HHS Public Access

Author manuscript

J Atten Disord. Author manuscript; available in PMC 2020 November 01.

Published in final edited form as:

J Atten Disord. 2020 November ; 24(13): 1891–1904. doi:10.1177/1087054715627488.

Test-retest reliability and measurement invariance of executive function tasks in young children with and without ADHD

Sarah L. Karalunas, Ph.D.[1], **Karen Bierman, Ph.D.**[2], **Cynthia L. Huang-Pollock, Ph.D.**[2]

[1]Oregon Health & Science University, Department of Psychiatry

[2]The Pennsylvania State University, Department of Psychology

Abstract

Objective—Measurement reliability is assumed when executive function (EF) tasks are used to compare between groups or to examine relationships between cognition and etiologic and maintaining factors for psychiatric disorders. However, the test-retest reliabilities of EF tasks have rarely been examined in young children. Further, measurement invariance between typically-developing and psychiatric populations has not been examined.

Method—Test-retest reliability of a battery of commonly-used EF tasks was assessed in a group of children between the ages of 5–6 years old with (n=63) and without (n=44) ADHD.

Results—Few individual tasks achieved adequate reliability. However, CFA models identified two factors, working memory and inhibition, with test-retest correlations approaching 1.0. Multiple indicator multiple causes (MIMIC) models confirmed configural measurement invariance between the groups.

Conclusion—Problems created by poor reliability, including reduced power to index change over time or to detect relationships with functional outcomes, may be mitigated using latent variable approaches.

There is increasing recognition that behaviorally-based diagnostic categories, such as those used in DSM-5, result in the creation of groups that are phenotypically and mechanistically heterogeneous (Insel et al., 2010; Sanislow et al., 2010). To resolve the issues created by diagnostic heterogeneity, researchers have increasingly turned to endophenotype measures and biomarkers (Kendler & Neale, 2010; Lenzenweger, 2013; Nolen-Hoeksema & Watkins, 2011). Neurocognitive processes, such as working memory, inhibition, and other executive functions (EF), have been specifically highlighted by the recent NIMH Research Domain Criteria Initiative (RDoC) as potential endophenotypes or biomarkers that may help elucidate mechanisms of psychiatric disorders, aid in treatment matching, and facilitate development of novel treatments (Insel et al., 2010; Nolen-Hoeksema & Watkins, 2011; Sanislow et al., 2010). However, the psychometric properties of these measures may limit their use for these purposes, especially when used with young children.

Corresponding Author: Sarah L. Karalunas, PhD, ADHD Research Study UHN80R1, Oregon Health & Science University, 3181 SW Sam Jackson Park Rd, Portland, OR 97239-9979, Phone: 503-494-5780, Fax: 503-418-8416, karaluna@ohsu.edu.
Additional Authors: Karen Bierman, The Pennsylvania State University, 252 Moore Bldg, University Park, PA 16802
Cynthia Huang-Pollock, The Pennsylvania State University, Department of Psychology, 254 Moore Bldg, University Park, PA 16802

Attention Deficit Hyperactivity Disorder (ADHD) is emblematic of the problems created by etiologic heterogeneity within DSM diagnostic categories and measures of EF, speed/variability of response, and response to reward contingencies feature prominently in theoretical models of ADHD (Barkley, 1997; Castellanos, Sonuga-Barke, Milham, & Tannock, 2006; Diamond, 2005). However, a central concern and substantial obstacle to using these measures as endophenotypes or biomarkers is that their test-retest reliability is not always known (Kuntsi, Neale, Chen, Faraone, & Asherson, 2006). When test-retest reliability is low, it not only attenuates between group differences but also reduces statistical power to detect associations with genes, disease symptoms, or other outcome measures (Green et al., 2004; Kendler & Neale, 2010). Thus, better characterization of the reliability of cognitive tasks is essential.

Studies that have directly assessed test-retest reliability of EF tasks in children have focused primarily on middle-childhood and adolescence. In this age range, there is at least some evidence of adequate reliability for the most commonly used neurocognitive measures, including working memory span tasks, reaction time measures, and computerized measures of inhibitory control (Archibald & Kerns, 1999; Bishop, Aamodt-Leeper, Creswell, McGurk, & Skuse, 2001; Kindlon, Mezzacappa, & Earls, 1995; Kuntsi, Andreou, Ma, Borger, & Van der Meere, 2005; Kuntsi, Stevenson, Oosterlaan, & Sonuga-Barke, 2001; Soreni, Crosbie, Ickowicz, & Schachar, 2009; Thorell, 2007). However, reliability estimates are population specific, and tasks that demonstrate adequate or better reliability in middle childhood and adolescence may not be adequate for younger children. In particular, the rapid pace of maturation and development in early childhood may result in lower reliability estimates if the rate of development is not consistent across individuals. Similarly, individual differences in learning effects (which occur when the initial task exposure results in improved performance on later administrations) may also lower test -retest reliability. While learning and maturational effects are both reflected in test-retest data as improvements in performance between testing sessions, shorter test-retest intervals (several weeks) are primarily influenced by learning effects whereas longer intervals (several months) also capture the effects of maturation.

Neither maturation nor learning effects prevents a measure from being reliable, as long as those effects were consistent across the entire sample (Rousson, Gasser, & Burkhardt, 2002); however, how each of these processes affect task reliability remains unclear because few studies have assessed the reliability of neurocognitive tasks in early childhood. Gnys & Willis (1991) found good test-retest reliability for the Tower of Hanoi and verbal fluency tests (r_{xx}<.70) in a sample of 96 typically-developing preschool and kindergarten children. Beck et al. (2011) similarly found good test-retest reliabilities (ICCs<.69) for a series of tasks measuring inhibitory control for typically-developing children ranging from 2–5 years old. However, both studies used same-day test-retest intervals, potentially leading to inflated reliabilities, and both studies noted the need to examine longer retest intervals. Thorell et al. (2006) found adequate or better test-retest reliabilities for inhibitory control and working memory tasks in a group of 4–5 year-old children over a two-week retest interval, however the sample was small (n=22). One of the largest studies of executive task reliability in young children to-date comes from Willoughby & Blair (2011). Here, authors found moderate reliabilities (r_{xx}=.52 – .66) over a 2–4 week time period for a test battery of inhibitory

control, working memory, and attention shifting tasks in an epidemiological sample of over 100 preschool-age children. Thus while there is evidence of moderate reliability for at least some tasks in early childhood, the number of studies examining these effects is small and studies have used relatively short test-retest intervals, which emphasize learning effects, leaving the additional effects of maturation on task reliability unclear.

An additional issue is that traditional test-retest analyses rely on the correlations between individual tasks to assess reliability, which conflates true score and error variance and does not provide an adequate measure of the reliability of the underlying construct being measured. In contrast, latent variable models, such as confirmatory factor analysis, partition variance into common and specific (task specific and error) variance (Kline, 2013; Vandenberg & Lance, 2000). Thus, the test-retest reliability of the factor scores may better reflect the stability of the underlying EF abilities as compared to individual tests. Consistent with this hypothesis, Willoughby & Blair (2011) applied confirmatory factor analysis in their sample of preschool children to demonstrate that reliability of the underlying EF factor approached unity even when individual task reliabilities did not.

Although this suggests that latent variable approaches may be useful for improving psychometric properties of individual neurocognitive tasks, several questions remain. First, all of the studies described focus on the measurement of EF factor structure and reliability in a single, typically-developing population. If these methods are to be applied to studying psychiatric populations, it is necessary to establish that the tasks function similarly in each group. In other words, measurement invariance must be established (Muthén, 1989; Vandenberg & Lance, 2000; Woods, 2009). Measurement invariance refers to whether a measure, in this case tests of EF, are psychometrically equivalent in different groups. If they are not, then groups cannot be compared because the tests may be capturing fundamentally different processes in each of the groups, rendering comparisons meaningless.

A critical first step in assessing measurement invariance is the demonstration of configural invariance, which requires that the number and pattern of factor loadings must be the same between groups (Muthén, 1989; Sass, 2011; Vandenberg & Lance, 2000). The question of configural invariance is particularly relevant for young children with and without ADHD. In typically-developing preschool-age children, a single factor is often adequate to capture EF abilities (Wiebe, Espy, & Charak, 2008; Wiebe et al., 2011; Willoughby & Blair, 2011; but see Schoemaker et al., 2012), but during middle childhood, EF becomes more differentiated and is better represented by multiple factors (Lee, Bull, & Ho, 2013; Miyake, Friedman, Emerson, Witzki, & Howerter, 2000; Shing, Lindenberger, Diamond, Li, & Davidson, 2010). It remains unclear at which age multiple as opposed to single factor models become appropriate and whether factors representing different executive abilities are equally reliable. In addition, neuroimaging studies have found that children with ADHD are characterized by protracted development of prefrontal areas of the brain supporting EF (Gilliam et al., 2011; Kofler et al., 2013; Lijffijt, Kenemans, Verbaten, & van Engeland, 2005; Mackie et al., 2007; Shaw et al., 2007), and the disorder is often conceptualized as a maturational lag, suggesting that the appropriate factor model may differ between children with and without ADHD of the same age. If this were the case, latent variable approaches would be inappropriate for

between-group comparisons, and so establishing invariance is particularly important to inform additional studies of EF.

The current study examined the reliability of a battery of common neurocognitive tasks in a sample of kindergarten-age children with and without ADHD with assessments in the fall and spring of the kindergarten year. The study adds to a small literature examining individual task reliability in this age range and expands the range of test-retest intervals that have been examined. Further, we expand on prior research by directly comparing reliability in typically-developing and ADHD populations on individual tasks, as well as by using confirmatory factor analysis models to test for configural measurement invariance between children with and without ADHD and to establish the stability of latent EF factors.

Method

All data were collected as a part of a larger study examining the impact of a social-emotional intervention program on self-regulatory skills in young childhood. For the larger study, children either participated in: 1) a 30-session (16–18 week) small group social skills training intervention condition directly aimed at building social-emotional competency, self-regulatory skills, and EF; or 2) a control condition with the same number of sessions that focused on tutoring in emergent literacy skills (e.g. letter identification, letter-sound correspondence). The control condition was not expected to affect EF, self-regulatory skills, or social-emotional and behavioral outcomes. Because children in the intervention condition were intentionally provided instruction meant to improve executive processes, they were excluded from the current study.

Participants

One hundred and seven children ages 5–6 were recruited in two successive cohorts from 48 kindergarten classrooms in six Pennsylvania school districts that included both urban and rural areas. Brochures describing the study were distributed to parents of all 5–6 year-old children in the participating classrooms. Interested parents provided their contact information, as well as informed consent for child participation in the study. All procedures received approval from the university Institutional Review Board.

Initially, teachers completed two behavioral rating scales: Conners' ADHD Rating Scale, Short Form—Revised (CTRS-R) (Conners, 2003) and the DuPaul ADHD Rating Scale (ADHD-RS) (DuPaul, Power, Anastopoulos, & Reid, 1998). Children with elevated teacher ratings, as well as children without teacher-rated ADHD symptoms were identified. Their parents then completed a structured diagnostic interview (the Diagnostic Interview Schedule for Children (DISC-IV) (Shaffer, Fisher, Lucas, Dulcan, & Schwab-Stone, 2000) and parent-report rating forms (the Conners' Parent Rating Scale, Long Form—Revised (CPRS-R) (Conners, 2003) and the Behavioral Assessment Scale for Preschool Children, 2nd Edition (BASC-2) (Reynolds & Kamphaus, 2004).

Children were considered to have ADHD (n= 48) if: (a) they met full clinical criteria for a diagnosis of ADHD on the DISC-IV, including criteria for impairment, chronicity, and cross-situational severity; *and* (b) both the parent *and* teacher reported age-inappropriate

levels of inattention or hyperactivity defined as at least one T-score ≥ 60 (84th percentile) on the Cognitive Problems/Inattention, Hyperactivity, ADHD Index, or DSM-IV Total Index of the Conners' or the Hyperactivity or Attention Problems Indices of the BASC-2), or ≥ 3 inattentive symptoms or ≥ 3 hyperactive/ impulsive symptoms or ≥ 4 total symptoms endorsed as "often" or "very often" on the]ADHD-RS. Children were considered to have emerging ADHD (n=15) if they did not meet full diagnostic criteria on the DISC-IV, but did have elevated levels of inattention or hyperactivity based on at least one parent-report measure *and* at least one teacher-report measure (criterion b above). Finally, children were considered non-ADHD controls (n= 44) if: (a) they did not meet diagnostic criteria for ADHD on the DISC-IV; *and* (b) teacher ratings of behavior on all relevant indices of the Conners' and BASC-2 T-Scores ≤ 59; *and* (c) and the total number of symptoms endorsed following the "or" algorithm yielded ≤ 2 inattentive symptoms, and ≤ 2 hyperactive/ impulsive symptoms, and ≤ 3 total symptoms. For all children, dimensional scores of inattention and hyperactivity symptom counts were determined following DSM-IV field trials (Lahey et al., 1994) using an "or" algorithm between parent report on the DISC-IV and teacher report on the ADHD-RS (where a rating of "often" or "almost always" would indicate that a symptom was present).

Given that many symptoms, particularly inattention symptoms, have low endorsement at this age but are highly endorsed as children reach middle childhood (Curchack-Lichtin, Chacko, & Halperin, 2013) and that diagnostic stability is improved across early and middle childhood when sub-threshold symptoms are considered (Bauermeister et al., 2011), children with full and emerging ADHD were grouped together for between-group analyses. However, as noted in the Results section, primary results were all confirmed using a continuous ADHD symptom count in addition to the categorical diagnostic indicator to ensure that this grouping strategy did not account for results. Description of sample characteristics can be found in Table 1.

Exclusionary criteria—Exclusionary criteria for the larger study included: a) parent report of a sensorimotor disability, frank neurological disorder, or psychosis; b) estimated FSIQ < 70 as measured by a 2-subtest short form (Vocabulary and Matrices) of the Stanford-Binet, 5th Edition; c) low levels of English proficiency that preclude children from completing the assessment battery; or d) if they were in a temporary custody situation with uncertain outcome. Children taking psychotropic medications were not excluded from the study. One child was prescribed Focalin and asked to discontinue medication 24 hours prior to each testing session. A second child was prescribed Strattera, which could not be safely discontinued, and participated while taking their regular dose at both the test and retest visits.

Cognitive Testing

Children were assessed at two time points at the school, in a quiet room outside of the classroom setting and away from peers. The time between assessments ranged between 15–26 weeks (mean= 21.13, SD= 1.69). All children were tested individually by trained examiners.

Inhibitory Control—Inhibitory control was assessed with five tasks: the Walk-a-line Slowly task (Kochanska, Murray, Jacques, Koenig, & Vandegeest, 1996), the Peg Tapping Task (Diamond & Taylor, 1996), the Head-Toes-Knees-Shoulders task (HTKS) (Ponitz et al., 2008), a Choice Delay Task (Sonuga-Barke, Taylor, Sembi, & Smith, 1992), and a Go/No-Go Task (Berlin & Bohlin, 2002),.

For the Walk-a-line Slowly task children were asked to walk along a six-foot piece of string taped to the floor as the examiner timed them. Children were then asked to repeat the task twice, walking slower, and then walking even slower -- as slowly as they could. The total score represented the average percentage by which a child reduced his/her speed on successive trials. This task has demonstrated adequate inter-rater reliability with preschool children (intra-class correlation = .98), indicating that raters are able to accurately determine the timing difference between trials (Smith-Donald, Raver, & Hayes, 2007).

In the Peg Tapping Task, children were asked to tap their peg twice when the interviewer tapped once, and vice versa. After a short set of practice items, their final score was the number of correct trials out of 16 total trials.

The HTKS task is a more complex version of the Head-to-Toes task. In this task, children habituated to several oral commands (e.g., “touch your head” and “touch your toes”). They were then asked to play “a silly game” in which,, in response to the command, “Touch your toes,” they were to touch their head. Then, they played another silly game in which they were to touch their knees when asked to touch their shoulders and vice versa. Children earned 2 points for a correct response, 0 points for an incorrect response, and 1 point if they made any motion to the incorrect response but then self-corrected. The outcome variable used was the total number of correct points, with a maximum of 40 points possible.

In the Choice-Delay Task (Sonuga-Barke et al., 1992), children chose between two rewards: 1) a one-point reward available after two seconds or 2) a two-point reward available after 30-seconds. Each trial began immediately after the reward was received from the preceding trial. Children had 20 trials in which they were instructed to earn as many points as possible. The variable used in analyses was the percentage of choices for the 2-point, delayed reward.

In the Go/No-go task children viewed four stimuli (blue triangle, blue square, red triangle, and red square) and were asked to make a key press every time they saw a blue shape (target, 75% of trials), but to withhold a response when they viewed a red shape (25% of trials). Each stimulus appeared for 1000 ms; children were allowed a total of 2000 ms to respond. Inhibitory control outcome measures included percent correct hits and commission errors. Reaction time on correct hits and standard deviation of reaction time for correct hits were also recorded as measures of processing speed, which is a separable EF factor, at least in the middle childhood age range (Rose, Feldman, & Jankowski, 2011).

Working Memory—Both verbal and visuospatial working memory were assessed. Verbal working memory was assessed using the Backward Word Span task (Carlson, 2005; Davis & Pratt, 1995). For this task, children listened to a list of words read out loud and then were asked to repeat the words in backwards order. The list started with one word, and increased

by one additional word with successive trials. Children received a score equal to the highest number of words they were able to repeat correctly in reverse order.

Visuospatial working memory was assessed using the Finger Windows task from the Wide Range Assessment of Memory and Learning, Second Edition (WRAML-2) (Sheslow & Adams, 2003). In the forwards condition of this task, the child watched as the examiner put a pencil in a series of holes on a card. The child then recreated this series. The WRAML-2 Finger Windows task includes only a forwards condition; however, a backwards condition was also created for this study in which children needed to point to the holes the examiner identified in reverse order. In both conditions, children received one point for every correct sequence recalled.

The Dimensional Change Card Sort (DCCS) (Frye, Zelazo, & Palfai, 1995) was also included as a measure of working memory, consistent with prior literature indicating that performance in early childhood is related to ability to use higher-order if-then rules (Zelazo, 2004; Zelazo & Frye, 1998) and use working memory to overcome response conflict (Munakata, 2001). Children were shown picture cards that varied along the dimensions of color and shape (e.g. red and blue, rabbits and boats). After learning to sort the cards according color, children were then asked to sort the cards according to shape instead. The score represented the number of trials (out of 6) in which the child correctly shifted sets after the sorting criteria changed.

Data Analysis

Two children initially recruited into the study did not participate in either the pre- or post-test due to absence from school on the day of testing, and were excluded from all analysis. Seven additional children completed only part of the pre- or post-test battery and 13 children's files were lost to file corruption on the computerized go/no-go task at either Time 1 or Time 2. Finally, two scores were identified as outliers (> 5 Standard Deviations from the sample mean): one on the Go/No-go Reaction Time measure and one on the Walk-a-line Slowly task. The outlying scores were treated as missing data for these tasks. In assessing reliability of individual tasks, children were only included if they completed the task at both time points (pre- and post-test). The final Ns for each task are reported in Table 2. In assessing factor structure and factor reliability, all available children were included in analyses using the full information maximum likelihood algorithms in MPLUS to handle missing data.

Learning and Maturation Effects—Learning/maturation effects were assessed with a multivariate repeated-measures ANOVA including the full battery of EF tasks. Time was the within-subjects factor and ADHD diagnosis as the between-subjects factor. A main effect of Time indicates the presence of learning/maturation effects and a Time*ADHD interaction indicates differences in these effects based on ADHD status.

Reliability of Individual Tasks—Test-retest reliability for each group was calculated as the age-partial inter-class product-moment correlation coefficients (Pearson correlation) between the two test administrations using SPSS (Rousson et al., 2002). There are no firm criteria for what constitutes "good" reliability. Prior research examining reliability of

neurocognitive tasks has adopted the criteria that reliabilities between .50–.70 are "adequate" and those above .70 are "good" (Kindlon et al., 1995; Kuntsi, Stevenson, et al., 2001). We adopt the same criteria here. The age effects were included to account for differences in age at the initial assessment time period (Kail, 2007; Williams, Ponesse, Schachar, Logan, & Tannock, 1999). Fisher's r-to-z tests were used to compare the correlation coefficients and determine whether these differed significantly between the diagnostic groups.

Factor Structure and Reliability of Factors—Confirmatory factor analyses (CFA) were conducted in MPLUS v.7.2. First, a series of CFA models were tested in the two diagnostic groups separately. Then, a multiple-indicators multiple-causes (MIMIC) model (Muthén, 1989) in which ADHD diagnostic status was included as a covariate was used to test for measurement invariance (Muthén, 1989; Woods, 2009). MIMIC models can be used to test for configural invariance by including grouping variables as covariates in the factor model, rather than testing separate models for each group as is required for multiple-group CFA (Kim, Yoon, & Lee, 2012; Muthén, 1989). MIMIC models assume equivalent factor loading across groups, rather than testing this directly (Muthén, 1989), thus they cannot be used for testing metric and other types of invariance. However, MIMIC models are preferred for testing measurement invariance in small samples (Muthén, 1989; Woods, 2009), and configural invariance must be established for subsequent tests of metric and other types of measurement invariance to be meaningful (Vandenberg & Lance, 2000). Further, using a MIMC model approach, both categorical and continuous measures of ADHD could be used in tests of measurement invariance, which is not possible with a multiple groups approach. Thus, using the MIMIC model approach to establish configural invariance is a first and critical step in determining the utility of latent variable models for comparing between typically-developing and psychiatric populations. MIMIC models were estimated using ADHD diagnostic status as a covariate. Each direct effect was estimated in a separate model with false discovery rate correction employed (Benjamini & Hochberg, 1995). Results were confirmed with total ADHD symptoms (continuous) as a covariate using the same procedures. As a final step in the analyses, pre- and post-test assessments were used in a single CFA model to determine the test-retest reliability (stability) of the identified factors across time.

Power analysis—MIMIC models estimate a much smaller number of parameters than multiple group CFA approaches by treating the grouping variable as a covariate, thus making them more appropriate in small samples. Monte Carlo simulation power analysis in MPLUS indicated adequate power (>.80) for all models, including the MIMIC models. Thus, models were adequately powered to detect configural invariance where it existed.

Results

Learning and Maturational Effects

A multivariate repeated-measures ANOVA including the EF measures revealed a significant main effect of Time ($p < .001$) and Condition ($p < .001$), but no significant Time*Diagnosis interaction effect ($p = .205$), indicating the presence of learning and maturational effects

(i.e., improvements over time) but no difference in these effects based on diagnostic status. Follow-up univariate tests for the main effect of Time indicated significant learning effects for eight of the twelve measures. In all cases performance improved at the second administration of the test. See Table 2 for means and standard deviations at each time point and summary of significance tests for learning/maturation effects. Results were also confirmed using the continuous ADHD symptom count rather than categorical diagnosis.

Reliability of Individual Measures

Table 3 shows the correlation matrix for all EF tasks in the full sample and Table 4 shows the age-partial test-retest correlations by group. In the typically-developing group, three tasks reached adequate or better levels of reliability: mean and standard deviation of RT from the Go/No-go task and HTKS. In the ADHD sample, six tasks reached adequate or better reliability: DCCS, Finger Windows Backwards, HTKS, as well as Go/No-go Hit Accuracy and commission errors, and Peg Tapping. Test-retest correlations for Word Span, Go/No-go Hit rate, and Peg Tapping were significantly higher in the ADHD than in the typically-developing group.

Confirmatory Factor Analysis and Reliability of Latent Variables

Confirmatory Factor Analysis (CFA)—A three factor CFA with 1) Inhibitory Control (GNG Accuracy, GNG Commissions, Peg Tapping, Delay Aversion, Walk-a-Line Slowly), 2) Working Memory (Backward Word Span, DCCS, HTKS, Finger Windows Forward, Finger Windows Backward), and 3) Processing Speed (GNG RT, GNG SDRT) was fit in the full sample. The model did not converge and provided a poor fit to the data. Problems with model convergence were a result of the linear dependency between the two indicators of the Processing Speed factor: RT and SDRT. This factor was also problematic in that CFA factors defined by only two indicators are not identified and so a minimum of three indicators is recommended to include a factor in CFA (Muthen & Muthen, 2009). Thus, in remaining analyses we focus on a two-factor CFA model using only the Working Memory and Inhibition factors.

In the two-factor CFA model for the full sample, the Delay Aversion and Walk-a-line Slowly tasks did not load significantly on the Inhibitory Control factor at either the test or re-test time point. Further, a model excluding these tasks showed significantly better fit to the data at both times, and so these tasks were excluded from additional analyses. Thus, the final two-factor CFA model included an Inhibitory Control factor (GNG Accuracy, GNG Commissions, Peg Tapping) and a Working Memory factor (Backward Word Span, DCCS, HTKS, Finger Windows Forward, Finger Windows Backward). Based on modification indices provided in MPLUS, the residual covariances for Finger Windows Forward and Finger Windows Backward were allowed to correlate. The two-factor CFA model fit well in the full sample (Adjusted BIC= 2073.0; $\chi^2[18]=10.1$, $p=.928$; RMSEA=0.0; CFI=1.00) and is shown in Figure 1. The 2-factor model fit significantly better than a 1-factor CFA model (Adjusted BIC= 2126.5; $\chi^2[20]=66.7$, $p<.001$; RMSEA=.15; CFI=0.84).

Measurement Invariance—The two-factor model fit well in both the typically-developing and ADHD groups separately at both Time 1 and Time 2 (all $p> .05$ for χ^2 tests,

all RMSEA< .03, all CFI> 0.98) and fit significantly better than the one factor model for both groups at both time points. We next combined the two groups into a single MIMIC analysis in which the ADHD diagnostic indicator was regressed onto the factors to assess measurement invariance between the diagnostic groups. A significant direct effect of ADHD diagnosis on the factor indicators (i.e. on observed task performance) would indicate a lack of measurement invariance. At Time 1, the model fit was good (Adjusted BIC= 2048.1; $\chi^2[24]=19.8$, $p=.711$; RMSEA=0.0; CFI=1.00). ADHD diagnosis was significantly related to the factors, indicating that the well-documented ADHD-related deficits in working memory and inhibitory control are captured by the latent variables; however, there were no significant direct effects of ADHD diagnosis on any of the factor indicators (all $p>.05$). The lack of direct effects indicates measurement invariance for this time point. All results were replicated for Time 2, confirming measurement invariance across the diagnostic groups at both time points. All results were also confirmed using total ADHD symptoms as a continuous covariate instead of the categorical diagnostic indicator.

Stability of Factor Scores—Finally, a two time point, two-factor CFA model was fit to establish the stability of the factor scores across time. Residual covariances between Time 1 and Time 2 tasks scores were allowed to correlate. The model was a good fit for the data (Adjusted BIC= 3927.0; $\chi^2[90]=93.0$, $p=.393$; RMSEA=.018; CFI=.99). The full model is shown in Figure 2. Consistent with prior studies the stability of the latent variables approached unity. The correlation (stability coefficient) for the Inhibitory Control factor was .98 and for the Working Memory factor was .99.

Discussion

Adequate test-retest reliability for neurocognitive tasks is not only critical to the accurate estimation of between group effect sizes (Huang-Pollock, Karalunas, Tam, & Moore, 2012), but also to the ability to detect associations between these cognitive processes, putative genetic mechanisms, symptom domains, and other outcome measures (Green et al., 2004; Kendler & Neale, 2010; Kuntsi et al., 2005; Kuntsi et al., 2006). In the current study, individual measures used to assess neurocognitive functioning in young children showed a wide range of reliabilities, with only a handful achieving adequate levels. Despite moderate or worse reliabilities for many individual tasks, latent variable modeling indicated test-retest correlations approaching 1.0 for factors measuring both Inhibition and Working Memory. High reliability was achieved over a longer test-retest interval that has previously been examined in this age range, suggesting that maturational effects did not limit the reliability of the underlying EF constructs as measured by the latent variable models.

In addition, MIMIC models confirmed configural invariance across typically-developing and ADHD samples, which is a critical for establishing that latent variable approaches can be used for comparison of these populations. The question of configural invariance between diagnostic groups in this age range is not trivial. In particular, in preschool-age children EF appears to be best represented as a unidimensional construct captured by a single factor (Hughes, Ensor, Wilson, & Graham, 2009; Wiebe et al., 2011; Willoughby & Blair, 2011). However, in middle childhood a multidimensional factor structure with three to five factors is most often found, suggesting that EF abilities become more differentiated with age (Lee et

al., 2013; Miyake et al., 2000; Shing et al., 2010). Children in our age range are on the cusp of these two time periods. Further, children with ADHD are often conceptualized as having maturational delays in prefrontal regions supporting EF (Gilliam et al., 2011; Kofler et al., 2013; Lijffijt et al., 2005; Mackie et al., 2007; Shaw et al., 2007), which implies that different factor structures may be needed to capture EF in typically-developing and ADHD children of the same age. However, this was not the case. In the current study, the same multi-dimensional factor solution fit the ADHD and typically-developing groups equally well. Although groups differed in factor means, capturing well-documented ADHD-related deficits on both working memory and inhibitory control, there were no indirect effects of ADHD diagnosis on individual tasks. The lack of indirect effects confirms that the tasks functioned similarly in both groups. Future studies with larger samples will be needed to establish strong measurement invariance, including metric and scalar invariance. However, these results suggest that latent variable approaches are a viable solution for addressing problems created by low individual task reliabilities, which reduce power for detecting between-group effects, limit the ability to detect developmental change, and interfere with the use of neurocognitive measures in endophenotype studies.

Although the clear recommendation from this study is to capitalize on latent variable approaches to maximize reliability, this may not always be possible and so several individual tasks that did not achieve adequate reliability are worth highlighting. Consider, first, the delay aversion task. Previous delay tasks for which adequate reliability has been reported in preschool-aged children either used test-retest intervals that were exceptionally short (~15 minutes), which would have artificially inflated their reliabilities, or used tangible rewards (cookies, pennies) (Beck et al., 2011; Thorell & Wahlstedt, 2006). Thus, it may be that young children require tangible rewards to elicit reliable reward choices.

Second, in middle childhood, mean RT and SDRT have each demonstrated adequate reliability and shown promising associations with behavioral symptoms and putative genetic mechanisms of ADHD (Kuntsi et al., 2005; Kuntsi, Oosterlaan, & Stevenson, 2001; Wood, Asherson, van der Meere, & Kuntsi, 2010). In contrast, in this study, reaction time measures failed to achieve adequate reliability in young children with ADHD and other studies have found that they do not differentiate young children with and without ADHD (Kalff et al., 2005). The current results suggest that the inability to differentiate groups may be due to low reliability, rather than because young children with ADHD do not have deficits in processing speed and efficiency.

This difference in interpretation has important implications for the search for endophenotypes in particular. One requirement of cognitive endophenotypes is that they should be stable over time regardless of disease course (Gottesman & Gould, 2003). Thus, if speed and variability of information processing do not characterize young children with ADHD, then this would argue against their mediating between gene action and symptom domains. However, if measures of speed and variability of information processing are unreliable in young children with ADHD, then this suggests that reliable measures first need to be identified before developmental trends can be assessed.

This problem with interpretation is not limited to reaction time measures. There are a growing number of studies aimed at characterizing the relationships between neurocognitive processes and ADHD symptom domains. These include studies comparing the relative heritability of different cognitive processes (Stins et al., 2005), the strength of relations between endophenotype and symptom domains in different age groups (Brocki, Fan, & Fossella, 2008), and the strength of relations between different neurocognitive processes and symptom domains. In each case, specific attention to task reliability is critical, given that any differences in the strength of association may reflect differences in task reliability rather than the underlying construct being assessed.

The current study confirmed configural invariance only for the working memory and inhibitory control factors. A third processing speed factor could not be tested in the CFA models because there were too few and too highly correlated indicators for this factor. Future test batteries that incorporate a larger number of tasks assessing processing speed will be required to address this limitation. In addition, the Delay Aversion and Walk-a-Line Slowly tasks did not load on the Inhibitory Control factor. The inhibition of gross motor movement required for Walk-a-Line Slowly makes it substantially different from the other inhibitory control tasks, which emphasize inhibition of prepotent fine motor movements and require stimulus discrimination and choice between two responses options. The low factor loading of the Delay Aversion task is consistent with prior work suggesting reward delay tasks tap a different type of inhibitory control than non-reward based tasks (Willoughby, Kupersmidt, Voegler-Lee, & Bryant, 2011; Zelazo & Carlson, 2012). In particular, whereas the majority of inhibitory tasks for the current study tapped “cool” executive processes (i.e., inhibitory control elicited in relatively emotion-free contexts) the delay aversion tasks taps a “hot” inhibitory process, elicited by the emotionally-salient rewards offered. Future studies in which more “hot” executive tasks are administered could test for the presence and reliability of an additional “hot” executive factor. Finally, additional studies with larger sample sizes will be required to confirm the current results and to test for other types of measurement invariance, including invariance of factor loadings between groups.

Conclusions and Future Directions

Current results provide evidence that many measures of cognitive function used with older children and adults do not achieve adequate levels of reliability in preschool-aged children. The use of measures with poor reliability hinders the field’s ability to identify associations with symptom domains and may lead to erroneous conclusions about the developmental stability of cognitive phenotypes in psychiatric disorders. Selection of reliable measures is increasingly important as greater emphasis is placed on the use of neurocognitive measures to identify genetic mechanisms with relatively small individual effects and the current study suggests that latent variable approaches are a viable solution to problems created by low reliability of individual tasks.

References

Archibald SJ, Kerns KA. Identification and description of new tests of executive functioning in children. Child Neuropsychology. 1999; 5(2):115–129.

Barkley RA. Behavioral inhibition, sustained attention, and executive functions: Constructing a unifying theory of ADHD. Psychological Bulletin. 1997; 121(1):65–94. [PubMed: 9000892]

Bauermeister JJ, Bird HR, Shrout PE, Chavez L, Ramírez R, Canino G. Short-Term Persistence of < i>DSM-IV</i> ADHD Diagnoses: Influence of Context, Age, and Gender. Journal of the American Academy of Child & Adolescent Psychiatry. 2011; 50(6):554–562. [PubMed: 21621139]

Beck DM, Schaefer C, Pang K, Carlson SM. Executive function in preschool children: test-retest reliability. Journal of Cognitive Development. 2011; 12(2):169–193.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological). 1995:289–300.

Berlin L, Bohlin G. Response Inhibition, Hyperactivity, and Conduct Problems Among Preschool Children. Journal of Clinical Child and Adolescent Psychology. 2002; 31(2):242–251. [PubMed: 12056107]

Bishop DVM, Aamodt-Leeper G, Creswell C, McGurk R, Skuse DH. Individual Differences in Cognitive Planning on the Tower of Hanoi Task: Neuropsychological Maturity or Measurement Error? Journal of Child Psychology and Psychiatry. 2001; 42(4):551–556. [PubMed: 11383971]

Brocki K, Fan J, Fossella J. Placing neuroanatomical models of executive function in a developmental context: imaging and imaging genetics strategies. Annals of the New York Academy of Sciences. 2008; 1129:246–255. [PubMed: 18591485]

Carlson SM. Developmentally sensitive measures of executive function in preschool children. Developmental Neuropsychology. 2005; 28(2):595–616. [PubMed: 16144429]

Castellanos FX, Sonuga-Barke E, Milham MP, Tannock R. Characterizing cognition in ADHD: Beyond executive dysfunction. Trends in Cognitive Sciences. 2006; 10(3):117–123. [PubMed: 16460990]

Conners, CK. Conners' rating scales: Revised technical manual. New York, NY: Multi-Health Systems; 2003.

Curchack-Lichtin JT, Chacko A, Halperin JM. Changes in ADHD Symptom Endorsement: Preschool to School Age. Journal of Abnormal Child Psychology. 2013:1–12. [PubMed: 22773360]

Davis HL, Pratt C. The development of children's theory of mind: The working memory explanation. Australian Journal of Psychology. 1995; 47(1):25–31.

Diamond A. Attention-deficit disorder (attention-deficit/hyperactivity disorder without hyperactivity): A neurobiologically and behaviorally distinct disorder from attention-deficit/hyperactivity disorder (with hyperactivity). Development and Psychopathology. Special Issue: Integrating Cognitive and Affective Neuroscience and Developmental Psychopathology. 2005; 17(3):807–825.

Diamond A, Taylor C. Development of an aspect of executive control: Development of the abilities to remember what I said and to"Do as I say, not as I do". Developmental psychobiology. 1996; 29(4):315–334. [PubMed: 8732806]

DuPaul, G, Power, T, Anastopoulos, A, Reid, R. ADHD Rating Scale—IV: Checklists, Norms, and Clinical Interpretation. NY, NY: Guilford Press; 1998.

Frye D, Zelazo PD, Palfai T. Theory of mind and rule-based reasoning. Cognitive Development. 1995; 10(4):483–527.

Gilliam M, Stockman M, Malek M, Sharp W, Greenstein D, Lalonde F, Shaw P. Developmental trajectories of the corpus callosum in attention-deficit/hyperactivity disorder. Biological Psychiatry. 2011; 69(9):839–846. [PubMed: 21247556]

Gnys JA, Willis G. Validation of executive function tasks with young children. Developmental Neuropsychology. 1991; 7(4):487–501.

Gottesman II, Gould TD. The endophenotype concept in psychiatry: Etymology and strategic intentions. American Journal of Psychiatry. 2003; 160(4):636–645. [PubMed: 12668349]

Green MF, Nuechterlein KH, Gold JM, Barch DM, Cohen J, Essock S, Heaton RK. Approaching a consensus cognitive battery for clinical trials in schizophrenia: the NIMH-MATRICS conference to select cognitive domains and test criteria. Biological Psychiatry. 2004; 56(5):301–307. [PubMed: 15336511]

Huang-Pollock CL, Karalunas SL, Tam H, Moore AN. Evaluating Vigilance Deficits in ADHD: A Meta-Analysis of CPT Performance. Journal of Abnormal Psychology. 2012; 121(2):360–371. [PubMed: 22428793]

Hughes C, Ensor R, Wilson A, Graham A. Tracking executive function across the transition to school: A latent variable approach. Developmental Neuropsychology. 2009; 35(1):20–36.

Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, Wang P. Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. The American Journal of Psychiatry. 2010; 167(7):748–751. [PubMed: 20595427]

Kail R. Longitudinal evidence that increases in processing speed and working memory enhance children's reasoning. Psychological Science. 2007; 18(4):312–313. [PubMed: 17470254]

Kalff AC, De Sonneville LMJ, Hurks PPM, Hendriksen JGM, Kroes M, Feron FJM, Jolles J. Speed, speed variability, and accuracy of information processing in 5 to 6-year-old children at risk of ADHD. Journal of the International Neuropsychological Society. 2005; 11(2):173-173-183. [PubMed: 15962705]

Kendler K, Neale M. Endophenotype: a conceptual analysis. Molecular Psychiatry. 2010; 15(8):789–797. [PubMed: 20142819]

Kim ES, Yoon M, Lee T. Testing Measurement Invariance Using MIMIC Likelihood Ratio Test With a Critical Value Adjustment. Educational and Psychological Measurement. 2012; 72(3):469–492.

Kindlon DJ, Mezzacappa E, Earls F. Psychometric properties of impulsivity measures: Temporal stability, validity and factor structure. Journal of Child Psychology and Psychiatry. 1995; 36(4):645–661. [PubMed: 7650088]

Kline, R. Exploratory and Confirmatory Factor Analysis. In: Petscher, Y, Schattsschneider, C, editors. Applied Quantitative Analysis in the Social Sciences. New York, NY: Routledge; 2013. 171–207.

Kochanska G, Murray K, Jacques TY, Koenig AL, Vandegeest KA. Inhibitory control in young children and its role in emerging internalization. Child Development. 1996; 67(2):490–507. [PubMed: 8625724]

Kofler MJ, Rapport MD, Sarver DE, Raiker JS, Orban SA, Friedman LM, Kolomeyer EG. Reaction time variability in ADHD: A meta-analytic review of 319 studies. Clinical Psychology Review. 2013; 33(6):795–811. [PubMed: 23872284]

Kuntsi J, Andreou P, Ma J, Borger N, Van der Meere J. Testing assumptions for endophenotype studies in ADHD: Reliability and validity of tasks in a general population sample. BMC Psychiatry. 2005; 5(40):e.1–e.11.

Kuntsi J, Neale BM, Chen W, Faraone SV, Asherson P. The IMAGE project: methodological issues of the molecular genetic anslysis of ADHD. Behavioral and Brain Functions. 2006; 2(27)

Kuntsi J, Oosterlaan J, Stevenson J. Psychological mechanisms in hyperactivity: I Response inhibition deficit, working memory impairment, delay aversion, or something else? Journal of Child Psychology and Psychiatry. 2001; 42(2):199–210. [PubMed: 11280416]

Kuntsi J, Stevenson J, Oosterlaan J, Sonuga-Barke EJS. Test-retest reliability of a new delay aversion task and executive function measures. British Journal of Developmental Psychology. 2001; 19(3):339–348.

Lahey BB, Applegate B, McBurnett K, Biederman J, Greenhill L, Hynd GW, Shaffer D. DSM-IV Field Trials for Attention-Deficit Hyperactivity Disorder in Children and Adolescents. American Journal of Psychiatry. 1994; 151(11):1673–1685. [PubMed: 7943460]

Lee K, Bull R, Ho RM. Developmental changes in executive functioning. Child Development. 2013; 84(6):1933–1953. [PubMed: 23550969]

Lenzenweger MF. Thinking clearly about the endophenotype–intermediate phenotype–biomarker distinctions in developmental psychopathology research. Development and Psychopathology. 2013; 25(4pt2):1347–1357. [PubMed: 24342844]

Lijffijt M, Kenemans JL, Verbaten MN, van Engeland H. A Meta-Analytic Review of Stopping Performance in Attention-Deficit/Hyperactivity Disorder: Deficient Inhibitory Motor Control? Journal of Abnormal Psychology. 2005; 114(2):216–222. [PubMed: 15869352]

Mackie S, Shaw P, Lenroot R, Pierson R, Greenstein D, Nugent T, Rapoport J. Cerebellar development and clinical outcome in attention deficit hyperactivity disorder. American Journal of Psychiatry. 2007; 164(4):647–655. [PubMed: 17403979]

Miyake A, Friedman NP, Emerson MJ, Witzki AH, Howerter A. The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. Cognitive Psychology. 2000; 41(1):49–100. [PubMed: 10945922]

Munakata Y. Graded representations in behavioral dissociations. Trends in Cognitive Sciences. 2001; 5(7):309–315. [PubMed: 11425620]

Muthen B, Muthen L. Topic 1. Introductory - advanced factor analysis and structural equation modeling with continuous outcomes. 2009 May.2014

Muthén BO. Latent variable modeling in heterogeneous populations. Psychometrika. 1989; 54(4):557–585.

Nolen-Hoeksema S, Watkins ER. A heuristic for developing transdiagnostic models of psychopathology: Explaining multifinality and divergent trajectories. Perspectives on Psychological Science. 2011; 6:589–609. [PubMed: 26168379]

Ponitz CEC, McClelland MM, Jewkes AM, Connor CM, Farris CL, Morrison FJ. Touch your toes! Developing a direct measures fo behavioral regulation in early childhood. Early Childhood Research Quarterly. 2008; 23:141–158.

Reynolds, C, Kamphaus, R. Behavioral Assessment System for Children, 2nd Ed. Manual. Circle Pines, MN: AGS Publishing; 2004.

Rose SA, Feldman JF, Jankowski JJ. Modeling a cascade of effects: the role of speed and executive functioning in preterm/full-term differences in academic achievement. Developmental Science. 2011; 14(5):1161–1175. [PubMed: 21884331]

Rousson V, Gasser T, Burkhardt S. Assessing intra-rater, interrater and test-retest reliability of continuous measurement. Statistics in Medicine. 2002; 21:3431–3446. [PubMed: 12407682]

Sanislow CA, Pine DS, Quinn KJ, Kozak MJ, Garvey MA, Heinssen RK, Cuthbert BN. Developing constructs for psychopathology research: Research domain criteria. The Journal of Abnormal Psychology. 2010; 119(4):631–639. [PubMed: 20939653]

Sass D. Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. Journal of Psychoeducational Assessment. 2011

Schoemaker K, Bunte T, Wiebe SA, Espy KA, Deković M, Matthys W. Executive function deficits in preschool children with ADHD and DBD. Journal of Child Psychology and Psychiatry. 2012; 53(2):111–119. [PubMed: 22022931]

Shaffer D, Fisher P, Lucas C, Dulcan M, Schwab-Stone M. NIMH Diagnostic Interview Schedule for Children Version IV (NIMH DISC-IV): Description, differences from previous versions, and reliability of some common diagnoses. Journal of the American Academy of Child & Adolescent Psychiatry. 2000; 39(1):28–38. [PubMed: 10638065]

Shaw P, Eckstrand K, Sharp W, Blumenthal J, Lerch JP, Greenstein D, Rapoport JL. Attention-deficit/hyperactivity disorder is characterized by a delay in cortical maturation. PNAS Proceedings of the National Academy of Sciences of the United States of America. 2007; 104(49):19649–19654.

Sheslow, D, Adams, W. Wide Range Assessment of Memory and Learning 2nd Ed (WRAML-2): Administration and Technical Manual. Wilmington, DE: Wide Range; 2003.

Shing YL, Lindenberger U, Diamond A, Li S-C, Davidson MC. Memory maintenance and inhibitory control differentiate from early childhood to adolescence. Developmental Neuropsychology. 2010; 35(6):679–697. [PubMed: 21038160]

Smith-Donald R, Raver CC, Hayes T. Preliminary construct and concurrent validity of the Preschool Self-Regulation Assessment (PSRA) for field-based research. Early Childhood Research Quarterly. 2007; 22(2):173–187.

Sonuga-Barke E, Taylor E, Sembi S, Smith J. Hyperactivity and delay aversion: I. The effect of delay on choice. Journal of Child Psychology and Psychiatry. 1992; 33(2):387–398. [PubMed: 1564081]

Soreni N, Crosbie J, Ickowicz A, Schachar R. Stop Signal and Conners' Continuous Performance Tasks : Test-Retest Reliability of Two Inhibition Measures in ADHD Children. Journal of Attention Disorders. 2009; 13:137–143. [PubMed: 19429883]

Stins J, Leo MJdS, Groot A, Polderman T, Caroline GCMvB, Boomsma D. Heritability of Selective Attention and Working Memory in Preschoolers. Behavior Genetics. 2005; 35(4):407–416. [PubMed: 15971022]

Thorell LB. Do delay aversion and executive function deficits make distinct contributions to the functional impact of ADHD symptoms? A study of early academic skill deficits. Journal of Child Psychology and Psychiatry. 2007; 48(11):1061–1070. [PubMed: 17995481]

Thorell LB, Wahlstedt C. Executive functioning deficits in relation to symptoms of ADHD and/or ODD in preschool children. Infant and Child Development. 2006; 15:503–518.

Vandenberg RJ, Lance CE. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. Organizational research methods. 2000; 3(1):4–70.

Wiebe SA, Espy KA, Charak D. Using confirmatory factor analysis to understand executive control in preschool children: I. Latent structure. Developmental Psychology. 2008; 44(2):575. [PubMed: 18331145]

Wiebe SA, Sheffield T, Nelson JM, Clark CA, Chevalier N, Espy KA. The structure of executive function in 3-year-olds. Journal of Experimental Child Psychology. 2011; 108(3):436–452. [PubMed: 20884004]

Williams BR, Ponesse JS, Schachar RJ, Logan GD, Tannock R. Development of inhibitory control across the life span. Developmental Psychology. 1999; 35(1):205–213. [PubMed: 9923475]

Willoughby M, Blair C. Test-retest reliability of a new executive function battery for use in early childhood. Child Neuropsychology. 2011; 17(6):564–579. [PubMed: 21714751]

Willoughby M, Kupersmidt J, Voegler-Lee M, Bryant D. Contributions of hot and cool self-regulation to preschool disruptive behavior and academic achievement. Developmental Neuropsychology. 2011; 36(2):162–180. [PubMed: 21347919]

Wood AC, Asherson P, van der Meere JJ, Kuntsi J. Separation of genetic influences on attention deficit hyperactivity disorder symptoms and reaction time performance from those on IQ. Psychological Medicine: A Journal of Research in Psychiatry and the Allied Sciences. 2010; 40(6):1027–1037.

Woods CM. Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. Multivariate Behavioral Research. 2009; 44(1):1–27. [PubMed: 26795105]

Zelazo PD. The development of conscious control in childhood. Trends in Cognitive Sciences. 2004; 8(1):12–17. [PubMed: 14697398]

Zelazo PD, Carlson SM. Hot and cool executive function in childhood and adolescence: Development and plasticity. Child Development Perspectives. 2012; 6(4):354–360.

Zelazo PD, Frye D. Cognitive complexity and control: II. The development of executive function in childhood. Current Directions in Psychological Science. 1998:121–126.

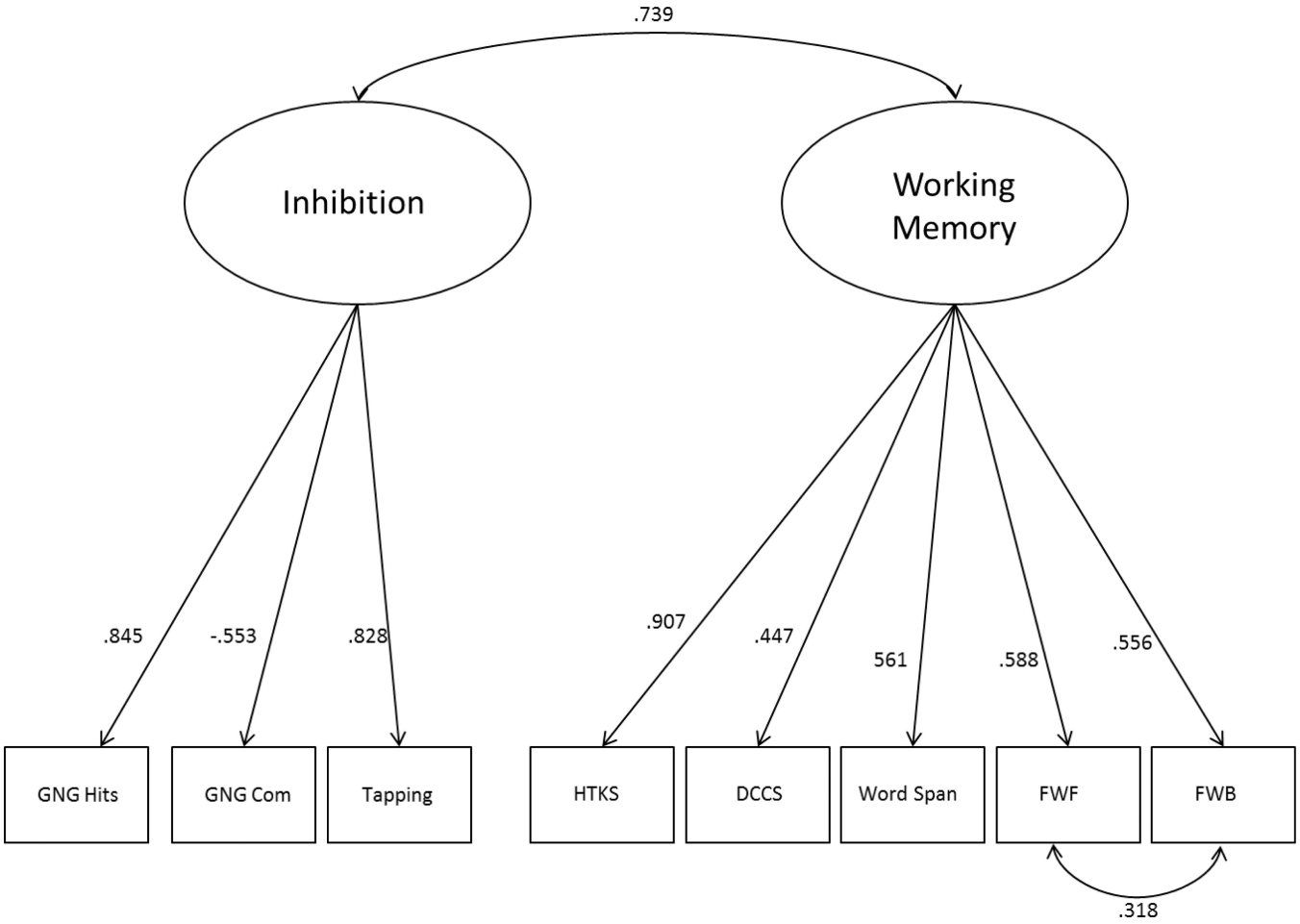

Figure 1.

shows the two-factor CFA model for the full sample at Time 1, including the factor loading for each indicator and the correlation between factors. GNG Acc= Go/no-go Accuracy, GNG Com= Go/no-go Commission Errors, HTKS= Head-Toes-Knees-Shoulder, DCCS= Dimensional Change Card Sort, Word Span= Backward Word Span, FWF= Finger Windows Forward, FWB=Finger Windows Backward.

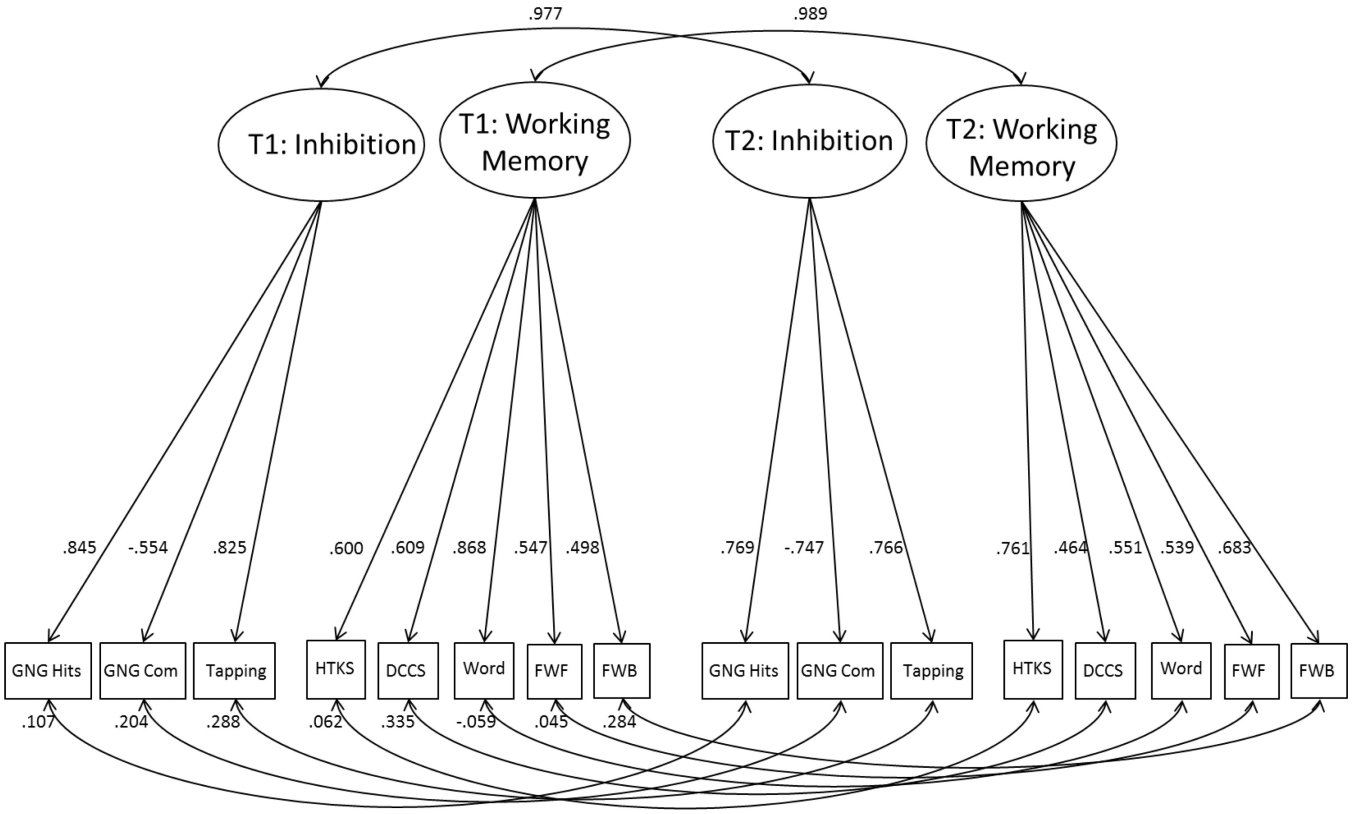

Figure 2.

shows the two-time point, two-factor CFA model, including factor loading at each time point and the correlation (stability) of factors across the two testing occasions. T1= Time 1 (test), T2= Time 2 (re-test), GNG Acc= Go/no-go Accuracy, GNG Com= Go/no-go Commission Errors, HTKS= Head-Toes-Knees-Shoulder, DCCS= Dimensional Change Card Sort, Word Span= Backward Word Span, FWF= Finger Windows Forward, FWB=Finger Windows Backward.

Table 1

Sample Characteristics.

Measure	***Control*** **(n=44)**	***ADHD*** **(n=63)**	***F***	***ES*** **(η^2)**
Gender (% male)	55.3%	68.1%	$\chi^2(1)=1.03$	
Age (in months)	70.16 (3.55)	69.45 (3.99)	0.81	0.01
Stanford-Binet Abbreviated IQ	106.79 (9.78)	97.45 (14.50)	12.46**	0.11
Conners' Parent Cognitive Problems/Inattention	46.16 (4.55)	59.68 (12.28)	42.56***	0.29
Conners' Parent Hyperactivity	46.23 (5.36)	62.33 (11.79)	63.05***	0.38
Conners' Teacher Cognitive Problems/Inattention	46.97 (3.34)	64.27 (16.69)	39.76***	0.28
Conners' Teacher Hyperactivity	44.68 (3.05)	64.80 (14.35)	72.50***	0.41

Note.

** = $p<.01$,

$p<.001$

Table 2

Learning Effects.

Measure	N	Time 1	Time 2	Learning Effect (t)
Delay Aversion (% choice for large reward)	103	0.31 (.19)	.31 (.22)	0.32
Backward Word Span (Total points)	97	2.25 (.72)	2.44 (.65)	2.15*
DCCS (% correct post-switch)	104	0.82 (.35)	.94 (.20)	3.91***
Finger Windows				
Forwards (Total Points)	104	6.73 (3.29)	7.84 (3.24)	3.17**
Backwards (Total points)	104	3.11 (2.35)	3.96 (2.42)	4.12***
Go/No-Go				
Hits	91	53.27 (6.53)	53.76 (5.32)	0.87
Commissions	91	3.76 (3.80)	3.97 (3.61)	0.65
RT (ms)	90	510.21 (95.84)	472.03 (85.32)	3.86***
SDRT (ms)	90	374.97 (86.02)	351.37 (74.85)	2.83***
Peg Tapping (Total points)	104	14.04 (3.95)	14.45 (3.03)	1.50
Walk-a-line (Total points)	103	0.85 (.86)	1.30 (1.44)	2.83**
Head-Toes-Knees Shoulder (Total points)	103	27.19 (13.07)	32.30 (8.14)	5.23***

Note. DCCS= Dimensional Change Card Sort; RT= Reaction Time; SDRT= Standard Deviation of Reaction Time.

* = $p < .05$,

** =$p < .01$,

*** =$p < .001$

J Atten Disord. Author manuscript; available in PMC 2020 November 01.

Table 3

Correlation Matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1 Delay Aversion	--																						
2 Word Span	−.06	--																					
3 DCCS	−.08	.21*	--																				
4 HTKS	.00	.50***	.40***	--																			
5 GNG Hits	.04	.35***	.23*	.53***	--																		
6 GNG Commissions	−.17	−.20*	−.24*	−.35**	−.47***	--																	
7 Peg Tapping	.09	.41***	.28**	.57***	.67***	−.41***	--																
8 FWF	−.07	.38***	.27**	.53***	.36***	−.19	.33**	--															
9 FWB	.01	.27**	.34**	.51***	.32**	−.17	.27**	.64***	--														
10 GNG RT	.11	−.28*	.03	−.35**	−.30	−.18	−.27	−.25*	−.19	--													
11 GNG SDRT	.11	−.49**	−.15	−.61***	−.84***	−.17	−.62***	−.46**	−.40**	.98***	--												
12 Walk-a-Line Slowly	.09	.12	.24*	.14	.04	−.06	.13	.11	.14	−.01	.11	--											
13 Delay Aversion	.28**	.12	.11***	.14	.08	−.07	.09	.08	.10	.06	.00	.24*	--										
14 Word Span	−.01	.25*	.31***	.44**	.40**	−.14	.37**	.42***	.46***	−.13	−.39*	.33**	.08	--									
15 DCCS	.13	.20*	.48*	.39***	.30**	−.09	.24*	.19	.20*	−.03	−.35*	.14	.14	.21	--								
16 HTKS	−.03	.40***	.53**	.70***	.63***	−.36**	.61***	.40***	.44***	−.26	−.66***	.16	.13	.44**	.45***	--							
17 GNG Hits	.20	.35**	.27**	.52***	.66***	−.48***	.66***	.31**	.24*	−.08	−.54**	.03	.20	.30*	.35**	.60***	--						
18 GNG Commissions	−.18	−.27*	−.29*	−.50***	−.62***	.56***	−.63***	−.21*	−.19	.06	.36*	−.10	−.15	−.26	−.35**	−.58***	−.98***	--					
19 Peg Tapping	.12	.30**	.35**	.56***	.60***	−.42***	.74***	.34**	.28**	−.30*	−.60***	.10	.17	.34*	.33**	.55***	.63***	−.67***	--				
20 FWF	.13	.31**	.21	.46***	.31**	−.14	.38***	.38***	.42***	−.19	−.40**	.19	.03	.30**	.29**	.32**	.37**	−.33**	.36**	--			
21 FWB	.07	.34**	.32	.60***	.40***	−.19	.34**	.52***	.60***	−.21	−.43**	.16	.26*	.37**	.29**	.47***	.42***	−.35**	.35**	.47***	--		
22 GNG RT	.02	−.33**	−.15	−.35**	−.36*	.04	−.20	−.28*	−.23*	.57***	.79**	.01	−.11	−.22	−.16	−.33*	−.34*	.14	−.22	−.29*	−.23*	--	
23 GNG SDRT	−.01	−.39**	−.20	−.38**	−.41**	.11	−.33*	−.32**	−.28**	.39**	.76**	.07	−.15	−.28*	−.23	−.40**	−.48***	.20	−.30*	−.32***	−.27*	.92***	--
24 Balance	.09	.11	.21*	.18	.09	−.05	.09	.18	.29**	−.12	−.14	.06	.02	.28*	.02	.11	.09	−.01	.07	.16	.28**	−.30**	−.33**

Note. DCCS= Dimensional Change Card Sort; HTKS= Head-Toes-Knees-Shoulders; GNG= Go/No-go; FWF= Finger Windows Forwards; FWB= Finger Windows Backwards.

* = $p < .05$,

** =$p < .01$,

*** = $p < .001$

Table 4

Individual task reliability coefficients.

Measure	Control			ADHD			Comparison
	N	Pearson r	Age-Partial r	N	Pearson r	Age-Partial r	Fisher's r-to-z (p)
Delay Aversion(% choice for large reward)	44	.080	.083	59	.375	.377	.127
Backward Word Span (Total points)	44	−.063	−.067	60	.460	.465	.005
DCCS (% correct post-switch)	44	.479	.486	60	.493	.501	.922
Finger Windows							
Forwards (Total Points)	44	.303	.324	60	.284	.289	.850
Backwards (Total Points)	44	.519	.463	60	.543	.525	.688
Go/No-go							
Hits	37	.130	.125	54	.596	.588	.013
Commissions	37	.360	.361	54	.568	.559	.253
RT (ms)	37	.625	.586	53	.350	.303	.106
SDRT (ms)	37	.575	.540	53	.446	.364	.316
Peg Tapping (Total Points)	44	.093	.088	60	.727	.720	<.001
Walk-a-line	44	.021	.067	59	.064	.064	.834
Head-Toes-Knees Shoulder (Total points)	44	.683	.685	59	.590	.584	.408

Note. DCCS= Dimensional Change Card Sort; RT= Reaction Time; SDRT= Standard Deviation of Reaction Time.

* = $p < .05$,

** =$p<.01$,

*** =$p<.001$