**Symposium Title:** The NIH Toolbox Cognitive Battery for Intellectual and Developmental Disabilities: The Latest Findings and Potential Utility for Clinical Trials

**Chair:** David Hess

**Discussant:** Audrey Thurm

**Overview:** Accurate measurement of cognitive functions in people with intellectual disability, especially those with greater than mild impairment, has been marked by numerous assessment limitations. Translational studies in our field have generated several preclinical models of specific causes of neurodevelopmental disorders, some of which have been taken to human trials, but with limited success when the outcomes of interest have focused on maladaptive behavior. Now, however, there is a push to utilize more objective cognitive tests to detect efficacy. The four presentations in this symposium will focus on the utility of the National Institutes of Health Toolbox Cognitive Battery (NIH-TCB), a component of the NIH Toolbox for Assessment of Neurological and Behavioral Function, which was developed to standardize evaluations in specific clinical populations for investigations of neurological development and change, disease recovery, and therapeutic interventions (http://www.nihtoolbox.org). The first presentation by Rebecca Shields (UC Davis) describes the latest results from a large NICDH-funded study aimed to provide accommodations and modifications to the battery to better suit people with disabilities (in this case fragile X syndrome and Down syndromes) as well as novel findings regarding reliability and validity of subtests in this battery. The second talk by Aaron Kaat (Northwestern University) will drill down on the executive function tests within the battery and describe efforts to extend measurement to lower functioning or younger participants – indeed there is a major dearth of cognitive measures for these individuals. The third talk (by Haleigh Scott at Davis) examines questions about how executive dysfunction at the cognitive level may impact maladaptive behavior, particularly in the areas of irritable and dysregulated mood and ADHD symptoms. The final talk (by Elizabeth Berry-Kravis of Rush University) will present an application of the NIH-TCB as outcomes in the context of treatment of a Niemann-Pick type C (NPC), a lysosomal storage disease characterized by progressive motor and cognitive deterioration, examining whether the measures might help to demonstrate stability or protection from deterioration. All together, these presentations will provide a good overview of the successes and challenges of validation of cognitive tests for challenging populations characterized by intellectual impairments, and a good update on progress of the potential utility of the NIH-TCB specifically as potential outcome measures in treatment trials.

**Paper 1 of 4**

**Paper Title:** Feasibility, Validity, and Reliability of the NIH Toolbox Cognitive Battery in Intellectual Disability

**Authors:** Rebecca Shields1, Elizabeth Berry-Kravis2, Karen Riley3, Keith Widaman4, Aaron Kaat5, Jeanine Coleman5, Claire Michalak2, Forrest McKenzie1, Andrea Drayton1, Richard Gershon5, David Hess1

**Introduction:** Due to floor effects of many tests, behavior challenges, and other assessment or measurement issues, individuals with intellectual disability (ID) can be a difficult population in which to accurately measure cognition. This is becoming an increasingly important issue as treatment trials begin to focus on cognitive outcomes. The NIH Toolbox Cognitive Battery (NIH-TCB), a tablet-based assessment developed for individuals ages 3 to 85, has strong potential as a series of outcome measures for this population (Hessl et al., 2016). The seven tests of the NIH-TCB measure the following constructs: attention and inhibitory control, cognitive flexibility, processing speed, working memory, episodic memory, receptive vocabulary, and reading/letter

---

1 University of California, Davis, MIND Institute
2 Rush University Medical Center
3 University of Denver, Morgridge College of Education
4 University of California Riverside, Graduate School of Education
5 Northwestern University, Feinberg School of Medicine
6 National Institute of Mental Health

Page 1 of 6
identification. This multi-site study aims to evaluate the feasibility, reliability, and convergent validity of the NIH-TCB in ID (R01HD076189). Here we present results from the full sample’s first two study visits. While analyses are still underway, we will also present results within each diagnosis group, as well as on group differences on Toolbox test performance.

Methods: Participants include 244 individuals with ID: 92 with Down syndrome, 77 with fragile X syndrome, and 75 with ID of other or unknown cause, with a required mental age of at least 3 years. Participants were seen across three sites (Rush University, UC Davis, and University of Denver) and completed a battery of cognitive assessments. To validate the seven constructs measured by the NIH-TCB, the following measures were also collected: Stanford Binet 5 (SBS), Leiter-3 Forward Memory, PPVT-4, WJ-4 Letter Word ID (WJ-LW), Conners Kiddie Continuous Performance Test 2 (KCPT2), WPPSI-4 Bug Search, and NEPSY-2 Inhibition. Participants returned approximately 3-6 weeks later and completed the NIH-TCB again to examine test-retest reliability.

Results: Feasibility, measured by the percent of participants who passed practice and received a valid score as judged by study validity criteria, ranged from 59.4% to 95.9% of the sample when considering only the usual test scores. Flanker and DCCS also have experimental Developmental Extension (DEXT) versions that extend the test range developmentally downward when needed. When including the DEXT scores for those tests, feasibility was better, ranging from 68.6% to 95.9%. On Flanker, 77.2% of the sample received a valid computed score and 96.7% received a valid score when also including DEXT scores. On DCCS, 59.4% received a valid computed score and 94.6% received a valid score when also including DEXT scores. A problem with inconsistent understanding of Pattern Comparison in some participants, also observed in very young typically developing children, necessitated development of a new processing speed task, which will be described.

The following tests demonstrated moderate convergent validity: Flanker (computed score with KCPT2 hit SD of reaction time), DCCS (computed score with NEPSY Inhibition errors), Pattern Comparison (with WPPSI-4 Bug Search), List Sort (with SBS Verbal Working Memory), and Picture Sequence Memory (with Leiter-3 Forward Memory). Picture Vocabulary and Oral reading demonstrated strong convergent validity (with the PPVT-4 and WJ-LW, respectively). All Toolbox tests had a stronger correlation with their convergent measure than their discriminant measure except DCCS.

With a mean retest interval of 33 days, following tests demonstrated moderate test-retest reliability: DCCS (computed score) and Picture Sequence Memory. The remaining tests demonstrated strong test-retest reliability: Flanker (computed score), Pattern Comparison, List Sort, Picture Vocabulary, and Oral Reading.

Discussion: The results of the study thus far demonstrate good feasibility of the NIH-TCB for ID, especially above about a mental age of 4.0, moderate to strong reliability, and adequate validity for most tests. Analyses by group will help to determine consistency or differences in these psychometric properties in each disorder, and future longitudinal sensitivity to change analyses will help to inform their potential selection as outcomes for future trials. Also the Fluid Reasoning composite score, combining performance across multiple tests, may yield stronger psychometric properties and consideration as clinical outcomes – we will be examining its performance as well. A limitation of the study was finding an appropriate convergent measure for DCCS and Flanker with a low enough floor such that it would be valid and feasible in this population. Additionally, DEXT scores are not scaled with the computed scores and thus the two score types cannot be combined. The smaller samples for DCCS and Flanker scores thus likely impact the strength of the effects for these two tests. Finally, it is noted that some of the convergent validity tests have measurement limitations as well. Despite these challenges, the NIH-TCB holds promise for use to obtain accurate measurement in ID, and perhaps future improvements to scoring or test items on DCCS DEXT and Flanker DEXT might improve these two tests’ validity and feasibility.
References/Citations:

**Paper 2 of 4**

**Title**: Assessing Early Executive Functioning in Intellectual Disability with the NIH Toolbox

**Authors**: Aaron J. Kaat⁶, Rebecca H. Shields⁷, Jeanine Coleman⁸, Karen Riley⁴, Elizabeth M. Berry-Kravis⁹, Richard C. Gershon¹, David R. Hessl²

**Introduction**: Executive functioning (EF) is a broad domain involving several cognitive functions. EF tasks are not fully developed until young adulthood, but early EF skills begin to emerge in infancy and toddlerhood (Carlson, 2005). The NIH Toolbox includes two EF measures: Flanker Inhibitory Control, and the Dimensional Change Card Sort Test (Gershon et al., 2013; Zelazo et al., 2013). Among typically-developing samples, a high proportion of 3- and 4-year-olds were not able to pass these EF tasks, and as such versions were created to be more developmentally appropriate (referred to as developmental extension, or DEXT versions). Early investigations suggest lower (relative) feasibility for the standard versions of these EF tasks among children and young adults with intellectual and developmental disabilities (I/DD; Hessl et al., 2016). The purpose of this study was to investigate the utility of the DEXT versions of these tests in I/DD.

**Methods**: Participants were 64 children and young adults with Fragile X, Down Syndrome, or Other I/DD recruited to a longitudinal study evaluating the NIH Toolbox. DEXT measures were offered in 122 assessments for DCCS and 124 assessments for Flanker (5 children with 1; 58 with 2; 1 child with 3 assessments). We provide a descriptive evaluation of DEXT utilization, progression through DEXT phases, and potential scoring approaches.

**Results**: DEXT was utilized in 32% and 61% of available assessments for Flanker and DCCS, respectively. On Flanker, 33 (27%) entered DEXT after practice, whereas 7 (8% of eligible) entered DEXT after taking the 20 standard Fish items but were well-below chance and entered DEXT. Flanker DEXT always begins at the least difficult level, where 17 (42% of eligible) assessments were unable to match on directionality and were never required to inhibit flanking stimuli. Of the assessments that required inhibiting flanking fish, 3 (13% of eligible) failed the highest salience (a different size, wider spacing, and a different color); 7 (30% of eligible) failed with moderate salience (spacing and color); 6 (26% of eligible) failed with the lowest salience (color). After the additional training within DEXT, the regular/standard difficulty level was passed in 9 (39% of eligible) assessments. On DCCS DEXT, standard practice was failed in 51 (42%) of assessments (entering DEXT at the lowest level), whereas practice was passed in 23 (32% of eligible) assessments but the individual was unable to successfully shift sets (entering DEXT at the “Separated” phase). Of the assessments which included the lowest phase of DCCS DEXT, 15% were unable to consistently match (20 exact matches). The second phase required matching by size and then reversing (big-to-small or small-to-big), which had a lower success rate (13% of eligible) than other phases of DEXT. Separated and integrated phases, which require matching by shape or color and a set shift between the two, were rarely administered. However, the highest DCCS DEXT level was met in 11 assessments, where 6 (55% of eligible) were able to successfully pass set shifting, suggesting that the additional training was effective.

---

⁶ Northwestern University Feinberg School of Medicine
⁷ University of California, Davis MIND Institute
⁸ University of Denver, Morgridge College of Education
⁹ Rush University Medical Center
Discussion: Assessing executive functioning in young children is a known difficulty. This was one of the first studies to assess executive functioning with the NIH Toolbox among individuals with a younger mental age. Similar to studies with chronologically young samples, passing rates on standard executive functioning tasks is reduced. Using developmental extensions to create more salient differences appears to help many of these children and young adults with disabilities. The DEXT version of Flanker provides matching on directionality and then increases the salience between target and flanking fish, allowing more individuals to be successful. The DEXT version of DCCS provides additional practice with matching prior to requiring set shifting, which allows greater group differentiation. However, it appears that some DCCS DEXT phases are out of order for relative difficulty. Future research is necessary to optimize scoring models and further refine the NIH Toolbox EF measures in I/DD.

References/Citations:


Paper Title: Impact of Hyperactivity and Irritability Symptoms on NIH Toolbox Executive Function Tasks in Intellectual Disability

Authors: Haleigh M. Scott10, Rebecca H. Shields1, Jeanine Coleman2, Karen Riley3, Elizabeth M. Berry-Kravis4, Richard C. Gershon2, Aaron J. Kaat2, David Hessl3

Introduction: Executive function (EF) refers to a set of cognitive controls such as flexibility, inhibition, and working memory. EF skills begin to emerge in early childhood and development continues into young adulthood. Deficits in EF have been linked to a variety of negative outcomes such as poor academic functioning and a variety of mental health problems (Snyder, Miyake & Hankin, 2015). While people with developmental disabilities have EF deficits as a component of their disability, people with developmental disabilities are also likely to have additional concerns such as behavior problems and/or ADHD. Comorbid difficulties such as ADHD and behavior problems are often the target of both psychological and pharmacological treatment and they may compound the EF deficits in this population. This study explored symptoms of irritability and hyperactivity in relation to Toolbox EF task performance in people with developmental disabilities.

Methods: Participants were 242 children and young adults with Down syndrome, Fragile X or nonspecific intellectual disability (ID) ages 8-25 years involved in a longitudinal study of the NIH Toolbox Cognitive Battery. The two EF measures on the NIH Toolbox are the Flanker Inhibitory Control- measuring inhibitory control and attention, and the Dimensional Change Card Sort Test (DCCS)- measuring cognitive flexibility (Zelazo et al., 2013). Parent report on symptoms of hyperactivity and irritability using the Aberrant Behavior Checklist (ABC) were used to determine if these symptoms were related to performance on the EF based tasks while controlling for age and functioning level using the Vineland Adaptive Behavior Scales (VABS).

10 University of California, Davis MIND Institute
2 Northwestern University Feinberg School of Medicine
3 University of Denver, Morgridge College of Education
4 Rush University Medical Center
**Results:** Age and functioning level were the only significant predictors of participant’s ability to pass training and complete both the Flanker and DCCS as opposed to being dropped to the developmental extension tasks or being rated as having an invalid assessment. Among those who completed the tasks, approximately 61% of participants for Flanker and 78% of participants for DCCS, age and functioning level remained significant predictors of performance on both tasks. For DCCS, parent reported irritability was a significant predictor of task performance above and beyond age and functioning. For the Flanker task, both parent reported irritability and parent reported hyperactivity were significant as predictors.

**Discussion:** The current study demonstrated that for people able to pass the training phases, performance on EF tasks was related to symptoms of hyperactivity and irritability as rated by parents. This suggested that EF deficits in people with developmental disabilities may be related to both of their cognitive functioning and comorbid behavioral concerns. Given the nature of the study design we are unable to determine directionality of this relationship. If deficits in EF are directly related to symptomology then treatments designed to target behavioral comorbidities may also improve EF. Further research, specifically longitudinal study designs, are needed to explore this topic as well as research exploring these topics in different groups, for example people with comorbid ASD.

**References/Citations:**

**Paper Title:** Use of the NIH Toolbox to Follow Cognition in Niemann-Pick Type C Patients Receiving Intrathecal 2-hydroxypropyl-beta-cyclodextrin (VTS-270) to Limit Progression of Disease

**Authors:** Elizabeth Berry-Kravis MD PhD, Jamie Chin BS, Claire Michalak BS, Rush University Medical Center

**Introduction:** The NIH Toolbox Cognitive Battery (NIH-TCB) has been used in studies of neurodegenerative disease such as Alzheimer disease in elderly populations to track rate of progression and response to interventions, but has not been implemented as of yet in studies of Pediatric neurodegenerative disorders. We piloted the NIH-TCB to track long-term cognitive outcomes in a cohort of patients with Niemann-Pick type C (NPC), a lysosomal storage disease with a broad age of presentation and relentlessly progressive motor and cognitive deterioration, being treated with VTS-270 to attempt to stabilize disease progression. Potential advantages of use of this measure over standard cognitive batteries in this population and setting were 1) reduction of motor burden for individuals with motor limitations that might hamper performance on standard cognitive batteries; 2) capability for more frequent administration than standard IQ tests; 3) ability to use the same test for a broad range of function across the cohort as opposed to different versions of tests for different ages (e.g. WISC/WAIS); 4) attractiveness of the iPad administration for children.

**Methods:** The NIH-TCB was administered to participants in a treatment trial and those in an expanded access program at baseline prior to initiation of intrathecal VTS-270 every 2 weeks by LP infusion, and after initiation of treatment every 6 months. The NIH-TCB version administered included 7 subtests, Dimensional Change Card Sort (cognitive flexibility), Flanker (inhibition and visual attention), Picture Sequence Memory (episodic memory), List Sorting (working memory), Pattern Comprehension (processing speed), Oral Reading Recognition (letter identification and reading), and Picture Vocabulary (receptive vocabulary). Raw scores on subtests (to evaluate for loss of function) and crystallized and fluid cognition composite standard scores (to evaluate relative developmental pace) were tracked over time from initiation of treatment. Standard IQ scores on the WPPSI/WISC/WAIS were also obtained yearly for the patients in this cohort.
Results: A total of 23 of 37 patients with NPC treated with VTS-270 at the site were able to complete the NIH-TCB at multiple visits (5 were too young, 7 were too impaired, 2 have not yet completed a retest due to recent enrollment). These patients ranged in age at baseline from 2-31 years of age. They have been followed with the NIH-TCB for a time range of 6 months to 4.5 years. Fluid cognitive composite range at baseline was 25-90 and crystallized composite range at baseline was 40-112, consistent with the expected more preserved functional ability in crystallized cognition in this disease. Most patients had stable performance on raw scores on most measures across multiple years. NIH-TCB results appear to track with IQ results in follow up testing, however full analyses are to be completed as more patients complete retesting.

Discussion: Preliminary evidence from this study suggests the NIH-TCB is a feasible measure with which to track cognitive progress in Pediatric and young adult patients in natural history and treatment studies for neurodegenerative diseases and has better ease of administration and better accommodates wide ranges of function and age than standard IQ tests and may provide advantages for assessing cognition more accurately in motor-impaired populations.