Psychometrics of hierarchical drift diffusion modeling for Eriksen flanker task: Reliability and validity in two samples

Brent I. Rappaport1, Allison Letkiewicz1, Anna Weinberg2, Savannah Buchanan1, & Stewart Shankman1

1 Department of Psychiatry, Feinberg School of Medicine, Northwestern University

2 Department of Psychology, McGill University

Author note

The authors made the following contributions. Brent I. Rappaport: Conceptualization, Formal Analysis, Methodology, Writing - Original Draft Preparation, Writing - Review & Editing; Allison Letkiewicz: Conceptualization, Formal Analysis, Methodology, Writing - Original Draft Preparation, Writing - Review & Editing; Anna Weinberg: Data Curation, Investigation, Project Administration, Writing - Review & Editing; Savannah Buchanan: Writing - Original Draft Preparation, Writing - Review & Editing; Stewart Shankman: Conceptualization, Funding Acquisition, Methodology, Project Administration, Resources, Supervision, Writing - Original Draft Preparation, Writing - Review & Editing.

Correspondence concerning this article should be addressed to Brent I. Rappaport, 680 N. Lakeshore Drive, Suite 1520, Chicago, IL 60611. E-mail: [brent.rappaport@northwestern.edu](mailto:brent.rappaport@northwestern.edu)

Abstract

Despite advantages of computational models to study cognitive functions that underlie psychiatric disorders, little research has compared the psychometric properties of these models to simpler behavioral measures. We examined the reliability and validity of three behavioral measures of a cognitive control task: the Eriksen flanker, specifically 1) raw accuracy, 2) a NIH Toolbox derived score that incorporates reaction time and accuracy, and 3) parameters of a hierarchical drift diffusion model (HDDM). Participants from two independent studies—one cross-sectional sample (N=392) and one longitudinal sample with three time points (N=79, 70, 68, respectively)—completed the Flanker task while electroencephalography data was collected. Behavioral data were processed using with HDDM yielding five parameters: drift rate to congruent/incongruent stimuli, separation between decision boundaries, non-decision time, and starting bias. In the two studies, drift rate, particularly for congruent stimuli, demonstrated better split-half and test-retest reliability than the NIH Toolbox score. Drift rate also demonstrated better convergent validity with brain measures (i.e., the error-related negativity event-related potential component) and neuropsychological measures of inhibition and executive function. Results showed that faster accumulation of evidence (drift rate) was related to 1) larger ERN amplitudes and 2) faster and more accurate inhibition and shifting over and above the NIH Toolbox and raw accuracy scores and covariates (IQ, motor speed). Given enough trials, these models can be fit to extant data. Thus, findings suggest that these models may be powerful tools in identifying aberrations in cognitive functioning associated with psychiatric disorders.

*Keywords:* psychometric, computational modeling, EEG, cognitive control, executive function

Executive functioning, and cognitive control in particular, is a richly studied area of cognitive psychology and neuroscience. Broadly, executive functioning includes several related cognitive processes that contribute to effortfully guide behavior toward goals (Banich, 2009). One such process is cognitive control. Understanding cognitive control, that is how humans are able to *“marshal…functions including working, semantic, and episodic memory; perceptual attention; and action selection and inhibition”* (M. Botvinick & Braver, 2015), is key to understanding human behavior. Alterations in cognitive control have far reaching consequences, for instance impacting healthy aging (Braver & Barch, 2002) and mental health (McTeague et al., 2016; Smucny et al., 2019). Studies often rely on a few notable tasks (e.g., Stroop, Go/No-go, and Eriksen Flanker) to explore how cognitive control arises and what it can tell us about human cognition and behavior.

One commonly used task to assess cognitive control is the Eriksen flanker task (Eriksen & Eriksen, 1974). This task involves asking participants to identify the central symbol in a string of symbols which are either congruent (i.e., the same) or incongruent (i.e., different) with the central symbol (for further history and details see Ridderinkhof et al. (2021)). The task has been implemented with a variety of symbols, though arrows have been most widely used (e.g., congruent: < < < < < ; incongruent: < < > < < ). Participants are often instructed to balance responding correctly with responding quickly. This leads to errors in incongruent trials and response slowing due to response competition. Participants’ ability therefore to respond both quickly and correctly is thought to quantify meaningful information about their cognitive control abilities, lending itself to individual differences research. Its robustness along with its simplicity make it an ideal tool for researchers to adapt to answer specific research questions.

Historically, researchers have used accuracy and reaction time as their primary measures of interest (e.g., Hsieh & Fang, 2012; Huyser et al., 2011; Imburgio et al., 2020; Luks et al., 2010; McDermott et al., 2007; Moore et al., 2015; Scudder et al., 2014; Wylie et al., 2009), despite research showing poor psychometrics. For example, two studies found poor test-retest reliability for these measures (Meyer et al., 2014; Sanders et al., 2018) and one ommitted such analyses altogether (Wöstmann et al., 2013). Accuracy also shows poor internal consistency, in one case ranging from Cronbach’s alphas of -0.18 to 0.45 depending on the condition (Wöstmann et al., 2013). In terms of validity, it shows variable convergent validity, being significantly correlated with performance on Stroop (Keye et al., 2009) but not with performance on the mini-mental state examination in one study of patients with dementia (Sanders et al., 2018). Moreover, examining accuracy and reaction time separately limits one’s ability to infer meaningful individual differences, such as participants’ ability to maintain high accuracy while increasing their reaction time.

In response to such criticism, researchers developed a score that combines reaction time and accuracy (Zelazo et al., 2014) and is included in the NIH Toolbox (hereafter referred to as the NIH Toolbox score). This score ought to improve psychometrically upon accuracy, given its inclusion of reaction time which appears to be more reliable than accuracy (Meyer et al., 2014). Despite validation for this calculation (Zelazo et al., 2014, 2013), it still has a number of inherent limitations. First, it only includes reaction time on *incongruent* trials, thus excluding information about participants differential reaction time to incongruent relative to congruent stimuli. Second, like accuracy, it conflates many underlying psychological processes. For example, participants may slow down in responding due to a greater emphasis on accuracy, greater difficulty accumulating evidence for a response, or simply distraction. Measures that rely only on accuracy and/or reaction time conflate these causes. Third, by calculating a summary statistic (e.g., mean), information about the trial-to-trial variability—crucial for improving reliability (Chen et al., 2021) and convergence with brain measures (Ratcliff et al., 2009; Wiecki et al., 2013)—is lost.

In responses to these limitations, a novel way to assess cognition in this task is drift diffusion modeling [DDM; Ratcliff et al. (2016); Ratcliff and Smith (2004); Ratcliff and McKoon (2008)], or hierarchical DDM (HDDM; a variety that uses hierarchical Bayesian parameter estimation) (Gutchess et al., 2021; Ulrichsen et al., 2020; Wiecki et al., 2013). It offers a clear advantage–the modeling of multiple different cognitive processes. This is done by using trial-by-trial accuracy and reaction times to differing stimuli (i.e., congruent and incongruent) to estimate parameters associated with and affected by specific changes in cognition, separate from changes in performance. This results in one value per parameter per participant. Importantly, each parameter controls for the effects of the other parameters included in the model (White et al., 2010a, 2010b). Thus, where other measures conflate the many cognitive processes that give rise to cognitive control, DDM differentiates them. This is important, for example, in cognitive studies interested in how motivation affects cognitive control (M. Botvinick & Braver, 2015), or clinical studies looking to identify specific, subtle processes aberrant in psychiatric disorders that may not lead to broad cognitive impairments (e.g., Moser et al., 2013). Indeed, DDM parameters tend to be more sensitive to differences in task performance (Pe et al., 2013; White et al., 2010b) as well as complimentary to accuracy and reaction time (Hall et al., 2021). In one case, DDM has already been used to improve the psychometric properties of other cognitive tasks (Price et al., 2019).

A validated method is to fit many different variations of DDM models, which include different parameters (e.g., drift rate, boundary separation, non-decision time, and bias (see (Ratcliff et al., 2016) for full discussion of DDM parameters)), to identify the best fitting model. The processes particularly relevant for cognitive control are evidence accumulation and speed-accuracy tradeoff (Bogacz et al., 2010). These processes can be estimated with two parameters from HDDM: drift rate and boundary separation, respectively. Drift rate is the average rate to approach a boundary (see Figure 1) and measures the quality/strength of evidence gleaned from the stimulus. A fast drift rate leads to fast and accurate responses. Boundary separation can be thought of as an indicator of speed-accuracy trade off or response caution. A large boundary separation means the two boundaries are farther apart, leading to more accurate albeit slower responses. A smaller boundary separation indicates closer boundaries and thus faster but less accurate responses. This is because with less boundary separation, there is more opportunity for noise in the drift to lead to reaching the incorrect boundary (an error).

Despite the promise of DDM, its psychometric properties have not been compared to more traditional measures like accuracy or the NIH Toolbox score. Therefore, one goal of the current study is to compare the reliability and validity of these measures. This paper focuses on multiple aspects of reliability and multiple aspects of validity. Specifically, we measured split-half and test-retest reliability, as well as convergent validity with brain and behavioral validators. As a brain validator, we used the error-related negativity (ERN), an event-related potential (ERP) component is typically assessed as a response-related component that responds to incorrect muscle responses, leading to an ERP that begins even slightly prior to the participant’s actual response, and is more negative for errors than correct responses (Falkenstein et al., 1991; Gehring et al., 1993). Further, studies have clarified the ERN as both a potential indicator of performance monitoring (Holroyd & Coles, 2002) and of cognitive control (Meyer & Hajcak, 2019), showing convergence with brain activity on other cognitive control tasks (Meyer et al., 2014). Importantly, the flanker task elicits an ERN that, though correlated with the ERN from similar tasks (Riesel et al., 2013), is more robust to the number of errors and has better internal consistency than other cognitive control tasks such as Go/No-go, Stoop, and picture/word tasks (Foti et al., 2013; Meyer et al., 2014, 2013). Clinical meta-analyses have purported the ERN as a potential transdiagnostic indicator, spanning anxiety (Cavanagh et al., 2017; Moser et al., 2013; Riesel, 2019) externalizing (Hall et al., 2021; Meyer & Hajcak, 2019), and psychotic disorder dimensions (Foti et al., 2013).

As a behavioral validator, we used neuropsychological measures of cognitive control specifically and executive function broadly based on performance across multiple tasks. These neuropsychological measures are lauded for their robust reliability and validity (Delis et al., 2004; Strong et al., 2011), making them ideal standards against which to test these measures. We further supplemented these analyses; specifically, for reliability, we aimed to identify the number of trials needed to achieve a stable parameter estimate. For validity, we leveraged sibling pairs in one study to compare the familiality of DDM parameters, accuracy, and the NIH Toolbox score.

Finally, responses to the recent replication crisis in psychology have emphasized the importance of replicating results in independent samples (Open Science Collaboration, 2015, 2012). Moreover, there have also been concerns about the test-retest reliability of tasks with strong condition-level effects (Hedge et al., 2018). Thus, a final goal was to replicate the findings in an independent sample. The use of a large cross-sectional study and a longitudinal study meant that we could test within- and between-person replication, an unique strength of the current study.

Thus in sum, we sought to examine the relative reliability and validity of four measures of behavioral performance on a well-studied behavioral task of cognitive control: the Eriksen flanker task (Eriksen & Eriksen, 1974). To do so, we compared the psychometric properties of two HDDM parameters, raw accuracy, and a NIH Toolbox derived score across two independent adult samples. One is longitudinal with three time points, allowing us to test-retest reliability. The other is a large cross-sectional study that also included neuropsychological measures and sibling pairs, allowing us to test convergent validity and familiality. Importantly, both studies assess split-half reliability and convergent validity with brain measures (i.e., the ERN).

# Methods

## Participants

### Study 1.

For Study 1, individuals were recruited from flyers posted around a university campus and screened over the phone to determine eligibility. Inclusion criteria were being 18–60-years-old and right-handed. Exclusion criteria were a history of major medical or neurological problems, or head trauma with loss of consciousness for greater than 15 minutes. Left-handed or ambidextrous individuals were also excluded. Participants were scheduled for 5 laboratory visits, ensuring that consecutive visits occurred 2–14 days apart (median=7). A total of 86 participants provided informed consent and completed the study, with a majority completing all five sessions (n = 74, 86%) and six participants completing only one session (7%). Within-subject sessions were excluded from analyses for poor accuracy on the flanker task (i.e., < 50%) or poor EEG data quality (i.e., fewer than 10 artifact-free trials per condition) or missing data. One participant was excluded across all sessions due to below 50% accuracy on the flanker task. See Table 1 for full demographic information on the final sample.

### Study 2.

For Study 2, individuals were recruited from mental health clinics and the local community. Inclusion criteria were being 18–30-years-old, having a biological sibling within the same age range able to participate, and being right-handed. Exclusion criteria were being left-handed, being unable to read or write in English, having a history of a head trauma with loss of consciousness, or having a first-degree family member with a history of manic, hypomanic, or psychotic symptoms (for full method details, see Gorka et al., 2016; Weinberg et al., 2015). Participants were oversampled for severe internalizing psychopathology using the Depression, Anxiety, and Stress Scale [DASS; Lovibond and Lovibond (1995)] during initial screening.

## Procedure

### Flanker task.

Across Study 1 and 2, an arrowhead version of the flanker task was administered using Presentation software (Neurobehavioral Systems, Berkeley, CA). On each trial, participants were presented with a row of five arrowheads for 200 ms and were asked to indicate the direction of the central arrowhead with the left or right mouse button as quickly and accurately as possible. Half of the trials were congruent, and half were incongruent with trial order randomized. Participants completed 11 blocks of 30 trials (330 trials total), with short breaks and performance-based feedback given in between blocks. At the end of each block, participants received one of three types of performance feedback: if accuracy was 75% or lower, the message “Please try to be more accurate” was displayed; if accuracy was above 90%, the message “Please try to respond faster” was displayed; if accuracy was between 75% and 90%, the message “You’re doing a great job” was displayed.

### EEG Data Collection.

#### Study 1.

Continuous EEG was recorded with a Neuroscan Synamp2 system (Compumedics, Charlotte, NC, USA) using six midline electrodes (Fz, FCz, Cz, CPz, Pz, & Poz). The electrooculogram (EOG) generated from eye movements and blinks was recorded using facial electrodes placed approximately 1 cm above and below the left eye and 1 cm to the right and left of the eyes. All electrode impedances were below 5 k, and data were recorded with a sampling rate of 1,000 Hz. Pre and postprocessing were conducted offline in MATLAB using EEGLAB (Delorme & Makeig, 2004) and ERPLAB (Lopez-Calderon & Luck, 2014). EEG data was imported to EEGLAB, heart rate channel was removed, and data was resampled to 500Hz. A bandpass filter from 0.1 to 30Hz was applied and data was rereferenced to the average of the mastoids. Stimulus and response-locked epochs were segmented from -500 (before response) to 1000ms (after response) and baseline corrected from -500 to -300ms. Eyeblink and ocular artifacts were corrected (Gratton et al., 1983) and artifact detection and rejection was conducted on all scalp electrodes. Specifically, the criteria applied were a voltage step of more than 50 μV between sample points, a voltage difference of 175 μV within a trial, or a minimum voltage difference of less than 0.50 μV within 100ms intervals. These intervals were rejected from individual channels in each trial. ERPs were computed as mean amplitude 0 to 80ms following a response at a frontocentral electrode (i.e., FCz) per prior research (e.g., Meyer et al., 2017, 2013; Moser et al., 2013; Riesel et al., 2013). ERPs were computed separately for correct and error responses and a residualized difference score was calculated to isolate activity to errors (Meyer et al., 2017), yielding the error-related negativity (ERN) component. ERPs had to include at least 10 error and 10 correct artifact free trials per session; thus within-participant sessions were excluded if there were too few artifact free error or correct trials (N=11 sessions across 7 participants).

#### Study 2.

All signal processing was conducted offline in MATLAB using customized scripts and EEGLAB. First, EEG data were referenced to the common average, downsampled to 500 Hz, and the DC offset was removed from each channel. A band-pass filter was then applied from 1 to 100 Hz. The 60 Hz line noise was removed using the cleanLineNoise function, which uses a sliding window to adaptively estimate and subtract the line noise component (Bigdely-Shamlo et al., 2015). Next, using the clean\_rawdata function, artifactual channels were removed, defined as those (a) containing more than 5 s of flat signal, (b) correlating less than .8 with surrounding channels, (c) containing high frequency noise to signal ratio greater than 4 SD (Kothe & Makeig, 2013). Artifact subspace reconstruction (ASR; Mullen et al., 2015) was then applied to correct for significant noise bursts, also implemented within clean\_rawdata. ASR is a principal-component-analysis-based (PCA-based) technique in which data within a 500 ms sliding window (window step=250 ms) are PCA-decomposed. Noisy components, defined as those with variance greater than 20 SD above that of the clean portions of the data, were removed and the data were reconstructed from the remaining components. Further, time windows were removed if more than 25% of the channels contained high-power artifacts, defined as greater than 7 SD above the clean power estimates in the channel. All artifactual channels were replaced by whole head spline interpolation, and data were again re-referenced to the common average so that the sum across all channels was zero. Lastly, independent component analysis (ICA) was implemented to retain brain-related components only. These components were defined as (a) having greater probability to be brain than artifacts according to an automatic IC classifier [ICLabel; Pion-Tonachini et al. (2019)], (b) having residual variance (i.e., the difference between IC’s scalp projection and the projection of fitted equivalent current dipole) less than 15%, and (c) having fitted dipole location within the brain. After preprocessing, response-locked epochs were segmented from -1500 to 1500 ms. All epochs were then low-pass filtered at 30 Hz and baseline corrected (-200 to 0 ms for stimulus-locked epochs and -500 to -300 ms for response-locked epochs). Response-locked ERN was defined as the average voltage from 0 to 80 ms at FCz.

### Neuropsychological measures from Study 2.

Performance on neuropsychological tasks was summarized into two scores used to assess convergent validity. First, cognitive control was estimated from participants’ time to complete the inhibition condition from the D-KEFS color-word interference task [D-KEFS; Delis et al. (2012a)]. Of note, this was reverse scored to make it more comparable to the executive function composite. Second, executive function was estimated using a mean composite of four measures from the Delis-Kaplan Executive Function System Design Fluency, Verbal Fluency, Trail Making, and Color-Word Interference tasks. Specifically, measures used were 1) total number of successful designs made during the category switching condition on Design Fluency, and 2) on Verbal Fluency, 3) completion time for the number-letter sequencing condition multiplied by 1 for reverse-scoring on Trail Making, and 4) time to complete the inhibition/switching condition multiplied by 1 for reverse-scoring on Color-Word Inference. Higher scores indicate “better” (i.e., faster and more accurate) performance. The Wechsler Test of Adult Reading [WTAR; Wechsler (2001)] estimated participants’ full-scale IQ (FSIQ) and was included as a covariate in Study 2 analyses. See Letkiewicz et al. (2021) for further details about how these scores are calculated. Of import, scores on the WTAR are highly correlated with FSIQ [r = 0.73; Strauss et al. (2006)]. Motor speed was assessed using the motor speed condition of the D-KEFS Trail Making Test and was also included as a covariate to account for the potential impact of slowed reaction time.

## Data analysis

### Hierarchical drift diffusion modeling.

**ALLIE’S PORTION GOES HERE**

Because prior work using these models has focused on drift rate and boundary separation (e.g., Aylward et al., 2020; Ossola et al., 2021; Ziegler et al., 2016), we present findings for only these parameters.

### NIH Toolbox.

NIH Toolbox scores were calculated according to instructions laid out in the manual (National Institutes of Health and Northwestern University, 2021) and in prior research (Weintraub et al., 2013; Zelazo et al., 2014). The accuracy score was calculated as , where 0.01515152 is 5 points divided by 330 trials. Reaction time (RT) was calculated as

where 250ms was the slowest RT and 1000 was the fastest across participants. The NIH Toolbox manual specifies that trials outside of 500ms-3000ms should be truncated (i.e., reaction times below 500ms set equal to 500ms). We opted not to truncate data due to concerns that we would lose meaningful variability in responses between 250-500ms. These equations produce 2-vectors: one accuracy vector and one reaction time vector, each of which is scored from 0-5. The scores are added together resulting in a score that range from 0-10, with higher scores indicating “better” (i.e., faster and more accurate responses). Of note, for participants with less than or equal to 80% accuracy, only the accuracy vector is used.

### Raw accuracy.

Raw accuracy was calculated as the proportion of correct responses to total responses. Accuracy on congruent trials showed a large amount of skew given participants’ high rates of correct responses. Accuracy on incongruent trials showed a more normal distribution; therefore, we used accuracy on only incongruent trials.

Individual-level parameters were imported into R version 4.2.2 (2022-10-31) where remaining analyses were conducted, primarily using the following packages: tidyverse , ggplot2, DescTools, lme4, lmerTest, broom, broom.mixed, psych, here, confintr, irr, MASS, knitr, and papaja.

### Outliers and missing data.

Participant data was excluded if they had below 50% accuracy. Data was excluded on a session-by-session basis (N=3 sessions of 1 participant). Outliers on ERN, inhibition, and executive functioning were windsorized, such that values greater than 3 standard deviations from the mean were replaced the maximum/minimum allowed value (the mean ± 3 standard deviations). Correlations, multiple regressions, and ICCs use listwise deletion of participants if they are missing any measure/time point.

### Reliability: test-retest and split-half.

Measure reliability was examined two ways: 1) split-half and 2) test-retest. First, split-half reliability was calculated by splitting the trials in half, calculating the spearman rho correlation between the two halves, and using the Spearman-Brown prediction formula () to calculate the reliability (Infantolino et al., 2018; Luking et al., 2017). Split-half reliability was also examined at increments of 15 trials (e.g., first 15 trials vs second 15 trials, first 30 trials vs next 30 trials, etc.) to assess the point at which this measure reaches stability. Cutoffs from Henson (2001) were used to identify split-half reliability at different levels (<0.70 = poor , 0.70–0.79 = acceptable, 0.80–0.89 = good, and >0.90 = excellent). Test-retest reliability was calculated as the intraclass correlation coefficient (ICC) between all three sessions from Study 1.

### Validity: Correlations, multiple regressions, and familiality.

Raw two-sided Spearman rho correlation coefficients were initially examined between the ERN magnitude or neuropsychological score and each measure of flanker behavioral performance. Multiple regression models were used to assess the relative variance accounted for by DDM parameters, NIH Toolbox score, and raw accuracy when estimated as simultaneous predictors of the ERN or neuropsychological scores. That is, each model included as predictors: one ddm parameter, the NIH Toolbox score, and raw accuracy. In Study 1, linear multiple regression models were used; in Study 2, linear mixed effect multiple regression models were used to account for within-family random intercept between sibling pairs. Familiality was estimated as the ICC between siblings from the same family in Study 2 (Constantino et al., 2010; Weinberger et al., 1981).

# References

Anokhin, A. P., Luciana, M., Banich, M., Barch, D., Bjork, J. M., Gonzalez, M. R., Gonzalez, R., Haist, F., Jacobus, J., Lisdahl, K., McGlade, E., McCandliss, B., Nagel, B., Nixon, S. J., Tapert, S., Kennedy, J. T., & Thompson, W. (2022). Age-related changes and longitudinal stability of individual differences in ABCD Neurocognition measures. *Developmental Cognitive Neuroscience*, *54*, 101078. <https://doi.org/10.1016/j.dcn.2022.101078>

Aylward, J., Hales, C., Robinson, E., & Robinson, O. J. (2020). Translating a rodent measure of negative bias into humans: The impact of induced anxiety and unmedicated mood and anxiety disorders. *Psychological Medicine*, *50*(2), 237–246. <https://doi.org/10.1017/S0033291718004117>

Banich, M. T. (2009). Executive Function: The Search for an Integrated Account. *Current Directions in Psychological Science*, *18*(2), 89–94. <https://doi.org/10.1111/j.1467-8721.2009.01615.x>

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K.-M., & Robbins, K. A. (2015). The PREP pipeline: Standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformatics*, *9*. <https://doi.org/10.3389/fninf.2015.00016>

Bogacz, R., Wagenmakers, E.-J., Forstmann, B. U., & Nieuwenhuis, S. (2010). The neural basis of the speedaccuracy tradeoff. *Trends in Neurosciences*, *33*(1), 10–16. <https://doi.org/10.1016/j.tins.2009.09.002>

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652. <https://doi.org/10.1037/0033-295X.108.3.624>

Botvinick, M., & Braver, T. (2015). Motivation and Cognitive Control: From Behavior to Neural Mechanism. *Annual Review of Psychology*, *66*(1), 83–113. <https://doi.org/10.1146/annurev-psych-010814-015044>

Braver, T. S., & Barch, D. M. (2002). A theory of cognitive control, aging cognition, and neuromodulation. *Neuroscience & Biobehavioral Reviews*, *26*(7), 809–817. <https://doi.org/10.1016/S0149-7634(02)00067-2>

Cavanagh, J. F., Meyer, A., & Hajcak, G. (2017). Error-Specific Cognitive Control Alterations in Generalized Anxiety Disorder. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *2*(5), 413–420. <https://doi.org/10.1016/j.bpsc.2017.01.004>

Chen, G., Pine, D. S., Brotman, M. A., Smith, A. R., Cox, R. W., & Haller, S. P. (2021). Trial and error: A hierarchical modeling approach to test-retest reliability. *NeuroImage*, *245*, 118647. <https://doi.org/10.1016/j.neuroimage.2021.118647>

Constantino, J. N., Majmudar, P., Bottini, A., Arvin, M., Virkud, Y., Simons, P., & Spitznagel, E. L. (2010). Infant head growth in male siblings of children with and without autism spectrum disorders. *Journal of Neurodevelopmental Disorders*, *2*(1), 39–46. <https://doi.org/10.1007/s11689-009-9036-5>

Delis, D. C., Kaplan, E., & Kramer, J. H. (2012b). *Delis-Kaplan Executive Function System*. American Psychological Association. <https://doi.org/10.1037/t15082-000>

Delis, D. C., Kaplan, E., & Kramer, J. H. (2012a). *Delis-Kaplan Executive Function System*. American Psychological Association. <https://doi.org/10.1037/t15082-000>

Delis, D. C., Kramer, J. H., Kaplan, E., & Holdnack, J. (2004). Reliability and validity of the Delis-Kaplan Executive Function System: An update. *Journal of the International Neuropsychological Society: JINS*, *10*(2), 301–303. <https://doi.org/10.1017/S1355617704102191>

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. <https://doi.org/10.1016/j.jneumeth.2003.10.009>

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>

Evans, T. C., Walukevich, K. A., Seager, I., & Britton, J. C. (2018). A psychometric comparison of anxiety-relevant attention measures. *Anxiety, Stress, & Coping*, *31*(5), 539–554. <https://doi.org/10.1080/10615806.2018.1489536>

Falkenstein, M., Hohnsbein, J., Hoormann, J., & Blanke, L. (1991). Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks. *Electroencephalography and Clinical Neurophysiology*, *78*(6), 447–455. <https://doi.org/10.1016/0013-4694(91)90062-9>

Foti, D., Kotov, R., & Hajcak, G. (2013). Psychometric considerations in using error-related brain activity as a biomarker in psychotic disorders. *Journal of Abnormal Psychology*, *122*(2), 520–531. <https://doi.org/10.1037/a0032618>

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., & Donchin, E. (1993). A Neural System for Error Detection and Compensation. *Psychological Science*, *4*(6), 385–390. <https://doi.org/10.1111/j.1467-9280.1993.tb00586.x>

Gorka, S. M., Nelson, B. D., Phan, K. L., & Shankman, S. A. (2016). Intolerance of uncertainty and insula activation during uncertain reward. *Cognitive, Affective, & Behavioral Neuroscience*, *16*(5), 929–939. <https://doi.org/10.3758/s13415-016-0443-2>

Gratton, G., Coles, M. G. H., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography & Clinical Neurophysiology*, *55*(4), 468–484. <https://doi.org/10.1016/0013-4694(83)90135-9>

Gutchess, A., Ksander, J., Millar, P. R., Uzundag, B. A., Sekuler, R., & Boduroglu, A. (2021). Cultural Differences in Performance on Eriksen’s Flanker Task. *Attention, Perception & Psychophysics*, *83*(2), 882–898. <https://doi.org/10.3758/s13414-020-02117-9>

Hall, N. T., Schreiber, A. M., Allen, T. A., & Hallquist, M. N. (2021). Disentangling cognitive processes in externalizing psychopathology using drift diffusion modeling: Antagonism, but not disinhibition, is associated with poor cognitive control. *Journal of Personality*, *89*(5), 970–985. <https://doi.org/10.1111/jopy.12628>

Hauser, T. U., Skvortsova, V., De Choudhury, M., & Koutsouleris, N. (2022). The promise of a model-based psychiatry: Building computational models of mental ill health. *The Lancet Digital Health*, *4*(11), e816–e828. <https://doi.org/10.1016/S2589-7500(22)00152-2>

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>

Hedge, C., Vivian-Griffiths, S., Powell, G., Bompas, A., & Sumner, P. (2019). Slow and steady? Strategic adjustments in response caution are moderately reliable and correlate across tasks. *Consciousness and Cognition*, *75*, 102797. <https://doi.org/10.1016/j.concog.2019.102797>

Henson, R. K. (2001). Understanding Internal Consistency Reliability Estimates: A Conceptual Primer on Coefficient Alpha. *Measurement and Evaluation in Counseling and Development*, *34*(3), 177–189. <https://doi.org/10.1080/07481756.2002.12069034>

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*(4), 679–709. <https://doi.org/10.1037/0033-295X.109.4.679>

Hsieh, S., & Fang, W. (2012). Elderly adults through compensatory responses can be just as capable as young adults in inhibiting the flanker influence. *Biological Psychology*, *90*(2), 113–126. <https://doi.org/10.1016/j.biopsycho.2012.03.006>

Huyser, C., Veltman, D. J., Wolters, L. H., de Haan, E., & Boer, F. (2011). Developmental aspects of error and high-conflict-related brain activity in pediatric obsessive-compulsive disorder: A fMRI study with a Flanker task before and after CBT: fMRI during a Flanker task in pediatric OCD. *Journal of Child Psychology and Psychiatry*, *52*(12), 1251–1260. <https://doi.org/10.1111/j.1469-7610.2011.02439.x>

Imburgio, M. J., Banica, I., Hill, K. E., Weinberg, A., Foti, D., & MacNamara, A. (2020). Establishing norms for error-related brain activity during the arrow Flanker task among young adults. *NeuroImage*, *213*, 116694. <https://doi.org/10.1016/j.neuroimage.2020.116694>

Infantolino, Z. P., Luking, K. R., Sauder, C. L., Curtin, J. J., & Hajcak, G. (2018). Robust is not necessarily reliable: From within-subjects fMRI contrasts to between-subjects comparisons. *NeuroImage*, *173*, 146–152. <https://doi.org/10.1016/j.neuroimage.2018.02.024>

Keye, D., Wilhelm, O., Oberauer, K., & Ravenzwaaij, D. (2009). Individual differences in conflict-monitoring: Testing means and covariance hypothesis about the Simon and the Eriksen Flanker task. *Psychological Research Psychologische Forschung*, *73*(6), 762–776. <https://doi.org/10.1007/s00426-008-0188-9>

Kothe, C. A., & Makeig, S. (2013). BCILAB: A platform for braincomputer interface development. *Journal of Neural Engineering*, *10*(5), 056014. <https://doi.org/10.1088/1741-2560/10/5/056014>

Letkiewicz, A. M., Funkhouser, C. J., & Shankman, S. A. (2021). Childhood maltreatment predicts poorer executive functioning in adulthood beyond symptoms of internalizing psychopathology. *Child Abuse & Neglect*, *118*, 105140. <https://doi.org/10.1016/j.chiabu.2021.105140>

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*. <https://doi.org/10.3389/fnhum.2014.00213>

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behaviour Research and Therapy*, *33*(3), 335–343. <https://doi.org/10.1016/0005-7967(94)00075-U>

Luking, K. R., Nelson, B. D., Infantolino, Z. P., Sauder, C. L., & Hajcak, G. (2017). Internal Consistency of Functional Magnetic Resonance Imaging and Electroencephalography Measures of Reward in Late Childhood and Early Adolescence. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *2*(3), 289–297. <https://doi.org/10.1016/j.bpsc.2016.12.004>

Luks, T. L., Oliveira, M., Possin, K. L., Bird, A., Miller, B. L., Weiner, M. W., & Kramer, J. H. (2010). Atrophy in two attention networks is associated with performance on a Flanker task in neurodegenerative disease. *Neuropsychologia*, *48*(1), 165–170. <https://doi.org/10.1016/j.neuropsychologia.2009.09.001>

McDermott, J. M., Pérez-Edgar, K., & Fox, N. A. (2007). Variations of the flanker paradigm: Assessing selective attention in young children. *Behavior Research Methods*, *39*(1), 62–70. <https://doi.org/10.3758/BF03192844>

McTeague, L. M., Goodkind, M. S., & Etkin, A. (2016). Transdiagnostic impairment of cognitive control in mental illness. *Journal of Psychiatric Research*, *83*, 37–46. <https://doi.org/10.1016/j.jpsychires.2016.08.001>

Meyer, A., Bress, J. N., & Proudfit, G. H. (2014). Psychometric properties of the error-related negativity in children and adolescents: Psychometric properties of ERN in youth. *Psychophysiology*, *51*(7), 602–610. <https://doi.org/10.1111/psyp.12208>

Meyer, A., & Hajcak, G. (2019). A review examining the relationship between individual differences in the error-related negativity and cognitive control. *International Journal of Psychophysiology*, *144*, 7–13. <https://doi.org/10.1016/j.ijpsycho.2019.07.005>

Meyer, A., Lerner, M. D., De Los Reyes, A., Laird, R. D., & Hajcak, G. (2017). Considering ERP difference scores as individual difference measures: Issues with subtraction and alternative approaches: Difference scores and ERPs. *Psychophysiology*, *54*(1), 114–122. <https://doi.org/10.1111/psyp.12664>

Meyer, A., Riesel, A., & Proudfit, G. H. (2013). Reliability of the ERN across multiple tasks as a function of increasing errors: Reliability of the ERN across multiple tasks. *Psychophysiology*, *50*(12), 1220–1225. <https://doi.org/10.1111/psyp.12132>

Moore, R. D., Pindus, D. M., Drolette, E. S., Scudder, M. R., Raine, L. B., & Hillman, C. H. (2015). The persistent influence of pediatric concussion on attention and cognitive control during flanker performance. *Biological Psychology*, *109*, 93–102. <https://doi.org/10.1016/j.biopsycho.2015.04.008>

Moser, J. S., Moran, T. P., Schroder, H. S., Donnellan, M. B., & Yeung, N. (2013). On the relationship between anxiety and error monitoring: A meta-analysis and conceptual framework. *Frontiers in Human Neuroscience*, *7*. <https://doi.org/10.3389/fnhum.2013.00466>

National Institutes of Health and Northwestern University. (2021). *NIH Toolbox Scoring and Interpretation Guide for the iPad*.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. <https://doi.org/10.1126/science.aac4716>

Open Science Collaboration. (2012). An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science. *Perspectives on Psychological Science*, *7*(6), 657–660. <https://doi.org/10.1177/1745691612462588>

Ossola, P., Antonucci, C., Meehan, K. B., Cain, N. M., Ferrari, M., Soliani, A., Marchesi, C., Clarkin, J. F., Sambataro, F., & De Panfilis, C. (2021). Effortful control is associated with executive attention: A computational study. *Journal of Personality*, *89*(4), 774–785. <https://doi.org/10.1111/jopy.12614>

Pe, M. L., Vandekerckhove, J., & Kuppens, P. (2013). A diffusion model account of the relationship between the emotional flanker task and rumination and depression. *Emotion*, *13*(4), 739–747. <https://doi.org/10.1037/a0031628>

Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, *198*, 181–197. <https://doi.org/10.1016/j.neuroimage.2019.05.026>

Price, R. B., Brown, V., & Siegle, G. J. (2019). Computational Modeling Applied to the Dot-Probe Task Yields Improved Reliability and Mechanistic Insights. *Biological Psychiatry*, *85*(7), 606–612. <https://doi.org/10.1016/j.biopsych.2018.09.022>

Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Computation*, *20*(4), 873–922. <https://doi.org/10.1162/neco.2008.12-06-420>

Ratcliff, R., Philiastides, M. G., & Sajda, P. (2009). Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG. *Proceedings of the National Academy of Sciences*, *106*(16), 6539–6544. <https://doi.org/10.1073/pnas.0812589106>

Ratcliff, R., & Smith, P. L. (2004). A Comparison of Sequential Sampling Models for Two-Choice Reaction Time. *Psychological Review*, *111*(2), 333–367. <https://doi.org/10.1037/0033-295X.111.2.333>

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, *20*(4), 260–281. <https://doi.org/10.1016/j.tics.2016.01.007>

Ridderinkhof, K. R., Wylie, S. A., van den Wildenberg, W. P. M., Bashore, T. R., & van der Molen, M. W. (2021). The arrow of time: Advancing insights into action control from the arrow version of the Eriksen flanker task. *Attention, Perception, & Psychophysics*, *83*(2), 700–721. <https://doi.org/10.3758/s13414-020-02167-z>

Riesel, A. (2019). The erring brain: Error-related negativity as an endophenotype for OCD-A review and meta-analysis. *Psychophysiology*, *56*(4), e13348. <https://doi.org/10.1111/psyp.13348>

Riesel, A., Weinberg, A., Endrass, T., Meyer, A., & Hajcak, G. (2013). The ERN is the ERN is the ERN? Convergent validity of error-related brain activity across different tasks. *Biological Psychology*, *93*(3), 377–385. <https://doi.org/10.1016/j.biopsycho.2013.04.007>

Sanders, L. M. J., Hortobágyi, T., Balasingham, M., Van der Zee, E. A., & van Heuvelen, M. J. G. (2018). Psychometric Properties of a Flanker Task in a Sample of Patients with Dementia: A Pilot Study. *Dementia and Geriatric Cognitive Disorders Extra*, *8*(3), 382–392. <https://doi.org/10.1159/000493750>

Scudder, M. R., Lambourne, K., Drollette, E. S., Herrmann, S. D., Washburn, R. A., Donnelly, J. E., & Hillman, C. H. (2014). Aerobic Capacity and Cognitive Control in Elementary School-Age Children. *Medicine & Science in Sports & Exercise*, *46*(5), 1025–1035. <https://doi.org/10.1249/MSS.0000000000000199>

Smucny, J., Barch, D. M., Gold, J. M., Strauss, M. E., MacDonald, A. W., Boudewyn, M. A., Ragland, J. D., Silverstein, S. M., & Carter, C. S. (2019). Cross-diagnostic analysis of cognitive control in mental illness: Insights from the CNTRACS consortium. *Schizophrenia Research*, *208*, 377–383. <https://doi.org/10.1016/j.schres.2019.01.018>

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary* (3rd ed.). Oxford University Press.

Strong, C.-A. H., Tiesma, D., & Donders, J. (2011). Criterion validity of the Delis-Kaplan Executive Function System (D-KEFS) fluency subtests after traumatic brain injury. *Journal of the International Neuropsychological Society: JINS*, *17*(2), 230–237. <https://doi.org/10.1017/S1355617710001451>

Ulrichsen, K. M., Alnæs, D., Kolskår, K. K., Richard, G., Sanders, A.-M., Dørum, E. S., Ihle-Hansen, H., Pedersen, M. L., Tornås, S., Nordvik, J. E., & Westlye, L. T. (2020). Dissecting the cognitive phenotype of post-stroke fatigue using computerized assessment and computational modeling of sustained attention. *European Journal of Neuroscience*, *52*(7), 3828–3845. <https://doi.org/10.1111/ejn.14861>

Wechsler, D. (2001). *Wechsler test of adult reading: WTAR.* Psychological Corporation.

Weinberg, A., Liu, H., Hajcak, G., & Shankman, S. A. (2015). Blunted neural response to rewards as a vulnerability factor for depression: Results from a family study. *Journal of Abnormal Psychology*, *124*(4), 878–889. <https://doi.org/10.1037/abn0000081>

Weinberger, D. R., DeLisi, L. E., Neophytides, A. N., & Wyatt, R. J. (1981). Familial aspects of CT scan abnormalities in chronic schizophrenic patients. *Psychiatry Research*, *4*(1), 65–71. <https://doi.org/10.1016/0165-1781(81)90009-3>

Weintraub, S., Dikmen, S. S., Heaton, R. K., Tulsky, D. S., Zelazo, P. D., Bauer, P. J., Carlozzi, N. E., Slotkin, J., Blitz, D., Wallner-Allen, K., Fox, N. A., Beaumont, J. L., Mungas, D., Nowinski, C. J., Richler, J., Deocampo, J. A., Anderson, J. E., Manly, J. J., Borosh, B., … Gershon, R. C. (2013). Cognition assessment using the NIH Toolbox. *Neurology*, *80*(Issue 11, Supplement 3), S54–S64. <https://doi.org/10.1212/WNL.0b013e3182872ded>

White, C. N., Ratcliff, R., & Starns, J. J. (2011). Diffusion models of the flanker task: Discrete versus gradual attentional selection. *Cognitive Psychology*, *63*(4), 210–238. <https://doi.org/10.1016/j.cogpsych.2011.08.001>

White, C. N., Ratcliff, R., Vasey, M. W., & McKoon, G. (2010a). Using diffusion models to understand clinical disorders. *Journal of Mathematical Psychology*, *54*(1), 39–52. <https://doi.org/10.1016/j.jmp.2010.01.004>

White, C. N., Ratcliff, R., Vasey, M. W., & McKoon, G. (2010b). Anxiety enhances threat processing without competition among multiple inputs: A diffusion model analysis. *Emotion (Washington, D.C.)*, *10*(5), 662–677. <https://doi.org/10.1037/a0019474>

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics*, *7*. <https://doi.org/10.3389/fninf.2013.00014>

Wöstmann, N. M., Aichert, D. S., Costa, A., Rubia, K., Möller, H.-J., & Ettinger, U. (2013). Reliability and plasticity of response inhibition and interference control. *Brain and Cognition*, *81*(1), 82–94. <https://doi.org/10.1016/j.bandc.2012.09.010>

Wylie, S. A., van den Wildenberg, W. P. M., Ridderinkhof, K. R., Bashore, T. R., Powell, V. D., Manning, C. A., & Wooten, G. F. (2009). The effect of speed-accuracy strategy on response interference control in Parkinson’s disease. *Neuropsychologia*, *47*(8-9), 1844–1853. <https://doi.org/10.1016/j.neuropsychologia.2009.02.025>

Zelazo, P. D., Anderson, J. E., Richler, J., Wallner-Allen, K., Beaumont, J. L., Conway, K. P., Gershon, R., & Weintraub, S. (2014). NIH Toolbox Cognition Battery (CB): Validation of Executive Function Measures in Adults. *Journal of the International Neuropsychological Society*, *20*(6), 620–629. <https://doi.org/10.1017/S1355617714000472>

Zelazo, P. D., Anderson, J. E., Richler, J., Wallner-Allen, K., Beaumont, J. L., & Weintraub, S. (2013). Nih Toolbox Cognition Battery (cb): Measuring Executive Function and Attention. *Monographs of the Society for Research in Child Development*, *78*(4), 16–33. <https://doi.org/10.1111/mono.12032>

Ziegler, S., Pedersen, M. L., Mowinckel, A. M., & Biele, G. (2016). Modelling ADHD: A review of ADHD theories through their predictions for computational models of decision-making and reinforcement learning. *Neuroscience & Biobehavioral Reviews*, *71*, 633–656. <https://doi.org/10.1016/j.neubiorev.2016.09.002>