AMERICAN JOURNAL ON INTELLECTUAL AND DEVELOPMENTAL DISABILITIES
2017, Vol. 122, No. 3, 215–234

DOI: 10.1352/1944-7558-122.3.215

The Arizona Cognitive Test Battery for Down Syndrome: Test-Retest Reliability and Practice Effects

Jamie O. Edgin, Payal Anand, Tracie Rosser, Elizabeth I. Pierpont, Carlos Figueroa, Debra Hamilton, Lillie Huddleston, Gina Mason, Goffredina Spanò, Lisa Toole, Mina Nguyen-Driver, George Capone, Leonard Abbeduto, Cheryl Maslen, Roger H. Reeves, and Stephanie Sherman

Abstract

A multisite study investigated the test-retest reliability and practice effects of a battery of assessments to measure neurocognitive function in individuals with Down syndrome (DS). The study aimed to establish the appropriateness of these measures as potential endpoints for clinical trials. Neurocognitive tasks and parent report measures comprising the Arizona Cognitive Test Battery (ACTB) were administered to 54 young participants with DS (7–20 years of age) with mild to moderate levels of intellectual disability in an initial baseline evaluation and a follow-up assessment 3 months later. Although revisions to ACTB measures are indicated, results demonstrate adequate levels of reliability and resistance to practice effects for some measures. The ACTB offers viable options for repeated testing of memory, motor planning, behavioral regulation, and attention. Alternative measures of executive functioning are required.

Key Words: *Down syndrome; cognition; memory; hippocampus; cerebellum; reliability; clinical trials; neuropsychological assessment; intellectual disability*

In the past 10 years, landmark investigations have identified several promising pharmacological interventions that have the potential to ameliorate cognitive dysfunction in Down syndrome (DS) and other neurodevelopmental disorders including autism and fragile X syndrome (reviewed in Arnold et al., 2012; Bartesaghi et al., 2015; Fernandez et al., 2007). Given that clinical trials for cognitive and behavioral disorders are increasing in number, there is an immediate need to establish a set of valid and reliable instruments to assess neurocognitive function in individuals with DS and other syndromes that result in intellectual disability (ID). Heller et al. (2006), Edgin, Mason, et al. (2010), and Berry-Kravis et al. (2013) described several key challenges for outcome assessments in individuals with ID. In particular, these researchers acknowledged that very few outcome assessments have been validated specifically for this population, and that sample-specific estimates of test-retest reliability are rare. DS constitutes one of the most frequent ID syndromes and is associated with a unique cognitive phenotype (Pennington, Moon, Edgin, Stedron, & Nadel, 2003). As a precursor to clinical trials in DS, estimates of the psychometric properties of neuropsychological measures as they apply to DS are essential.

Given this need, in April of 2015, a working group convened at the National Institutes of Health (NIH) to focus on this problem. This group determined that test-retest reliability estimates generated in the timeframes required to inform clinical trials were virtually nonexistent in DS (Esbensen et al., this issue). In the time since this working group met, one study has been published describing the psychometrics and usability of neuropsychological measures for clinical trials in this population (d'Ardhuy et al., 2015). Participants in this study were older children and young adults with DS drawn from a large multinational sample. The study authors highlighted some measures that could be useful in the clinical trials context (see Tables 1 and 2 for the

DOI: 10.1352/1944-7558-122.3.215

Table 1
Arizona Cognitive Test Battery Measures From Edgin, Mason, et al. (2010), Task Demands, and Recommended Outcomes Based on Retesting Data From Current Study

Domain/test	Description	Primary ability assessed	Recommended outcomes based on retesting performance and findings of Edgin, Mason, et al. (2010)*
Primary			
Prefrontal			
Modified Dots task (Davidson, Amso, Anderson, & Diamond, 2006)	Presses a button below a cat, shifts to a new rule (pressing across the screen) for a frog, shifts between rules	Inhibitory control, working memory	*No appropriate outcome measures. High levels of floor effects*.
CANTAB IDED	Forced-choice discrimination task with change in relevant dimension	Set-shifting	*No appropriate outcome measures. Inadequate test-retest reliability.*
Hippocampal			
CANTAB Paired Associates Learning	Recall for hidden abstract patterns and associated locations	Spatial associative memory	Stages completed, total errors adjusted are recommended. Utilize the alternate forms provided by CANTAB to limit practice effects.
Virtual computer-generated arena (Thomas, Hsu, Laurence, Nadel, & Jacobs, 2001)	Navigation of a virtual arena (via joystick) to find a fixed hidden target	Spatial memory	*No appropriate outcome measures due to poor test-retest reliability.*
Cerebellar			
Finger Sequencing task (Edgin, 2010b)	Sequences generated by tapping a number of fingers (1,2,3,4) to a lever in succession	Motor sequencing	Use maximum sequence reached.
NEPSY Visuomotor Precision (ages 3–4 years; Korkman, Kirk, & Kemp, 1998)	Follow two tracks with a pen	Visuomotor tracking, hand-eye coordination	Use total score.

(Table 1 continued)

DOI: 10.1352/1944-7558-122.3.215

Table 1
Continued

Domain/test	Description	Primary ability assessed	Recommended outcomes based on retesting performance and findings of Edgin, Mason, et al. (2010)*
CANTAB Simple Reaction Time (SRT)	Participants press a button in response to a box presented on a screen	Motor response time and attention	Use median correct latency, omission errors, commission errors.
Secondary			
KBIT-II Verbal Subscales (verbal knowledge, riddles; Kaufman & Kaufman, 2004)	Points to pictures based on the word or phrase, answers riddles	Verbal comprehension, production	Use subscale raw score totals.
KBIT-II Matrices	Semantic or visuospatial pattern completion	Problem solving	Use subscale raw score total.
CANTAB Spatial Span	Touching of boxes in order changing color on the screen,	Immediate memory for spatial-temporal sequences	Use span.
Scales of Independent Behavior- Revised (Bruininks, Woodcock, Weatherman, & Hill, 1996)	Parent report of everyday skills	Adaptive behavior	Use Standard Score.
Behavioral Outcome Measures			
Behavioral Rating Inventory of Executive Function-School Age (Gioia, Isquith, Guy, & Kenworthy, 2000)	Eight parent-report subscales of everyday executive ability	Domains of prefrontal function, behavioral regulation, and metacognition	All scales can be employed. Practice effects were detected on the working memory T-score.
Nisonger Child Behavior Rating Form-Parent (Aman, Tassé, Rojahn, & Hammer, 1996)	Eight parent-report subscales of adaptive and maladaptive skills	Conduct Problems, Hyperactivity, Anxiety, Sensitivity, Ritualistic, Stereotypic, Social Adaptive Skills, and Compliance	All scales can be used except Self-Injury/ Stereotypic behavior due to poor retest reliability.

Note. CANTAB = Cambridge Neuropsychological Testing Automated Battery. IDED = Intradimensional/extradimensional set-shifting. KBIT-II= the Kaufman Brief Intelligence Test-Second Edition.
*As shown in this table, some measures from the Arizona Cognitive Test Battery (Edgin, Mason, et al., 2010) were noted as having poor psychometric characteristics in the current study. These measures have are listed as having "*No appropriate outcome measures*" in the last column of Table 1.

DOI: 10.1352/1944-7558-122.3.215

Table 2
Measures Recommended Based on d'Ardhuy et al. 2015 and the Edgin et al. Current Study

Domain	d'Ardhuy et al., 2015 Test Name	Edgin et al., current study Test Name
IQ	Leiter- 3	Kaufman Brief Intelligence Test, Second Edition (KBIT-II) • Verbal Knowledge raw score • Matrices raw score • Riddles raw score
Adaptive Behavior	Not tested	Scales of Independent Behavior-Revised (SIB-R)
Hippocampal Memory	RBANS - List learning OMQ-PF (See also Spanò & Edgin, 2016)	CANTAB Paired Associates Learning (using alternate forms)
Working Memory	CANTAB SSP Forward	CANTAB SSP Forward
Attention	Not tested	CANTAB Simple Reaction Time: Commission and Omission of Errors
Executive Function	Behavior Rating Inventory of Executive Function- Preschool (BRIEF-P) (not appropriate in adults)	Behavior Rating Inventory of Executive Function- School Age (BRIEF-School Age)
Language	CELF-P-2 Word classes RBANS Semantic Fluency	KBIT-II • Verbal Knowledge raw score • Riddles raw score
Motor	Not tested	Finger Sequencing NEPSY Visuomotor Precision CANTAB Simple Reaction time
Problem Behaviors	None tested	Nisonger Scales (all scales except Self- Injury and Stereotypic)

Note. RBANS = Repeatable Battery for the Assessment of Neuropsychological Status. OMQ-PF = Observer Memory Questionnaire-Parent Form. CANTAB = Cambridge Neuropsychological Testing Automated Battery. SSP = Spatial Span. CELF-P-2 = Clinical Evaluation of Language Fundamentals–Preschool-2.

Arizona Cognitive Test Battery outcome measures and comparisons with d'Ardhuy et al., 2015). This investigation focused specifically on memory and executive functioning measures, and the tests administered did not cover the full breadth of the neurocognitive phenotype of DS. Specifically, measures of motor planning and attention were omitted.

Other attempts have been made to highlight tests that could be useful for outcome studies in this population, including development of the TESDAD battery (de Sola et al., 2015; de la Torre et al., 2016) and the Arizona Cognitive Test Battery (ACTB) for DS (Edgin et al., 2010), a series of measures that target the primary areas of cognitive dysfunction in DS defined by animal investigations of pharmacological intervention. Our initial report of the ACTB highlighted the validity and preliminary reliability of neuropsychological measures mapped onto the known neurological phenotype of DS. ACTB measures included benchmark tests (e.g., IQ and adaptive behavior) as well as measures targeted to assess hippocampal, prefrontal, and cerebellar function (Table 1). To further establish the adequacy of this battery for use in clinical trials, there remains a need to administer these tests repeatedly in a large sample. Clinical investigations require measures

DOI: 10.1352/1944-7558-122.3.215

demonstrating adequate reliability and stability across multiple sessions spanning months of time (as in d'Ardhuy et al., 2015; usually 4–24 weeks). Therefore, the goal of the present study was to employ a large sample of individuals with DS (N= 54) to determine the reliability and practice effects of tests in the ACTB for this population.

The ACTB was designed to assess neuropsychological domains associated with brain systems that are likely to be targeted in future drug trials offered to individuals with DS. In preclinical animal studies, pharmacological interventions have targeted cognitive endpoints including hippocampal memory functions and prefrontal functions (e.g., attention, executive functioning). Regarding preclinical testing of learning and memory functions, Fernandez et al. (2007) reported that administration of pentylenetetrazole (PTZ, a GABA noncompetitive antagonist) lessened excessive inhibition in the dentate gyrus in a mouse model of DS (Ts65Dn). Furthermore, PTZ eliminated deficits on tests of hippocampal memory function. Other treatments have been developed to counteract the overexpression of specific orthologs of chromosome 21 genes in Ts65Dn mice (e.g., Dyrk1a), with effects including improved learning in the Morris Water Maze (Guedj et al., 2009). Dryk1A inhibitors, including plant extracts that include epigallocatechin-3-gallate, may also show effects on prefrontal function, given links between these deficits and Dryk1a overexpression (Thomazeau et al., 2014). In another report, Salehi et al. (2009) found that the administration of l-threo-3, 4-dihydroxyphenylserine, or xamoterol, a β1-adrenergic receptor partial agonist, normalized deficits in memory and learning in Ts65Dn mice, suggesting that modifications of the adrenergic system may be of additional benefit to improve cognitive outcomes.

Although Salehi et al. (2009) detected changes in the Ts65Dn hippocampus after drug administration, modification of adrenergic neurotransmitters has the potential to affect multiple brain systems, including the prefrontal cortex. Indeed, the cerebellum also shows extensive alteration in humans and mouse models of DS and is likely to be a target of outcome studies in the future (Roper et al., 2006). In this regard, Das et al. (2013) found that treating newborn Ts65Dn with a single treatment of a sonic hedgehog pathway agonist resulted in normalized cerebellar morphology and also improved memory function.

The ACTB is well-suited to the primary targets of the pharmacological interventions in children and adolescents with DS, as it includes nonverbal tests of prefrontal, hippocampal, and cerebellar function in addition to more general assessments of cognitive ability and behavior. The initial report of the ACTB focused primarily on a selection of measures that would be appropriate for the DS population (Edgin, Mason, et al., 2010). In particular, we employed nonverbal (visual) measures of learning, memory, and attention in order to enable participation of individuals with limited expressive language, who constitute a large percentage of this population (Abbeduto, Warren, & Conners, 2007). Further, several ACTB tests include normative data from individuals whose performance falls several standard deviations below average and, therefore, have lower "floor" performance levels than many traditional cognitive assessments. This property is essential for a measure to be sensitive to change when administered to individuals who have ID. ACTB measures also correlated with parent reports of adaptive skills and behavior, demonstrating concurrent validity and relevance of the measures to daily life functioning.

Given these initial findings, the primary goal of the present study was to assess the repeatability of the ACTB measures and the sensitivity of each measure to practice effects. In the context of a clinical trial, measures will be repeated within short time intervals (e.g., a few months) and across several study sites. Therefore, we measured individuals' performance on the ACTB and associated behavioral measures at two sessions approximately three months apart across several testing centers. Also important for the design of clinical trials is an understanding of the characteristics of participants who may be expected to demonstrate greater or less marked change across time. It has been recently noted that some measures may be susceptible to practice effects–and even fatigue effects–in this population (Fernandez & Edgin, 2016). Therefore, a better characterization of practice effects is required, including the participant characteristics that may lead to greater inconsistency across assessments. To answer this question, we examined the participant factors that were associated with large changes from baseline to post-test. Given past findings suggesting the importance of age and level of adaptive and maladaptive behavior characteristics of the individual in relation to performance on ACTB

DOI: 10.1352/1944-7558-122.3.215

measures (Edgin, Mason, et al., 2010), we expected that these factors might play an important role in predicting individual differences in performance variability from the baseline to post-test assessment. The cognitive deficits in DS and the proposed targets of these interventions are similar to those under investigation in a number of syndromes that result in ID, such as fragile X syndrome. Therefore, identifying reliable outcome assessments for individuals with DS may help support clinical trials not only in this group, but also in other syndromes that result in ID.

Methods

Participants

Participants were recruited via local advocacy and parent organizations, advertisement, and a university developmental disabilities research registry. All participants seen for a larger study of cognitive outcomes were asked to return for a second session 3 months after baseline (±4 weeks). Data for this retesting sample were drawn from the Down Syndrome Cognition Project (PIs Sherman and Reeves), a consortium study examining cognitive function in relation to health and genetics in 250 individuals with DS ages 6 to 25 years. Due to resource limitations, only a subset of this sample entered into repeated assessments to determine measure stability and reliability. The final retested sample included 54 individuals from Emory University ($n = 25$), University of Arizona ($n = 15$), Waisman Center at the University of Wisconsin ($n = 6$), Johns Hopkins University ($n = 4$), and Oregon Health Sciences University ($n = 4$). The mean age of the sample was 13.40 years at baseline ($SD = 3.30$, range 7–20 years). Overall, this sample was 56% male, 80% Caucasian, and included 10 families (19%) with income <$50,000 per year. The retested sample did not differ from the larger sample on the Kaufman Brief Intelligence Test (KBIT) IQ or age (M[SD] KBIT IQ retested = 46.55 [7.14], larger sample = 46.42 [10.05], t[246] = –0.09, $p = 0.93$; M(SD) age retested = 13.40 [3.30], larger sample = 13.47 [4.83], t[246] = 0.11, $p = 0.92$). The retested sample represented a range of IQ scores, from 40–69 standard score (SS). Two participants (3.7%) only completed partial assessments at time 2 due to behavioral difficulties. Caregiver-reported outcomes were not completed for 13% ($n = 7$) of the sample due to administration errors and the absence of normative scores in adult age participants for some measures (e.g., Behavior Rating Inventory of Executive Function-School Age). The sample of children with completed caregiver reported outcomes did not differ from the larger sample at baseline (M [SD] KBIT IQ caregiver's report = 46.35 [7.37], larger sample = 46.47 [9.92], t[246] = 0.08, $p = 0.94$; M [SD] age with caregiver reports = 13.17 [2.99], larger sample = 13.52 [4.83], t[246] = 0.47, $p = 0.35$). Table 3 details the full background characteristics of the retested sample across sites.

Exclusion criteria included the presence of Robertsonian translocation, mosaicism, past head injury resulting in a loss of consciousness greater than 5 minutes, other brain trauma (bleeds etc.), lack of oxygen at birth, untreated epilepsy or other seizure disorder, history of chemotherapy, accidental poisoning, or untreated severe hearing or vision loss. All children attended English-speaking

Table 3
Sample Characteristics of Participants With Down Syndrome Across Sites

Site	n	Average age at baseline (years)	Sex Characteristics (% Male)	Ethnicity
Emory University	25	14.48	44%	80% White, Non-Hispanic; 16% Black, Non-Hispanic; 4% Biracial/Multiracial
University of Arizona	15	11.20	67%	73% White, Non-Hispanic; 20% White, Hispanic; 7% Biracial
Waisman Center at the University of Wisconsin	6	12.00	83%	100% White, Non-Hispanic
Johns Hopkins University	4	13.50	75%	100% White, Non-Hispanic
Oregon Health Sciences University	4	13.75	25%	100% White, Non-Hispanic

DOI: 10.1352/1944-7558-122.3.215

schools. There was no restriction based on their level of verbal ability. We verified Trisomy 21 status by collecting karyotypes from participant medical records.

Procedure

All procedures were approved by the institutional review boards of the participating institutions. After informed consent and assent, participants completed each 2-hour testing session in either a laboratory setting or their homes with an examiner experienced in assessing individuals with ID. Testing was monitored and scored for fidelity to ensure that each tester administered the ACTB in a similar manner. Specifically, each tester had to submit videos to the parent site (Arizona) and complete a set of criteria at 80% or greater until the trainer was satisfied with the tester's administration. The criteria included establishing report and proper test administration (i.e., meeting proper basal and ceiling values). The parent site engaged in a 2-day training session with each new tester and then monitored at least three videos at regular intervals (every 6 to 12 months) until the 80% criteria was met. We presented the ACTB in two fixed counterbalanced orders. The Kaufman Brief Intelligence Test, Second Edition (KBIT-II) and the Cambridge Neuropsychological Testing Automated Battery (CANTAB) motor screening were set first in the order to prioritize the administration of IQ and task instructions for the computer at the beginning of the assessment. After that point, the two orders had the CANTAB measures alternating before and after a break to equate fatigue effects across orders. The orders were constructed to interweave desktop and computer administration and to avoid interference effects across tests. At retest, the same fixed order was used, and the measures were administered without the use of alternate forms. Retesting occurred approximately three months after the first assessment (within a window of 4 weeks around this interval). The timing of this interval was chosen because participants would be unlikely to demonstrate substantial improvement related to development alone (i.e., increased maturity) within the span of 3 months. Additionally, this interval is similar to the time frames for treatment in other clinical trials (d'Ardhuy et al., 2015). The testing locations and experimenter remained the same for each participant to the extent possible. During the laboratory assessment, the parents or caregivers were administered the questionnaires for completion. Study data were collected and managed using Research Electronic Data Capture (REDCap) tools hosted across the universities in our network (Harris et al., 2009). REDCap is a secure, web-based application designed to support data capture for research studies. This format enabled validated data entry and data sharing across sites.

Measures

The ACTB measures and justification for the use of each one are described in full detail in Edgin, Mason, et al. (2010) and are shown in Table 1. These measures fall into two categories. The first category comprised target measures to serve as primary outcomes in a clinical trial. These measures assessed brain functions that would be specifically targeted by drugs under development, such as hippocampal memory (Thomas, Hsu, Laurence, Nadel, & Jacobs, 2001) or prefrontal and cerebellar functions (Davidson, Amso, Anderson, & Diamond, 2006; Korkman, Kirk, & Kemp, 1998). The second category of measures was comprised of broader, potential secondary outcomes in a trial. These measures included IQ (Kaufman & Kaufman, 2004), adaptive behavior (Bruininks, Woodcock, Weatherman, & Hill, 1996), and maladaptive behavior scales (Aman, Tasse, Rojahn, & Hammer, 1996; Gioia, Isquith, Guy, & Kenworthy, 2000). As described in Edgin, Mason, et al. (2010), each neuropsychological measure was chosen based on data demonstrating links to brain function in the target region. In this investigation, measures were administered without the use of alternate forms because different forms were only available for a subset of the measures.

Primary measures.

Hippocampal (associative) memory.

CANTAB paired-associates learning (PAL). For this task, the participant was asked to learn associations between abstract visual patterns and hiding locations on a computer screen. Participants were first presented with six boxes, which opened up one at a time. A shape appeared in one of the boxes, and the participant was asked to remember where the shape was hidden. After the presentation, the shape appeared in the middle of the screen, and the examiner asked the participant to touch the box where the shape was hidden. Thus, this task required the subject to generate the spatial location associated with the stimulus. The task increased in difficulty from one to eight shapes to be remembered.

DOI: 10.1352/1944-7558-122.3.215

Based on functional neuroimaging data in healthy adults and patients with mild cognitive impairment, the hippocampus is activated during both encoding and retrieval on this task (de Rover et al., 2011). CANTAB PAL has been used as a benchmark measure for memory deficits in several patient groups, including individuals with DS, demonstrating low levels of noncompletion, adequate test-retest reliability, and sensitivity to detect differences between individuals with DS and control participants without the confounding influence of deficits in language (Edgin, Mason, et al., 2010; Edgin, Spanò, Kawa, & Nadel, 2014; Pennington et al., 2003; Visu-Petra, Benga, & Miclea, 2007). Further, performance on this task has been shown to correlate with parent-reported memory skills and ERP assessments (Spanò & Edgin, 2016; Van Hoogmoed, Nadel, Spanò, & Edgin, 2016).

Virtual computer-generated arena (Thomas et al., 2001). This task is an assessment of hippocampal function based on the Morris Water Maze paradigm from the animal literature (Morris, 1984; Thomas, Hsu, Laurence, Nadel, & Jacobs, 2001). Across several trials, participants learn to find a target hidden on the floor of a computer-generated arena, presented from a first-person perspective. The fixed target position can be learned by relating its position to landmarks (distal cues) surrounding the arena. This task has been successfully used in individuals with DS and other developmental disabilities in past investigations (Edgin & Pennington, 2005; Pennington et al., 2003).

Prefrontal tasks.

Modified dots task (Davidson et al., 2006). This task is a measure of inhibitory control and working memory for participants aged 4 years to adulthood consisting of three phases. In the first phase, participants learn the rule associated with the cat stimuli (the congruent location rule) by pressing a button located directly below an animated depiction of a cat on a computer screen. In the second phase, participants see frogs presented on the left or right-hand side and must touch the button located on the other side of the computer screen from the frog (the incongruent location rule). In the final phase, participants are asked to respond to trials in which the rules are alternated randomly. Each of the first two phases begins with practice trials. Scores are calculated based on the percentage of correct responses for each phase of the test (max score = 100%). This task requires behavioral inhibition to override the prepotent tendency to respond on the same side as the visual stimulus during the incongruent rule trials.

CANTAB intradimensional/extradimensional set-shifting (IDED). This task measures the participant's ability to learn a baseline rule and then to disengage from that response set to learn another rule. Participants are presented with two colored shapes during multiple trials. The shapes appear in four boxes. In the task, the participant must learn which shape is "correct" through simple trial-and-error; correct is designated through a "correct" label and the screen turning green. Once the rule is consistently recognized after several trials, the "correct" shape rule is reversed. The participant must now recognize this rule shift and adapt their choices to the new "correct" shape. In later trials, a second shape is transposed onto each shape, adding a second dimension that the participant must then take into consideration when determining which shape is "correct." Temporal lobe patients and those with Alzheimer's disease show relatively unaffected performance on the IDED task. However, frontal patients are impaired (Strauss, Sherman, & Spreen, 2006).

Cerebellar tasks.

Finger sequencing task (Edgin, 2010b). A computerized version of finger sequencing was developed by the Edgin lab and based on a sequencing task in the NEPSY battery (Korkman et al., 1998). The computerized version involves tapping a lever with one, two, three, or four fingers in sequence in the same manner that one would tap fingers to the thumb in the original paradigm. Both dominant and nondominant hands are tested. There is a 10-second practice period, followed by a 30-second test period for each trial. After each set is completed, the participants are rewarded by viewing a dog moving on the screen nearer to a goal. The computerized version records the number of correct sequences, the total taps, and the standard deviation between taps for each set. The test-retest reliability in a sample of 32 undergraduate students tested across a 6-week interval was excellent for the computerized version (intraclass correlation [ICC] for total taps generated = 0.91, ICC for correct sequences = 0.87, and ICC for tap standard deviation = 0.79; Edgin, 2010a).

NEPSY visuomotor precision (ages 3–4; Korkman, Kirk, & Kemp, 1998). This task involves the subject following a series of two tracks, a train and car track, from start to finish using a pen. Participants must keep the pen in contact with

DOI: 10.1352/1944-7558-122.3.215

the paper at all times and maintain their lines within the tracks. The errors (lines exiting the track) and completion time are considered together to generate a total score.

CANTAB simple reaction time (SRT). In the SRT task, participants press a button when a stimulus (a white box) appears on the computer screen. The onset timing of the stimulus varies between trials. Slowing of motor response time is typical with cerebellar dysfunction, and studies have reported slowed reaction times in DS in comparison to mental-age matched controls and those with other developmental disabilities, such as autism (Frith & Frith, 1974). In addition to the SRT measure of simple psychomotor speed, this measure generates variables associated with attention, including errors of omission and commission.

Secondary measures.

Kaufman Brief Intelligence Test, Second Edition (KBIT-II; Kaufman & Kaufman, 2004). The KBIT-II is a brief, individually administered measure of both verbal and nonverbal intelligence appropriate for individuals from 4 to 90 years old (Kaufman & Kaufman, 2004). The test consists of three subtests: Verbal Knowledge, Matrices, and Riddles. Verbal knowledge requires the participant point to the correct image after being given a verbal prompt. For the Matrices test, the participant must select the correct image to complete a pattern. In the Riddles test, there are two sections. First, the participant must select the correct image after being given a riddle style prompt and, in the second, the participant must verbally answer the riddle with one word after a prompt. Standard scores for the KBIT-II have a mean equal to 100, standard deviation of 15.

CANTAB Spatial Span. The CANTAB Spatial Span is a test of immediate spatial working memory. Participants copy a sequence of blocks that are displayed one at a time. The score is determined by the length of the longest sequence successfully recalled by the participant (span length; max. score = 9). A well-replicated finding in individuals with DS is a deficit in verbal short-term memory, with strength in spatial short-term memory tasks (Edgin, Pennington, & Mervis, 2010). This task was found to have acceptable psychometric characteristics, including adequate test-retest reliability, in d'Ardhuy et al., 2015.

Scales of Independent Behavior-Revised (SIB-R; Bruininks et al., 1996). The SIB-R is a caregiver-completed checklist-style rating scale designed to assess adaptive functioning and everyday skills. The SIB-R measures Motor, Social and Communication, Personal Living, and Community Living Skills. The measure spans a wide-range of ages, from infancy to adulthood.

Behavioral outcome measures.

Behavior Rating Inventory of Executive Function-School Age (BRIEF; Gioia et al., 2000). The BRIEF is a widely used caregiver questionnaire of everyday skills reflective of abilities in the executive domain. It generates a range of scales, including scales specific to working memory and inhibitory control. This measure has been used in several populations with developmental disabilities, including individuals with autism and frontal lesions. The test-retest reliability has been found to be adequate to high for the parent form in DS and typical groups ($r = .80–.89$ for most scales; d'Ardhuy et al., 2015; Strauss et al., 2006).

Nisonger Child Behavior Rating Form-Parent (CBRF; Aman et al., 1996). The Nisonger CBRF was developed to measure behavior problems known to occur in individuals with intellectual disabilities, including problems with hyperactivity and attention, social problems, and stereotypic behavior. The Nisonger CBRF also correlated highly with analogous subscales from the Aberrant Behavior Checklist (Aman et al., 1996).

Statistical Analysis

All analyses were conducted with SPSS 23.0. The distribution of measures was first tested for normality using the Shapiro-Wilk test (Table 4). Tests that were normally distributed ($p > 0.01$ Shapiro-Wilk) included the NEPSY visuomotor total score, KBIT-II raw scores, SIB-R standard scores, and most parent-reported behavioral outcomes on the BRIEF and Nisonger Scales. All other tests had statistically significant Shapiro-Wilk tests ($p < 0.01$). All test-retest data were analyzed with intraclass correlation to match previous investigations, and non-normal variables were also analyzed with Spearman's rho. For the tests with adequate reliability, paired sample *t*-tests were used to test for differences between baseline and post-test performance; Wilcoxson signed-rank test was employed for non-normal outcomes. Floor effects were measured at baseline and reflect the values of the children unable to complete the task or those who completed the task with the lowest possible score. Spearman's rho correlations were conducted to relate change from baseline to age and behavioral assessments. To account for multiple comparisons, we adopt-

DOI: 10.1352/1944-7558-122.3.215

Table 4
Test-Retest Reliability Estimates for ACTB Primary, Secondary, and Behavioral Outcome Measures Across 3 Months in Children With DS; Practice Effects Are Indicated by a Statistically Significant Change in Scores From the Baseline Assessment

Measure	Correlation *n*	Shapiro Wilk *p**	Retest Intraclass Correlation (ICC)	Retest Spearman's rho	*n* Floor	*p* Δ Baseline to 3 Months (Paired t/Wilcoxon)**
Primary Outcomes						
Hippocampal						
CANTAB PAL First Trials Memory Score	52	<0.01	0.69	0.65	4/54	0.009
CANTAB PAL Stages Completed	52	<0.01	0.72	0.69	4/54	0.03
CANTAB PAL Total Errors Adjusted	52	<0.01	0.75	0.75	0/54	0.02
Computer-Generated Arena Total Targets	50	<0.01	0.43	0.41	6/54	0.70
Computer-Generated Mean Path Length	50	0.46	0.33	0.39	6/54	0.30
Prefrontal						
CANTAB IDED pre-ED Errors	50	<0.01	0.48	0.53	4/54	0.10
CANTAB IDED Stages Completed	50	<0.01	0.36	0.48	9/54	0.12
Modified Dots Task Inhib. Control Phase Percent Correct	48	<0.01	0.59	0.48	24/54	0.02
Modified Dots Task Combined Phase Percent Correct	48	<0.01	0.69	0.50	29/54	0.32
Cerebellar						
CANTAB SRT Median Corr. Latency (ms)	49	<0.01	0.80	0.77	5/54	0.61
CANTAB SRT % Commission Errors	49	<0.01	0.78	0.68	5/54	0.63
CANTAB SRT % Omission Errors	49	<0.01	0.66	0.67	5/54	1.0
NEPSY Visuomotor Precision Total Errors	51	<0.01	0.89	0.88	3/54	0.04
NEPSY Visuomotor Precision Total Score	51	0.06	0.78	0.69	3/54	0.18

(Table 4 continued)

DOI: 10.1352/1944-7558-122.3.215

Table 4
Continued

Measure	Correlation *n*	Shapiro Wilk *p**	Retest Intraclass Correlation (ICC)	Retest Spearman's rho	*n* Floor	*p* Δ Baseline to 3 Months (Paired t/Wilcoxon)**
Finger Sequencing Maximum Sequence Reached	46	<0.01	0.85	0.71	8/54	0.37
Secondary Outcomes						
KBIT-II Verbal Standard Score	51	< 0.01	0.87	0.87	17/54	0.29
KBIT-II Nonverbal Standard Score	51	< 0.01	0.66	0.70	15/54	0.76
KBIT-II Standard Score	51	<0.01	0.81	0.85	20/54	0.46
KBIT-II Verbal Knowledge Raw Score	51	0.30	0.86	0.86	3/54	0.06
KBIT-II Riddles Raw Score	51	0.25	0.76	0.74	5/54	0.71
KBIT-II Matrices Raw Score	51	0.48	0.72	0.76	3/54	0.20
CANTAB Spatial Span	47	<0.01	0.78	0.72	17/54	0.05
CANTAB Spatial Span Errors	47	<0.01	0.42	0.54	8/54	0.28
SIB-R Standard Score	48	0.20	0.79	0.84	6/54	0.59
Behavioral Outcome Measures (n = 47)[a]						
BRIEF Emotional Control T-score	47	0.08	0.74	0.72	0/47	0.24
BRIEF Inhibit T-score	47	0.09	0.79	0.77	0/47	0.32
BRIEF Initiate T-score	47	0.23	0.78	0.78	0/47	0.24
BRIEF Monitor T-score	47	0.27	0.62	0.62	0/47	0.64
BRIEF Organization of Materials T-score	47	0.10	0.81	0.84	0/47	0.76
BRIEF Planning Organizing T-score	47	0.16	0.75	0.74	0/47	0.57
BRIEF Shift T-score	47	0.22	0.86	0.85	0/47	0.18
BRIEF Working Memory T-score	47	0.91	0.74	0.76	0/47	0.004

(Table 4 continued)

DOI: 10.1352/1944-7558-122.3.215

Table 4
Continued

Measure	Correlation *n*	Shapiro Wilk *p**	Retest Intraclass Correlation (ICC)	Retest Spearman's rho	*n* Floor	*p* Δ Baseline to 3 Months (Paired t/Wilcoxon)**
BRIEF Behavioral Regulation Index T-score	47	0.61	0.88	0.90	0/47	0.10
BRIEF Metacognition Index T-score	47	0.61	0.81	0.83	0/47	0.10
BRIEF GEC T-score	47	0.66	0.84	0.87	0/47	0.09
Nisonger Conduct Problems Raw Score	47	<0.01	0.76	0.78	0/47	0.46
Nisonger Hyperactive Raw Score	47	0.15	0.73	0.82	0/47	0.64
Nisonger Insecure Anxious Raw Score	47	<0.01	0.76	0.74	0/47	0.89
Nisonger Overly Sensitive Raw Score	47	<0.01	0.80	0.82	0/47	0.62
Nisonger Self-Isolated Ritualistic Raw Score	47	<0.01	0.71	0.71	0/47	0.42
Nisonger Self-Injury Stereotypic Raw Score	47	<0.01	0.24	0.38	0/47	0.33
Nisonger Adaptive Social Raw Score	47	0.21	0.73	0.71	0/47	0.60
Nisonger Compliant Calm Raw Score	47	0.09	0.66	0.69	0/47	0.80

Note. CANTAB = Cambridge Neuropsychological Testing Automated Battery. Paired Associates Learning. IDED = Intradimensional/Extradimensional Set-Shifting. ED = extradimensional. Inhib. = inhibitory control. Corr. = Correct. SRT = Simple Reaction Time. KBIT-II= the Kaufman Brief Intelligence Test-Second Edition. SIB-R=Scales of Independent Behavior-Revised. BRIEF = Behavior Rating Inventory of Executive Function-School Age. GEC = General Executive Composite.
[a]Caregiver reports were not returned for 7 participants. Floor effects were measured at baseline and indicate measure noncompletion as well as achieving the measure's lowest score.
*Statistically significant values for the Shapiro-Wilk test indicate the task distribution is non-normal.
**Normally distributed scores were analyzed with paired samples t-tests and non-normal scores were analyzed with the Wilcoxon signed-rank test.

ed a more conservative alpha level ($p \leq 0.01$) to determine significance.

Results

Test-Retest Reliability

Table 4 details the sample size, normality tests, and levels of test-retest reliability for the primary neuropsychological outcome measures, secondary measures, and parent/experimenter reports of behavior contained in the ACTB. Test-retest reliability intraclass and Spearman's correlations were evaluated using the following criteria: < 0.40 (poor), 0.40–0.59 (fair), 0.60–0.75 (good), and > 0.75 (very good) to match previous studies assessing the psychometric strength of cognitive

DOI: 10.1352/1944-7558-122.3.215

measures in DS (d'Ardhuy et al., 2015). Given that the Spearman's rho and ICC values were very similar, we detail the ICC values in the text, but both are available in Table 4 for reference. Of the primary outcomes, five out of six variables in the cerebellar category demonstrated very good test-retest reliability (ICC > 0.75) and the sixth variable was good (ICC = 0.66, CANTAB SRT omission errors).

Results for the hippocampus-dependent and prefrontal measures were mixed. The CANTAB PAL showed consistency in the strength of the correlations, with the highest reported value at ICC = 0.75 (i.e., PAL total errors adjusted). However, the test-retest correlations generated from the c-g arena were unacceptable, with fair or poor reliability (ICC = 0.43 for total targets and ICC = 0.33 for mean path length).

On the CANTAB IDED, errors prior to the extradimensional (ED) stage showed fair reliability (ICC = 0.48), but the stages completed measure was inadequate (ICC = 0.36). The modified dots task showed adequate test-retest correlations at the stage in which the rules alternate (combined phase % correct ICC = 0.69). The inhibitory control phase only showed fair reliability (ICC = 0.59), and this measure had the highest level of floor performance of any test in the ACTB (54% on the combined phase correct with 29/54 children showing the floor performance). Spearman's correlations for this measure were only in the fair range. Therefore, none of the measures in the "prefrontal" category of the ACTB were adequate.

The secondary measures included in the ACTB are tasks that measure global outcomes (e.g., IQ) or serve as control measures. The KBIT-II raw scores, SIB-R SS, and CANTAB Spatial Span score all demonstrated good or very good levels of reliability (ICC > 0.70 for all). In contrast, the CANTAB Spatial Span errors measure did not show adequate levels of reliability (fair at ICC = 0.42).

Parent report measures from the ACTB that are evaluated here (i.e., BRIEF report of Executive Function and the Nisonger CBRF-P, $n = 47$) showed good levels of reliability. The overall BRIEF GEC (General Executive Composite, ICC = 0.84), behavioral regulation index (ICC = 0.88), and metacognitive index (ICC = 0.81) displayed very good reliability, with all individual BRIEF scales showing good or very good retest correlations. The weakest scales were working memory (ICC = 0.74), emotional control (ICC = 0.74), and monitoring T-scores (ICC = 0.62), but correlations still fell into the adequate range. The Nisonger CBRF demonstrated primarily good or very good reliability, showing acceptable retest correlations at good or very good levels for most scales, but unacceptable reliability for the Self-Injury/Stereotypic Behavior raw score (ICC = 0.24).

Given the range of function in DS, it is also of interest to better understand how participants in the lower range of IQ may perform on these tests. Thus, we compared the test-retest reliability coefficients in individuals with IQ scores lower and higher than the mean score (set at 47 SS) on measures with strong test-retest reliability in the overall sample. Specifically, we compared test-retest reliability on reliable ACTB outcomes in individuals that attained scores above ($n = 22$) and below this IQ mean ($n = 30$) on key outcome measures. Of the 19 measures tested, five test-retest correlations fell below the cut-off for good reliability and into the "fair" range (retest rho 0.40–0.60). These included the test-retest correlation for the KBIT-II Matrices in the low IQ group (rho = 0.53) and the CANTAB SRT omission errors in the low IQ group (rho = 0.58). In only two of the five instances was there fair reliability in the group with the lowest IQ; the high IQ group had reliability scores below the cut-off on the maximum sequencing score on the Finger Tapping task (rho = 0.58), the Nisonger Adaptive Social subscale (rho = 0.43), and the Nisonger Insecure/Anxious subscale (rho = 0.52).

Practice Effects and Correlates of Change Over Time

Table 3 shows the statistical significance of the changes in scores from baseline to retest for each measure, and Figure 1 provides a graphical representation of the range of these changes. At $p \leq 0.01$, we observed statistically significant practice effects for the CANTAB PAL first trials memory score ($p = 0.009$) only. The other individually administered assessments did not show changes that reached statistical significance at $p \leq 0.01$. Most parent reports of behavior remained stable across time, with the exception of a significant difference on the BRIEF Working Memory T-score ($p = 0.004$). These practice effects all represented overall positive gains in performance from baseline. Figure 1 shows the range of performance gains and losses on key measures, as described in Table 2, including the median gain, *p*-

DOI: 10.1352/1944-7558-122.3.215

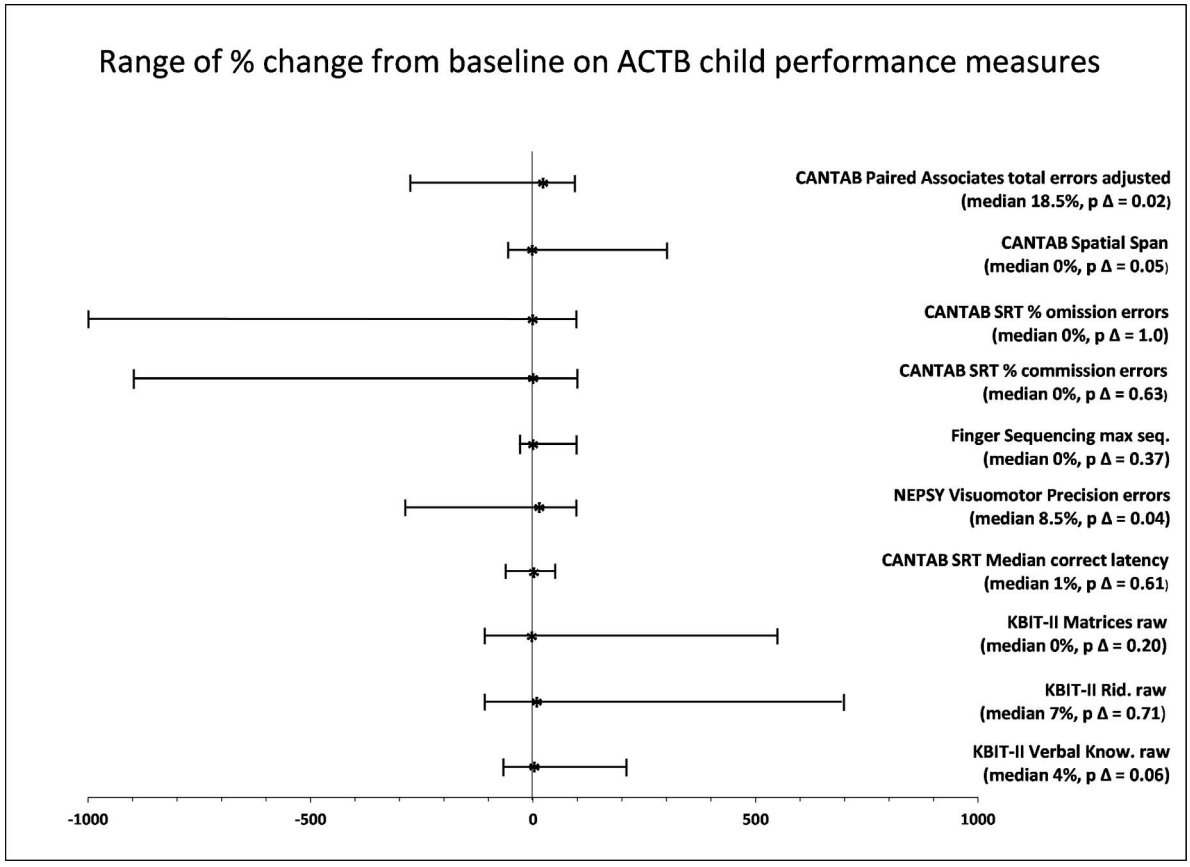

Figure 1. Median percent change, *p* values for change, and range of change across recommended measures of child performance from Table 2. ACTB = Arizona Cognitive Test Battery. CANTAB = Cambridge Neuropsychological Testing Automated Battery. SRT = Simple Reaction Time. KBIT-II = Kaufman Brief Intelligence Test-Second Edition. Rid. = Riddles. Know. = knowledge.

value for percent change, and the full range of scores. As can be seen from Figure 1, the median gain across test sessions was often at 0 or tended toward score increases (the highest median increase was 18.5% on the CANTAB PAL). However, the range of score changes was extensive, with a small number of children showing large fatigue effects, or losses on the CANTAB PAL, the NEPSY Visuomotor Precision test, and attention variables on the CANTAB Simple Reaction Time.

An individual's tendency to show large increases or decreases from baseline to post-test could be hypothesized to relate to his or her level of certain challenging behaviors (such as those measured on the Nisonger scales), or to other factors such as executive control abilities, IQ, or age. Therefore, we assessed the correlations between these factors and the amount of change from baseline on the tests that showed extensive fatigue-related percentage change in Figure 1. The mean proportion of change on each these measures (CANTAB PAL, NEPSY Visuomotor Precision, and SRT commission and omission variables) was not significantly correlated with any of these factors (Spearman's rho was nonsignificant at $p > 0.01$ for all). However, there was a marginal, but not statistically significant, correlation with age (rho = 0.28, $p = 0.04$), with negative changes in scores relating to younger ages. In other words, younger children were more likely (at the trend level only) to show fatigue-related change. However, there was no relation between IQ and change from baseline (rho = –0.06, $p = 0.65$).

Discussion

In the current report, we extended validation of the ACTB. Data cataloging the psychometric properties, including stability over intervals at this

DOI: 10.1352/1944-7558-122.3.215

timescale, are rarely collected in DS and other syndromes that result in ID, but are essential for evaluating the efficacy of cognitive and behavioral outcomes in these groups. In fact, we are aware of only one other published investigation to date that details retest reliability data on neuropsychological measures in a large group of individuals diagnosed with DS (d'Ardhuy et al., 2015). The current investigation lends support for the use of a subset of these measures in clinical trials of cognition in DS and possibly other intellectual disabilities in several ways. Principally, consistent with the results of Edgin et al. (2010) from an independent sample, many of these measures displayed adequate test-retest reliability and minimal practice effects across higher and lower levels of IQ. Although the test-retest results suggest promise for the use of several ACTB tasks administered individually and parent-report measures in clinical trials, they also indicate where refinements of the current battery are needed (see Table 1). These measures from the ACTB, in addition to those identified in d'Ardhuy et al. (2015), begin to form the basis for a toolkit that could be useful in clinical outcome studies of DS (see Table 2 for measures recommended across both studies). Our recommendations for primary and secondary measures and for parent- and experimenter-reported behavioral outcomes follow.

Primary Outcomes

Table 2 lists the measures that we recommend from the ACTB and how they are modified from the original battery (Table 1). In particular, the CANTAB PAL showed strong reliability with some practice effects, and measures of attention and motor planning were usable, including finger sequencing, the NEPSY Visuomotor Precision task, and the CANTAB SRT test, which offers a processing speed task as well as measures of attention (omission errors) and hyperactivity (commission errors). The c-g arena must be replaced with a more valid and reliable measure of contextual memory function, and none of the measures placed in the executive category were reliable.

Some secondary measures could also be employed as primary measures in a clinical trial based on the targets of that trial. For instance, the KBIT-II includes some useful measures that map onto important aspects of the DS phenotype as well, even though they are measured on an IQ scale. Although the KBIT-II is a brief IQ scale with a high standard score floor (SS = 40), the raw scores on the scale had low floor effects and strong retest reliability. In particular, the Matrices measure could be used to measure frontal function and planning, given that none of the executive function measures in the original ACTB showed adequate psychometrics to be used in a clinical trial. Although nuanced methods of language assessment are promising and currently under validation (Berry-Kravis et al., 2013), our data show that the KBIT Riddles test could offer another short and reliable measure of language competency with negligible practice or floor effects. Consistent with d'Ardhuy et al. (2015), we found the CANTAB Spatial Span forward measure to be reliable, with a trend toward a practice effect. The CANTAB Spatial Span has been employed in clinical trials of ADHD (Bedard, Martinussen, Ickowicz, & Tannock, 2004), and the forward measure of this test may allow for a good measure of frontal function. Given that this task has one of the highest floor effects on the forward version (31%), we caution against the use of the CANTAB Spatial backward span task, which was also noted to be inappropriate in d'Ardhuy et al. (2015).

The use of alternate forms when available (e.g., CANTAB measures, Spatial Span) could help further reduce practice effects on some of these measures. Indeed, d'Ardhuy et al. (2015) found it possible to administer a verbal list learning measure of memory with alternate forms (the Repeatable Battery for the Assessment of Neuropsychological Status [RBANS], Table 2) with no statistically significant practice effect. Although most measures did not have significant mean changes, the range of change was large for some measures and demonstrates a fatigue effect in a small number of children (Figure 1). These effects were related to age, and suggest that batteries for younger children need to be carefully devised to avoid fatigued responses; shorter batteries may be required in the youngest children.

Of interest is the consistent pattern of strong test-retest reliability on measures of latency and reaction time. Taking into account past research on extensive reaction time variability in children (Zabel, von Thomsen, Cole, Martin, & Mahone, 2009), this finding was unexpected, but suggests that measures that are quite simple in their demands, such as reaction time, may allow for

DOI: 10.1352/1944-7558-122.3.215

the consistent measurement of cognitive outcome in ID syndromes.

Secondary Measures

Secondary IQ and adaptive measures also proved to be quite stable. Reliability on the CANTAB Spatial Span task was acceptable, although floor effects were higher than other measures on the ACTB. To decrease floor effects, the CANTAB Spatial Span task could be replaced by a table-top or more engaging version of the task (CORSI blocks), which has been found to show low floor effects in past investigations, at least in adults (Edgin, Pennington, & Mervis, 2010).

A limitation of the ACTB is the use of abbreviated versions of both adaptive behavior and IQ assessments, which was necessary because of time constraints in this study. More comprehensive measures could be included in trials with a longer treatment interval during which more substantial global gains might be expected, or in cases where a more in-depth assessment is required. The current data suggest that these measures can be quite stable and resistant to large practice effects, even when there is verbal content and no alternate form. Although the focus of our development of the ACTB is primary outcome measures, more validation is required for secondary outcomes. d'Ardhuy et al. (2015) validated a full IQ assessment, finding that the Leiter 3 could be used effectively and demonstrated good psychometric properties when tested in accordance with the typical study design of a clinical trial.

Parent- and Experimenter-Rated Behavioral Outcomes

The parent report measures of adaptive and maladaptive behavior and executive function were usable and stable on the whole, with some single scales showing poor reliability or large practice effects (BRIEF Working Memory T-score and Nisonger Self-Injury/Stereotypic). These findings are consistent with work showing acceptable reliability on scales of memory and behavior for parent-reported outcomes in DS (Ji, Capone, & Kaufman, 2011; Pritchard, Kalback, McCurdy, & Capone, 2015). However, our study had a fairly high rate of measure noncompletion ($n = 7$, 13%), so the use of these measures may require significant follow-up efforts or incentives to ensure that these data are obtained without error. Further, parent-reported outcomes are more susceptible to placebo effects (Heller et al., 2010), so it is unlikely that they will be designated a sole primary outcome in a clinical investigation. Parent measures could nevertheless provide useful information as a secondary measure, and may improve ecological validity as these instruments measure constructs that are highly relevant to daily life functioning and well-being. One important finding of the current investigation is that most of the measures did not display significant practice effects, and the median percentage change was quite low for these tasks. Heller et al. (2010) discussed the need to closely examine individual differences in the response to drug treatment. Although it seems feasible that the extent of a placebo or practice effect may be highly influenced by individual differences, we found few correlations between background factors and the size of these effects that could help determine which participants may succumb to large gains or losses from baseline (beyond a trend-level correlation with age).

Some study limitations should be noted. First, although the study represents one of the largest samples to undergo retesting to measure reliability and practice effects in this population, the sample size remains modest. Notably, the retested sample did not differ statistically from the larger cohort in terms of IQ or age, suggesting that the re-testing results can be extended to the larger DS population. Future investigations should examine these outcomes in larger samples with enough power to stratify participants across important variables (e.g., cross-site differences). Along these lines, another limitation is the wide age range of our participants. Future investigations are needed to examine the psychometric properties of outcome measures in targeted age ranges. In particular, little work has been conducted with the validation of outcome assessments for preschool or young school-aged children. The adult age-range also has a great need for sensitive outcome assessments and, in particular, for the validation of tests that may be early indicators of cognitive decline.

Conclusions

As shown in Tables 2 and 4, the current results support the use of a selected set of ACTB measures for clinical trials in older children and young adults with DS. Particularly promising were measures of motor planning and attention and parent-reported scales of behavior. Regarding

DOI: 10.1352/1944-7558-122.3.215

memory outcomes, the CANTAB Spatial Span and PAL could be implemented with acceptable reliability, but alternate forms should be utilized to limit floor effects. No adequate individualized assessments of executive control were identified. Together, the findings from this study, d'Ardhuy et al. (2015), and the NIH working group (Esbensen et al., this issue) can help to devise a gold standard protocol for future clinical trials in older children and young adults with DS. Although some researchers have also emphasized the importance of targeting interventions to younger children with DS (Edgin, Clark, Massand, & Karmiloff-Smith, 2015), to our knowledge, no investigations have been conducted to assess the psychometric characteristics of cognitive outcome measures specifically for younger children with DS. Given the momentum of developing clinical trials for DS and other conditions, such as fragile X syndrome, more work is needed to develop tests of neuropsychological function that may be administered across a wider age range in people with ID in general. It will be important to continue documenting the factors relating to fatigue effects in this population and to minimize the length of outcome batteries to the extent possible. Further, as demonstrated by the current investigation, there is a great need to identify and validate measures of frontal lobe and executive function that would be suitable for individuals with DS and other intellectual disability groups.

References

Abbeduto, L., Warren, S. F., & Conners, F. A. (2007). Language development in Down syndrome: From the prelinguistic period to the acquisition of literacy. *Mental Retardation and Developmental Disabilities Research Reviews, 13*(3), 247–261. http://dx.doi.org/10.1002/mrdd.20158

Aman, M. G., Tassé, M. J., Rojahn, J., & Hammer, D. (1996). The Nisonger CBRF: A child behavior rating form for children with developmental disabilities. *Research in Developmental Disabilities*, *17*(1), 41–57. http://dx.doi.org/10.1016/0891-4222(95)00039-9

Arnold, L. E., Aman, M. G., Hollway, J., Hurt, E., Bates, B., Li, X., ... Ramadan, Y. (2012). Placebo-controlled pilot trial of mecamylamine for treatment of autism spectrum disorders. *Journal of Child and Adolescent Psychopharmacology, 22*(3), 198–205. http://dx.doi.org/10.1089/cap.2011.0056

Bartesaghi, R., Haydar, T. F., Delabar, J. M., Dierssen, M., Martínez-Cué, C., & Bianchi, D. W. (2015). New perspectives for the rescue of cognitive disability in Down syndrome. *The Journal of Neuroscience, 35*(41), 13843–13852. http://dx.doi.org/10.1523/JNEUROSCI.2775-15.2015

Bedard, A. C., Martinussen, R., Ickowicz, A., & Tannock, R. (2004). Methylphenidate improves visual-spatial memory in children with attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, *43*(3), 260–268. http://dx.doi.org/10.1097/00004583-200403000-00006

Berry-Kravis, E., Doll, E. D., Sterling, A., Kover, S. T., Schroeder, S. M., Mathur, S., & Abbeduto, L. (2013). Development of an expressive language sampling procedure in fragile X syndrome: A pilot study. *Journal of Developmental and Behavioral Pediatrics*, *34*(4), 245–251. http://dx.doi.org/10.1097/DBP.0b013e31828742fc

Bruininks, R. K., Woodcock, R. W., Weatherman, R. F., & Hill, B. K. (1996). *Scales of independent behavior- revised*. Itasca, IL: Riverside Publishing Company.

d'Ardhuy, X. L., Edgin, J. O., Bouis, C., de Sola, S., Goeldner, C., Kishnani, P., ... Spiridigliozzi, G. (2015). Assessment of cognitive scales to examine memory, executive function and language in individuals with Down syndrome: Implications of a 6-month observational study. *Frontiers in Behavioral Neuroscience*, *9*, 300. http://dx.doi.org/10.3389/fnbeh.2015.00300

Das, I., Park, J. M., Shin, J. H., Jeon, S. K., Lorenzi, H., Linden, D. J., ... Reeves, R. H. (2013). Hedgehog agonist therapy corrects structural and cognitive deficits in a Down syndrome mouse model. *Science Translational Medicine*, *5*(201), 201ra120. http://dx.doi.org/10.1126/scitranslmed.3005983

Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia*, *44*(11), 2037–2078. http://dx.doi.org/10.1016/j.neuropsychologia.2006.02.006

DOI: 10.1352/1944-7558-122.3.215

de Rover, M., Pironti, V. A., McCabe, J. A., Acosta-Cabronero, J., Arana, F. S., Morein-Zamir, S., ...& Sahakian, B. J. (2011). Hippocampal dysfunction in patients with mild cognitive impairment: a functional neuroimaging study of a visuospatial paired associates learning task. *Neuropsychologia*, *49*(7), 2060–2070. http://dx.doi.org/10.1016/j.neuropsychologia.2011.03.037

de la Torre, R., de Sola, S., Hernandez, G., Farré, M., Pujol, J., Rodriguez, J., ... Xicota, L. (2016). Safety and efficacy of cognitive training plus epigallocatechin-3-gallate in young adults with Down's syndrome (TESDAD): A double-blind, randomised, placebo-controlled, phase 2 trial. *The Lancet Neurology*, *15*(8), 801–810.

de Sola, S., de la Torre, R., Sánchez-Benavides, G., Benejam, B., Cuenca-Royo, A., del Hoyo, L., ... Hernandez, G. (2015). A new cognitive evaluation battery for Down syndrome and its relevance for clinical trials. *Frontiers in Psychology*, *6*, 708. http://dx.doi.org/10.3389/fpsyg.2015.00708

Edgin, J. O. (2010a). [Finger sequencing task]. *Unpublished raw data*, University of Arizona, Tucson, AZ.

Edgin, J. O. (2010b). [Finger sequencing task]. *Unpublished paradigm*, University of Arizona, Tucson, AZ.

Edgin, J. O., Clark, C. A., Massand, E., & Karmiloff-Smith, A. (2015). Building an adaptive brain across development: targets for neurorehabilitation must begin in infancy. *Frontiers in Behavioral Neuroscience*, *9*, 232. http://dx.doi.org/10.3389/fnbeh.2015.00232

Edgin, J. O., Mason, G. M., Allman, M. J., Capone, G. T., DeLeon, I., Maslen, C., ... Nadel, L. (2010). Development and validation of the Arizona Cognitive Test Battery for Down syndrome. *Journal of Neurodevelopmental Disorders*, *2*(3), 149–164. http://dx.doi.org/10.1007/s11689-010-9054-3

Edgin, J. O., Pennington, B. F., & Mervis, C. B. (2010). Neuropsychological components of intellectual disability: the contributions of immediate, working, and associative memory. *Journal of Intellectual Disability Research*, *54*(5), 406–417. http://dx.doi.org/10.1111/j.1365-2788.2010.01278.x

Edgin, J. O., & Pennington, B. F. (2005). Spatial cognition in autism spectrum disorders: Superior, impaired, or just intact? *Journal of Autism and Developmental Disorders*, *35*(6), 729–745. http://dx.doi.org/10.1007/s10803-005-0020-y

Edgin, J. O., Spanò, G., Kawa, K., & Nadel, L. (2014). Remembering things without context: development matters. *Child Development*, *85*(4), 1491–1502. http://dx.doi.org/10.1111/cdev.12232

Esbensen, A. J., Hooper, S. R., Fidler, D., Hartley, S., Edgin, J., d'Ardhuy, X. L., ... Urv, T. (in press). Outcome measures for clinical trials in Down syndrome. *American Journal on Intellectual and Developmental Disabilities, 122*(3).

Fernandez, F., & Edgin, J. O. (2016). Pharmacotherapy in Down's syndrome: which way forward? *The Lancet Neurology, 15*(8), 776–777. http://dx.doi.org/10.1016/S1474-4422(16)30056-4

Fernandez, F., Morishita, W., Zuniga, E., Nguyen, J., Blank, M., Malenka, R. C., ... Garner, C. C. (2007). Pharmacotherapy for cognitive impairment in a mouse model of Down syndrome. *Nature Neuroscience*, *10*(4), 411–413. http://dx.doi.org/10.1038/nn1860

Frith, U., & Frith, C. D. (1974). Specific motor disabilities in Downs syndrome. *Journal of Child Psychology and Psychiatry*, *15*(4), 293–301. http://dx.doi.org/10.1111/j.1469-7610.1974.tb01253.x

Gioia, G. A., Isquith, P. K., Guy, S. C., & Kenworthy, L. (2000). *BRIEF: Behavior Rating Inventory of Executive Function professional manual*. Lutz, FL: Psychological Assessment Resources, Inc.

Guedj, F., Sébrié, C., Rivals, I., Ledru, A., Paly, E., Bizot, J. C., ... Delabar, J. M. (2009). Green tea polyphenols rescue of brain defects induced by overexpression of DYRK1A. *PLoS One*, *4*(2), e4606. http://dx.doi.org/10.1371/journal.pone.0004606

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377–381.

Heller, J. H., Spiridigliozzi, G. A., Crissman, B. G., Sullivan-Saarela, J. A., Li, J. S., & Kishnani, P. S. (2006). Clinical trials in children with Down syndrome: Issues from a cognitive research perspective. *American Journal of Medical Genetics Part C: Seminars in Medical*

DOI: 10.1352/1944-7558-122.3.215

Genetics, *142*(3), 187–195. http://dx.doi.org/10.1002/ajmg.c.30103

Heller, J. H., Spiridigliozzi, G. A., Crissman, B. G., McKillop, J. A., Yamamoto, H., & Kishnani, P. S. (2010). Safety and efficacy of rivastigmine in adolescents with Down syndrome: Long-term follow-up. *Journal of Child and Adolescent Psychopharmacology*, *20*(6), 517–520. http://dx.doi.org/10.1089/cap.2009.0099

Ji, N. Y., Capone, G. T., & Kaufmann, W. E. (2011). Autism spectrum disorder in Down syndrome: Cluster analysis of Aberrant Behaviour Checklist data supports diagnosis. *Journal of Intellectual Disability Research, 55*(11), 1064–1077. http://dx.doi.org/10.1111/j.1365-2788.2011.01465.x

Kaufman A. S., & Kaufman, N. L. (2004). *Kaufman Brief Intelligence Test–Second edition* (manual). Circle Pine, MN: AGS Publishing.

Korkman, M., Kirk, U., & Kemp, S. (1998). *NEPSY: A developmental neuropsychological assessment*. San Antonio, TX: The Psychological Corporation.

Morris, R. (1984). Developments of a water-maze procedure for studying spatial learning in the rat. *Journal of Neuroscience Methods*, *11*(1), 47–60. http://dx.doi.org/10.1016/0165-0270(84)90007-4

Pennington, B. F., Moon, J., Edgin, J., Stedron, J., & Nadel, L. (2003). The neuropsychology of Down syndrome: evidence for hippocampal dysfunction. *Child Development*, *74*(1), 75–93. http://dx.doi.org/10.1111/1467-8624.00522

Pritchard, A. E., Kalback, S., McCurdy, M., & Capone, G. T. (2015). Executive functions among youth with Down Syndrome and coexisting neurobehavioural disorders. *Journal of Intellectual Disability Research, 59*(12), 1130–1141. http://dx.doi.org/10.1111/jir.12217

Roper, R. J., Baxter, L. L., Saran, N. G., Klinedinst, D. K., Beachy, P. A., & Reeves, R. H. (2006). Defective cerebellar response to mitogenic Hedgehog signaling in Down's syndrome mice. *Proceedings of the National Academy of Sciences*, *103*(5), 1452–1456. http://dx.doi.org/10.1073/pnas.0510750103

Salehi, A., Faizi, M., Colas, D., Valletta, J., Laguna, J., Takimoto-Kimura, R. E. E. A., ... Mobley, W. C. (2009). Restoration of norepinephrine-modulated contextual memory in a mouse model of Down syndrome. *Science Translational Medicine*, *1*(7), 7ra17–7ra17. http://dx.doi.org/10.1126/scitranslmed.3000258

Spanò, G., & Edgin, J. O. (2016). Everyday memory in individuals with Down syndrome: Validation of the Observer Memory Questionnaire-Parent Form. *Child Neuropsychology*, 1–13. http://dx.doi.org/10.1080/09297049.2016.1150446

Strauss, E., Sherman, E. M., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary*. New York, NY: American Chemical Society.

Thomas, K. G., Hsu, M., Laurance, H. E., Nadel, L., & Jacobs, W. J. (2001). Place learning in virtual space III: Investigation of spatial navigation training procedures and their application to fMRI and clinical neuropsychology. *Behavior Research Methods, Instruments, & Computers*, *33*(1), 21–37. http://dx.doi.org/10.3758/BF03195344

Thomazeau, A., Lassalle, O., Iafrati, J., Souchet, B., Guedj, F., Janel, N., ... Manzoni, O. J. (2014). Prefrontal deficits in a murine model overexpressing the Down syndrome candidate gene dyrk1a. *The Journal of Neuroscience*, *34*(4), 1138–1147. http://dx.doi.org/10.1523/JNEUROSCI.2852-13.2014

Van Hoogmoed, A. H., Nadel, L., Spanò, G., & Edgin, J. O. (2016). ERP correlates of object recognition memory in Down syndrome: Do active and passive tasks measure the same thing? *Neuropsychologia*, *82*, 39–53. http://dx.doi.org/10.1016/j.neuropsychologia.2016.01.004

Visu-Petra, L., Benga, O., & Miclea, M. (2007). Visual-spatial processing in children and adolescents with Down's syndrome: a computerized assessment of memory skills. *Journal of Intellectual Disability Research*, *51*(12), 942–952.

Zabel, T. A., von Thomsen, C., Cole, C., Martin, R., & Mahone, E. M. (2009). Reliability concerns in the repeated computerized assessment of attention in children. *The Clinical Neuropsychologist*, *23*(7), 1213–1231.

Received 2/1/2016, accepted 11/14/2016.

We thank the families who made this work possible. This study was supported by grants from the LuMind Research Down Syndrome Foundation, NIH

DOI: 10.1352/1944-7558-122.3.215

1R01HD088409 (to Edgin) and R01HD07434601 (to Abbeduto). We thank Adele Diamond for sharing the Dots Task with our network. Valerie Deleon and Iser Deleon were instrumental in supporting the KKI recruiting site. We are also grateful to Carolyn Mervis for comments on an earlier version of the manuscript.

Authors:
Jamie O. Edgin and **Payal Anand,** University of Arizona; **Tracie Rosser,** Emory University; **Elizabeth I. Pierpont,** University of Wisconsin-Madison and University of Minnesota; **Carlos Figueroa,** University of Arizona; **Debra Hamilton** and **Lillie Huddleston,** Georgia State University; **Gina Mason**, Cornell University; **Goffredina Spanò,** University of Arizona; **Lisa Toole,** Johns Hopkins University; **Mina Nguyen-Driver,** Oregon Health Sciences University; **George Capone,** Johns Hopkins University; **Leonard Abbeduto,** University of Wisconsin-Madison and University of California, Davis; **Cheryl Maslen,** Oregon Health Sciences University; **Roger H. Reeves,** Johns Hopkins University; and **Stephanie Sherman,** Emory University.

Correspondence concerning this article should be addressed to Jamie O. Edgin, Department of Psychology, University of Arizona, 1503 E University Blvd., Tucson, AZ 85721 (e-mail: jamie.edgin@gmail.com).