), though much more investigation is needed.

Conclusion

We examined whether well-established experimental tasks, but modified with theoretically-aligned variants and task manipulations, are viable tools for measuring individual differences in cognitive control. As previously reported (Tang, Bugg, et al., 2021), the experimental manipulations included in this task battery were validated to be highly robust at the group level, in inducing consistent shifts towards proactive and reactive control. Yet, with the use of traditional psychometric approaches, the experimental manipulations did not produce clear effects on either internal consistency (split-half) or temporal stability (test-retest) measures of reliability, which were observed to be moderate at best. In contrast, when the test-retest data were re-examined using hierarchical Bayesian modeling, the findings were quite different, with good to excellent reliability observed in most measures. Moreover, these reliability effects translated into improved strength of within-task correlations. Nevertheless, even with the Bayesian estimates, cross-task correlations were unaffected and remained uniformly poor. Together, these findings add to the growing literature suggesting the importance of Bayesian generative models when estimating individual differences, but also point to the need for further investigation into the source of low cross-task correlations among experimental tasks that attempt to measure putatively domain-general cognitive control constructs. Most generally, we encourage other researchers interested in cognitive individual differences to attend more closely to psychometric issues when conducting this important research.

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

Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2021). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behavior Research Methods*, *53*(4), 1407–1425. https://doi.org/10.3758/s13428-020-01501-5

Boudewyn, M. A., Long, D. L., Traxler, M. J., Lesh, T. A., Dave, S., Mangun, G. R., Carter, C. S., & Swaab, T. Y. (2015). Sensitivity to Referential Ambiguity in Discourse: The Role of Attention, Working Memory, and Verbal Ability. *Journal of Cognitive Neuroscience*, *27*(12), 2309–2323. https://doi.org/10.1162/jocn_a_00837

Braem, S., Bugg, J. M., Schmidt, J. R., Crump, M. J. C., Weissman, D. H., Notebaert, W., & Egner, T. (2019). Measuring Adaptive Control in Conflict Tasks. *Trends in Cognitive Sciences*, *23*(9), 769–783. https://doi.org/10.1016/j.tics.2019.07.002

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., 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., Kizhner, A., Tang, R., Freund, M. C., & Etzel, J. A. (2021). The Dual Mechanisms of Cognitive Control Project. *Journal of Cognitive Neuroscience*, *33*(9), 1990–2015. https://doi.org/10.1162/jocn_a_01768

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

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ*, *8*, e9414. https://doi.org/10.7717/peerj.9414

Bugg, J. M. (2012). Dissociating Levels of Cognitive Control: The Case of Stroop Interference. *Current Directions in Psychological Science*, *21*(5), 302–309. https://doi.org/10.1177/0963721412453586

Chapman, L. J., & Chapman, J. P. (1978). The measurement of differential deficit. *Journal of Psychiatric Research*, *14*(1–4), 303–311. https://doi.org/10.1016/0022-3956(78)90034-1

Chaytor, N. S., Barbosa-Leiker, C., Germine, L. T., Fonseca, L. M., McPherson, S. M., & Tuttle, K. R. (2021). Construct validity, ecological validity and acceptance of self-administered online neuropsychological assessment in adults. *The Clinical Neuropsychologist*, *35*(1), 148–164. https://doi.org/10.1080/13854046.2020.1811893

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

Cohen, J. D., Barch, D. M., Carter, C., & Servan-Schreiber, D. (1999). Context-processing deficits in schizophrenia: Converging evidence from three theoretically motivated cognitive tasks. *Journal of Abnormal Psychology*, *108*(1), 120–133. https://doi.org/10.1037/0021-843X.108.1.120

Cooper, S. R., Gonthier, C., Barch, D. M., & Braver, T. S. (2017). The Role of Psychometrics in Individual Differences Research in Cognition: A Case Study of the AX-CPT. *Frontiers in Psychology*, *8*, 1482. https://doi.org/10.3389/fpsyg.2017.01482

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

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon’s Mechanical Turk as a Tool for Experimental Behavioral Research. *PLOS ONE*, *8*(3), e57410. https://doi.org/10.1371/journal.pone.0057410

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

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

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

Dubois, J., & Adolphs, R. (2016). Building a Science of Individual Differences from fMRI. *Trends in Cognitive Sciences*, *20*(6), 425–443. https://doi.org/10.1016/j.tics.2016.03.014

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., Sison, M. L., Moffitt, T. E., Caspi, A., & Hariri, A. R. (2020). What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. *Psychological Science*, *31*(7), 792–806. https://doi.org/10.1177/0956797620916786

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.

Enock, P. M., Robinaugh, D. J., Reese, H. E., & McNally, R. J. (2012, November). *Improved reliability estimation and psychometrics of the dot-probe paradigm on smartphones and PC.* Annual Meeting of the Association of Behavioral and Cognitive Therapies, National Harbor, MD.

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

Etzel, J. A., Courtney, Y., Carey, C. E., Gehred, M. Z., Agrawal, A., & Braver, T. S. (2020). Pattern Similarity Analyses of FrontoParietal Task Coding: Individual Variation and Genetic Influences. *Cerebral Cortex*, *30*(5), 3167–3183. https://doi.org/10.1093/cercor/bhz301

Finn, E. S., Scheinost, D., Finn, D. M., Shen, X., Papademetris, X., & Constable, R. T. (2017). Can brain state be manipulated to emphasize individual differences in functional connectivity? *NeuroImage*, *160*, 140–151. https://doi.org/10.1016/j.neuroimage.2017.03.064

Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., Papademetris, X., & Constable, R. T. (2015). Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity. *Nature Neuroscience*, *18*(11), 1664–1671. https://doi.org/10.1038/nn.4135

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct Validation in Social and Personality Research: Current Practice and Recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Freund, M. C., Etzel, J. A., & Braver, T. S. (2021). Neural Coding of Cognitive Control: The Representational Similarity Analysis Approach. *Trends in Cognitive Sciences*, *25*(7), 622–638. https://doi.org/10.1016/j.tics.2021.03.011

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

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

Gärtner, A., & Strobel, A. (2019). *Individual differences in inhibitory control: A latent variable analysis* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/gnhmt

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

Gratton, C., Laumann, T. O., Nielsen, A. N., Greene, D. J., Gordon, E. M., Gilmore, A. W., Nelson, S. M., Coalson, R. S., Snyder, A. Z., Schlaggar, B. L., Dosenbach, N. U. F., & Petersen, S. E. (2018). Functional Brain Networks Are Dominated by Stable Group and Individual Factors, Not Cognitive or Daily Variation. *Neuron*, *98*(2), 439-452.e5. https://doi.org/10.1016/j.neuron.2018.03.035

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

Han, Y., & Adolphs, R. (2020). Estimating the heritability of psychological measures in the Human Connectome Project dataset. *PLOS ONE*, *15*(7), e0235860. https://doi.org/10.1371/journal.pone.0235860

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 ofd′. *Behavior Research Methods, Instruments, & Computers*, *27*(1), 46–51. https://doi.org/10.3758/BF03203619

Hedge, C., Powell, G., Bompas, A., & Sumner, P. (2021). Strategy and processing speed eclipse individual differences in control ability in conflict tasks. *Journal of Experimental Psychology. Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001028

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

Hussey, I., & Hughes, S. (2020). Hidden Invalidity Among 15 Commonly Used Measures in Social and Personality Psychology. *Advances in Methods and Practices in Psychological Science*, *3*(2), 166–184. https://doi.org/10.1177/2515245919882903

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., & Koch, I. (2010). Control and interference in task switching—A review. *Psychological Bulletin*, *136*(5), 849–874. https://doi.org/10.1037/a0019842

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

Kucina, T., Wells, L., Lewis, I., Salas, K. de, Kohl, A., Palmer, M., Sauer, J. D., Matzke, D., Aidman, E., & Heathcote, A. (2022). *A Solution to The Reliability Paradox for Decision-Conflict Tasks*. PsyArXiv. https://doi.org/10.31234/osf.io/bc6nk

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

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press.

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

Lin, Y., Brough, R., Tay, A., Jackson, J. J., & Braver, T. (2022). *Working memory capacity preferentially enhances implementation of proactive control*. PsyArXiv. https://doi.org/10.31234/osf.io/wvpbn

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

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429029608

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

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

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*(1), 1–18. https://doi.org/10.1016/0022-2496(66)90002-2

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., & 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

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

Posner, M., & Snyder, C. (1975). *Facilitation and inhibition in the processing of signals*. ResearchGate. https://www.researchgate.net/publication/243666218_Facilitation_and_inhibition_in_the_processing_of_signals

Pronk, T., Hirst, R., Wiers, R., & Murre, J. (2021). *Can we Measure Individual Differences in Cognitive Measures Reliably via Smartphones? A Comparison of the Flanker Effect Across Device Types and Samples*. PsyArXiv. https://doi.org/10.31234/osf.io/2kdca

Ramirez, G., Gunderson, E. A., Levine, S. C., & Beilock, S. L. (2013). Math Anxiety, Working Memory, and Math Achievement in Early Elementary School. *Journal of Cognition and Development*, *14*(2), 187–202. https://doi.org/10.1080/15248372.2012.664593

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

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

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., Kumar, A., & Haaf, J. M. (2019). *Why Most Studies of Individual Differences With Inhibition Tasks Are Bound To Fail*. PsyArXiv. https://doi.org/10.31234/osf.io/3cjr5

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

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

Snijders, T. A. B., & Bosker, R. J. (2011). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. SAGE.

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

Spearman, C. (1910). Correlation Calculated from Faulty Data. *British Journal of Psychology, 1904-1920*, *3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

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

Stan Development Team. (2020b). *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

Stawarczyk, D., Majerus, S., Catale, C., & D’Argembeau, A. (2014). Relationships between mind-wandering and attentional control abilities in young adults and adolescents. *Acta Psychologica*, *148*, 25–36. https://doi.org/10.1016/j.actpsy.2014.01.007

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

Tang, R., Bugg, J., Snijder, J.-P., Conway, A. R. A., & Braver, T. (2021). *The Dual Mechanisms of Cognitive Control (DMCC) project: Validation of an online behavioral task battery*. PsyArXiv. https://doi.org/10.31234/osf.io/ngwqc

Tang, R., Etzel, J. A., Kizhner, A., & Braver, T. S. (2021). Frontoparietal pattern similarity analyses of cognitive control in monozygotic twins. *NeuroImage*, *241*, 118415. https://doi.org/10.1016/j.neuroimage.2021.118415

Tucker-Drob, E. M. (2011). Individual Differences Methods for Randomized Experiments. *Psychological Methods*, *16*(3), 298–318. https://doi.org/10.1037/a0023349

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

Verbruggen, F., & Logan, G. D. (2009). Models of response inhibition in the stop-signal and stop-change paradigms. *Neuroscience & Biobehavioral Reviews*, *33*(5), 647–661. https://doi.org/10.1016/j.neubiorev.2008.08.014

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

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

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

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

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 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.

Appendix B

Table B1

Between-Task Spearman Rho Correlations of 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

Between-Task Spearman Rho Correlations of 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

Between-Task Spearman Rho Correlations of 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

Appendix C

Overview of DMC Task Battery Paradigms

Task Paradigms

Here we present the most pertinent information regarding the tasks and their manipulations. For a complete description of the tasks (e.g., manipulation rationales, stimulus parameters (e.g., ISI, etc.)), see Tang et al. (2021).

Stroop

Baseline Sessions. In a baseline session the trials were manipulated in a list-wide, mostly congruent (LW-MC) manner. Subjects completed a total of 288 trials during a 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 sessions were 66%. The sessions were divided into two blocks of 144 trials each, between which subjects were instructed to rest for one minute.

Reactive Sessions. In the reactive sessions 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 sessions, 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 sessions. Each reactive session was divided into three blocks of 160 trials each, between which subjects were instructed to rest for one minute.

Proactive Sessions. In the proactive sessions, the trials were manipulated in a list-wide, mostly incongruent (LW-MI) manner. Subjects completed a total of 288 trials during the proactive sessions, 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 sessions were 33%. A proactive session was divided into two blocks of 144 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 the biased set, 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. For brevity, the results of the PC-50 item set are not reported.

AX-CPT

Baseline Sessions. For all AX-CPT sessions, 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. After receiving task instructions, subjects performed a 12-trial practice block before beginning the actual task.

Reactive Sessions. 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 variants) 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 variants.

Proactive Sessions. In the proactive sessions, 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 online testing format, and deviated slightly from in-person versions, in which subjects responded verbally regarding the button they were preparing to press.

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

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

Cued Task Switching

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), also referred to as the biased set. The second set of stimuli (D4, E3, H5, I6, 4D, 3E, 5H, I6) were unbiased (50% congruent, 50% incongruent). 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 the task rules through a set of 16 practice trials.

Baseline Sessions. For the baseline session, 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 to the subjects.

Reactive Sessions. The reactive sessions of Cued-TS were 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, which were indicated by a green cue and target, they received a monetary penalty for that trial that was subtracted from their compensation amount.

Proactive Sessions. The proactive sessions of Cued-TS were identical to the baseline sessions except for the addition of a reward-based motivational incentive. This motivational incentive provides subjects with a reward cue, indicated by a cue in green font-color during presentation of the task cue. Non-incentive trials indicated by the task cue appearing in red font. 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.

Cognitive Control Measures. Average reaction times (RTs) on correct trials and error rates were calculated separately for congruent/incongruent biased items, for each subject in each session. Additionally, the TRCE (Task Rule Congruency Effect) was calculated as a difference score between incongruent and congruent trials and was computed for biased items. For brevity, the results of the unbiased set are not reported.

Sternberg

Baseline Sessions. The baseline sessions 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).

Reactive Sessions. In the reactive sessions, 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).

Proactive Sessions. In the proactive sessions, 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).

Cognitive Control Measures. Average reaction times (RTs) on correct trials and error rates were calculated per trial type (i.e., NN, NP, RN trials) for critical items (list-length 5). 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. For brevity, the results of the variable-load item set are not reported here.