J Neurodevelop Disord (2010) 2:149–164
DOI 10.1007/s11689-010-9054-3

Development and validation of the Arizona Cognitive Test Battery for Down syndrome

Jamie O. Edgin · Gina M. Mason · Melissa J. Allman · George T. Capone · Iser DeLeon · Cheryl Maslen · Roger H. Reeves · Stephanie L. Sherman · Lynn Nadel

Received: 14 January 2010 / Accepted: 11 June 2010 / Published online: 10 July 2010

Abstract Neurocognitive assessment in individuals with intellectual disabilities requires a well-validated test battery. To meet this need, the Arizona Cognitive Test Battery (ACTB) has been developed specifically to assess the cognitive phenotype in Down syndrome (DS). The ACTB includes neuropsychological assessments chosen to 1) assess a range of skills, 2) be non-verbal so as to not confound the neuropsychological assessment with language demands, 3) have distributional properties appropriate for research studies to identify genetic modifiers of variation, 4) show sensitivity to within and between sample differences, 5) have specific correlates with brain function, and 6) be applicable to a wide age range and across contexts. The ACTB includes tests of general cognitive ability and prefrontal, hippocampal and cerebellar function. These tasks were drawn from the Cambridge Neuropsychological Testing Automated Battery (CANTAB) and other established paradigms. Alongside the cognitive testing battery we administered benchmark and parent-report assessments of cognition and behavior. Individuals with DS (n=74, ages 7–38 years) and mental age (MA) matched controls (n=50, ages 3–8 years) were tested across 3 sites. A subsample of these groups were used for between-group comparisons, including 55 individuals with DS and 36 mental age matched controls. The ACTB allows for low floor performance levels and participant loss. Floor effects were greater in younger children. Individuals with DS were impaired on a number ACTB tests in comparison to a MA-matched sample, with some areas of spared ability, particularly on tests requiring extensive motor coordination. Battery measures correlated with parent report of behavior and development. The ACTB provided consistent results across contexts, including home vs. lab visits, cross-site, and among individuals with a wide range of socio-economic backgrounds and differences in ethnicity. The ACTB will be useful in a range of outcome studies, including clinical trials and the identification of important genetic components of cognitive disability.

Keywords Down syndrome · Intellectual disabilities · Genetics · Neuropsychology · Assessment · Clinical trials

Abbreviations

DS	Down syndrome
GWAS	Genome wide association studies
MA	Mental age

J. O. Edgin (✉) · G. M. Mason · L. Nadel
Department of Psychology, University of Arizona,
1503 E. University Blvd.,
Tucson, AZ 85721, USA
e-mail: jamie.edgin@gmail.com

M. J. Allman · I. DeLeon
Department of Behavioral Psychology, Kennedy Krieger Institute,
707 North Broadway,
Baltimore, MD 21205, USA

G. T. Capone
Pediatrics, Kennedy Krieger Institute,
707 North Broadway,
Baltimore, MD 21205, USA

C. Maslen
Department of Molecular and Medical Genetics,
Oregon Health Sciences University,
3181 S.W. Sam Jackson Park Rd.,
Portland, OR 97239-3098, USA

R. H. Reeves
Department of Physiology, Johns Hopkins University,
725 North Wolfe Street,
Baltimore, MD 21205, USA

S. L. Sherman
Department of Human Genetics, Emory University,
615 Michael Street, Suite 301,
Atlanta, GA 30322, USA

The last decade has seen substantial progress in our understanding of the neurobiological bases of the intellectual disability in Down Syndrome. Given recent findings highlighting promising avenues for pharmacological intervention in DS (Salehi et al. 2009; Fernandez et al. 2007; Roper et al. 2006; Salehi et al. 2006), clinical trials are imminent and will require assessments of the cognitive phenotype that are thoroughly validated. Likewise, new approaches such as genome-wide association studies (GWAS) add to the available tools for determining new therapeutic targets. While the field has reached a turning point in the understanding of this syndrome, precise measurements of the cognitive phenotype are required for further advances. To meet this need, we have developed and validated a neuropsychological battery for outcome studies in this population.

We begin with the perspective that outcome studies of this nature must employ a neuropsychological approach (see Pennington 2009 for a review). Previous outcome studies have utilized global cognitive assessments (e.g., IQ and adaptive behavior) and have not focused on tasks targeting the functions of specific brain regions affected in DS. Composite measures of cognition, such as IQ, may mask effects due to the non-specific nature of the assessment. To develop a clear picture of the biological pathways modified by an intervention, we need to understand which regions of the brain are tapped by particular outcome measures. Aside from direct measures of brain function or structure, the clearest associations between cognitive outcome and underlying brain function are reflected in established neuropsychological assessments. Our battery targets three areas of neuropsychological function, based on neuropathological findings in mouse models and humans with DS, including tests of prefrontal, hippocampal and cerebellar function (Reeves and Garner 2007; Crnic and Pennington 2000; Nadel 2003; Teipel and Hampel 2006).

Substantial information exists regarding neuropsychological measures in adult populations. Previous studies have addressed the optimal properties of clinical test batteries in adult populations, such as those with schizophrenia (Green et al. 2004). However, most neuropsychological assessments have not been properly validated in children or those with developmental disabilities. Special challenges are associated with testing individuals who have intellectual and developmental disorders, including substantial floor effects and the confounding effects of impaired language and attention (Table 1). Studies of intervention or genetic effects in individuals with or without developmental disabilities require a battery of measures with specific properties as follows:

1) *The battery should collectively assess a range of skills*, including broad skills (IQ, adaptive behavior) and skills attributable to particular neural systems. Multi-method assessment is also important, including informant scales, experimenter ratings and participant assessment. Clinical trials will require both the assessment of cognitive outcome as well as documentation of the state of the participant's skills on everyday tasks. Assessments should include at least two tests of each domain of interest to provide consistency in outcome. Assessment across multiple informants and with multiple measures is necessary to obtain an accurate picture of an individual's skills beyond any variability associated with the individual assessment.
2) *Control tasks should be included*. Control tasks are measures not expected to be affected by an intervention or related to a particular set of genes. These tasks may

Table 1 Assessment issues with DS and other intellectual disabilities

Challenge	Solutions
Floor effects	Use measures with graded difficulty measures that are error based. The standard scores on many measures do not provide a range at the lower end. Use measures with a low floor or use raw scores controlling for age
Language ability	Focus on tests that are primarily nonverbal, with nonverbal responses, adapt instructions for the population
Assessment variability due to behavior and cooperation	Include careful interview of caregivers (or teachers), experimenter ratings of behavior and cooperation on each test, multiple tests included as a performance check
Sensitivity of the measures to detect effects	Choose measures with continuous normally distributed outcomes; measures should demonstrate concurrent validity and show age-related change. Measures show differences between populations or have been documented to be impaired in the past literature
Flexibility of use	Choose measures that are easily adaptable across cultures, and that can be used in home or lab-based assessment
Applicability across a range of ages	Choose measures with graded levels of difficulty
Reproducibility, lack of validation of measures in populations with developmental disabilities	Use specially validated tests, collect test-retest reliability estimates that are sample-specific

represent cognitive strengths in the population. For instance, young individuals with DS show relatively strong skills in visuo-spatial construction and short-term spatial memory (Pennington et al. 2003; Vicari and Carlesimo 2006). The inclusion of tasks that measure relative strengths as well as weaknesses will allow studies to test for the specificity of effects. Findings of a genetic association will be more meaningful if the relationship is specific to a given domain. For example, a polymorphism in the *KIBRA* gene has been found to relate to neuropsychological assessments of hippocampal function in the typical population (Papassotiropoulos et al. 2006). This result is strengthened by the fact that this association was specific to several tasks measuring hippocampal function, while measures of prefrontal function were not associated.

3) *The test results should demonstrate variability and distributional properties that allow straightforward statistical analyses to detect treatment effects or exposure effects*. The measures included in a test battery ideally should have sufficient variability, evidenced by normally distributed scores and low levels of floor performance. Genome-wide association studies with developmental disorders, such as DS, may employ a design in which the genetic basis for a dichotomous trait (or extremes such as high/low IQ, high/low memory) is compared between groups that are generated based on cut-points. Floor effects must be minimized because they will mask changes or cause difficulties with group stratification. Tests that measure the number of errors are helpful in this regard as they limit floor performance levels.
4) *The tests should have the sensitivity to detect differences, including within sample variation and variation between populations*. The best evidence for sensitivity in measures comes from studies in which differences have been detected. Therefore, in devising this battery, we preferred to use tests that have shown differences in samples of individuals with DS in the past. Sensitivity is also evidenced by the presence of concurrent validity with other measures, demonstrated through correlations between each measure and parent ratings of development and behavior. Detection of cross-sectional differences in age, IQ and adaptive behavior also help to validate a measure's sensitivity.
5) *The tests should demonstrate adequate test-retest reliability in study samples of typically developing individuals as well as in study samples of those with intellectual disabilities*. Test-retest reliability data are essential to compare the change associated with retest variation to the change brought about by the intervention of interest. Test-retest reliability also directly relates to the power to detect change across time.
6) *The tests should relate, as specifically as possible, to particular brain regions*. While not always feasible, fMRI validation helps to determine that an assessment taps a particular neural system. With the exception of a few well-validated measures (e.g, the "Dots task", Davidson et al. 2006), very few tests are available for which substantial data exist on the neural basis of performance in children.
7) *The test battery must be resistant to confounding factors such as poor motivation, attention, or language skill*. Language skills are an area of great difficulty in several developmental disorders and in DS particularly. Therefore, tasks should be as simple as possible in their language components, including nonverbal measurements where possible. If verbal tasks are included, verbal and nonverbal versions of these tasks within the measurement domain are recommended. The inclusion of alternate versions increases the chances that genetic correlates of the task are related to the domain of interest (e.g., hippocampal memory) and not a confounding domain (e.g., language). While control for behavioral problems is not always possible, measures of attention and motivation should be completed alongside the assessment. The inclusion of ratings of behavior allows for the discrimination between deficits related to behavioral difficulties and deficits that represent true difficulties in cognitive function.
8) *The test battery must be applicable to a wide range of ages to allow easy comparison across various control populations and for the tracking of long-term change*. Adherence to this property is a challenge as the majority of neuropsychological tests involve a transition in test difficulty from the preschool to school-age years, approximately at the same developmental age at which many individuals with DS are currently functioning. Recently, tests have been developed to answer this measurement challenge, including the Dots task, a prefrontal measure appropriate from age 4 to adolescence (Davidson et al. 2006). In addition, tests from the CANTAB battery have been used in studies with children as young as age 4, providing continuous measures into adulthood (Luciana 2003).
9) *The test battery must be adaptable across a variety of contexts and cultures*. For large scale studies to be effective, testing will need to be executed across sites and locations, including home visits. Also, tests need to be language adaptable to properly verify the effect of certain genes in a variety of cultures that may have different environmental conditions. Since the incidence of many gene polymorphisms varies as a function of the genetic makeup of the people in a region, the relationship between outcomes and a particular gene may require validation across cultures.

With these properties in mind, we established the Arizona Cognitive Test Battery (ACTB). The tasks were drawn from the Cambridge Neuropsychological Testing Automated Battery (CANTAB) Eclipse battery or based on established paradigms (e.g., NEPSY, c–g arena and the Dots task (Davidson et al. 2006)). Tests from the CANTAB battery have been used in earlier studies of individuals with DS (Pennington et al. 2003; Visu-Petra et al. 2007). These studies have shown consistent impairment on several CANTAB tests, in particular the CANTAB Paired Associates Learning task. Similar spatial associative memory measures have been found to show age-related decline in DS (Alexander et al. 1997). Widely used batteries of tests, such as the CANTAB and NEPSY, benefit from the breadth of this use and hence the range of comparable data. For instance, the CANTAB has been used in several neuroimaging studies and with a wide-range of patient populations, including individuals with intellectual disability. Several CANTAB tests are currently included in a NIH study of brain development, adding to the validation of these measures in children (Waber et al. 2007; Luciana et al. 1999). Many of these tests are error-based, helping to limit floor effects, are applicable to children across a wide range of ages, and have alternate forms to decrease practice effects. Another positive aspect of the CANTAB tests chosen for the ACTB is that there is evidence for their effective use across languages and cultures (Luciana and Nelson 2002).

In validating the ACTB we detail its distributional properties, including the spread of individual scores generated by each task. Given the error-based nature of several of these measures, we expected that many of the tasks would generate normal distributions of data. To determine the sensitivity of the measures, we examine the ability of the tasks to relate to cross-sectional differences in age and IQ and to detect deficits in relationship to mental age (MA) matched controls. Given the extensive documentation of hippocampal, prefrontal and cerebellar deficits in this population, we expected that these measures would be impaired in relation to MA matched controls. We also detail the concurrent validity with parent ratings of behavior and development, preliminary test-retest reliability and examine the use of these measures across differing social and testing contexts.

Methods

Participants

Eighty individuals with DS were recruited across three sites including the University of Arizona, Tucson; Emory University, Atlanta; and Johns Hopkins University, Baltimore. Participants in Baltimore were recruited primarily through doctor's referral (GC), while individuals elsewhere were recruited via local and parent organizations and advertisement. Exclusion criteria included the presence of Robertsonian translocation (2 cases), mosaicism (3 cases), past head injury (0 cases), or, the presence of significant dementia in adults, as reported by participants' caregivers using the dementia scale for intellectual disabilities (1 case; Evenhuis 1992). Of the remaining 74 cases, 16 parents were unsure of the karyotype of their child with respect to the extra chromosome 21, and the remaining 58 endorsed that their child carried trisomy 21.

After exclusions, the total sample included 74 individuals (38 males; 36 females) with DS, ages 7–38 years old. The mean KBIT-II IQ of the full sample was 44.71±7.34 (range 40–75) and the mean SIB-R scaled score of adaptive behavior was 36.50±25.28 (range 0–89). The median total family income of the sample was $40–60,000 (corresponding to a score of "4.00" measured on a 10 point scale), with a substantial range starting from $0–15,000 to over $200,000. The ethnic distribution of the sample (i.e., maternal ethnicity) included 65.8% Caucasian, 24.7% Hispanic, 4.1% Black, 2.7% American Indian and 2.7% Asian American. The sample included 5/74 of individuals who had a parent report of a doctor's diagnosis of autism. Because this study constitutes a trial to test the validity of these assessments, the number of participants varies based on the introduction of the measures.

To test the sensitivity of these measures in detecting deficits in DS, we generated a matched sub-sample of individuals with DS and controls. Fifty typically developing mental age-matched children (ages 3–8 years) were recruited as controls at the Arizona site via advertisement and Experian corporation databases (i.e., a set of purchased marketing contacts). From this pool of 50 participants, we generated a sample of 36 (15 female, 21 male, ages 3–6 years) children in which the mean total of the verbal and nonverbal KBIT-II raw scores (a test of IQ; see below for description) was equivalent to a sample of 55 (26 female, 29 male) individuals with DS. Gender did not differ significantly between the matched samples, χ^2 (1, 91)= 0.28, p=0.60.

Procedure

All procedures were approved by the human participants committees of the participating institutions. Participants completed a 2 hr testing session in either a laboratory setting or in their homes with an examiner experienced in developmental disabilities. The ACTB was presented in a fixed order. The CANTAB procedures were modified using language appropriate to children. While the test order was predominantly fixed, the substitution and deletion of tests throughout the development of the battery resulted in a

varied order of tests across the full sample. After the task battery was solidified we implemented a two-version, counterbalanced order to avoid position effects. All computer measures were presented on a touch-screen computer; a button-box and joystick were utilized for the CANTAB SRT and Computer-generated Arena tasks. When approximately half of the tasks had been completed, the participant was allowed a break in which he or she was able to rest and choose a prize; in cases in which participants seemed to lose focus or become tired, more breaks were allowed. Experimenters recorded behavior during the session test, including reasons for non-completion. A subset of the sample (n=23) was also rated on behavior and cooperation for each individual cognitive test. The caregiver report measures, including background information, were completed as the participant was being tested.

Measures

Table 2 details the main task demands, usual age-range, outcome scores, test-retest reliabilities and links with brain function from past investigations. Reliability estimates in Table 2 were gathered from the standardized test manual, the study of Lowe and Rabbitt (1998) for CANTAB, and unpublished data from our group. While we have focused on a main outcome score for each of these tests it should be noted that several of these measures generate a detailed range of scores, such as completion time, reaction time, and errors. Many of the outcomes can be broken down over the different task phases to examine how errors change with alternating task demands. Furthermore, in the development of this battery several measures were piloted and eliminated that are not fully described here. Most notably, the CANTAB pattern and spatial recognition memory tasks and a version of the Wisconsin Card Sorting task generated extreme floor effects (>50% of subjects at floor) and were thus eliminated as candidates.

The Arizona Cognitive Test Battery

Tests of hippocampal function In the *CANTAB Paired-Associates Learning (PAL)* task (clinical version) the participant learns associations between abstract visual patterns and hiding locations on a computer screen. The task increases in difficulty from 1 to 8 patterns to be remembered. The outcome measure focused on here is the number of trials able to be completed on first view (score range=0–26). Impairments have been shown in several previous studies of DS (Pennington et al. 2003; Visu-Petra et al. 2007). Furthermore, a recent investigation of the heritability estimates of CANTAB measures has shown a substantial genetic influence on a memory composite which included the CANTAB PAL (i.e., a high heritability estimate of 57%) (Singer et al. 2006), suggesting that this measure may be useful for genetic studies.

The computer-generated arena (c–g arena) is an assessment of hippocampal function based on a paradigm from the animal literature (Thomas et al. 2001; Morris 1984). 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 the landmarks (distal cues) surrounding the arena. The main outcome variable is the percentage of time participants spend searching the quadrant of the arena in which the target is located (max= 100%). This task has been successfully used in individuals with DS and other developmental disabilities (Pennington et al. 2003; Edgin and Pennington 2005).

Tests of prefrontal function In the *CANTAB Intra-Extra Dimensional Set Shift (IED)* task participants are initially presented with two colored shapes, and must learn which shape is "correct" through trial-and-error. After several trials of recognizing the correct rule, the "correct" shape is reversed. Now the participant must recognize this rule shift and choose the new correct shape. In later trials, a second shape is transposed onto each shape, so that the participant must take another dimension into consideration when determining which shape is "correct." The number of errors per stage was the main outcome variable in this investigation. Temporal lobe and Alzheimer Disease (AD) patients show relatively unaffected performance on the ID/ED task, while frontal patients are impaired (Strauss et al. 2006).

The *Modified DOTS task* is a measure of inhibitory control and working memory suitable for participants aged 4 years to adulthood. There are 3 task phases. During the first phase, participants learn the rule associated with the cat stimuli (the congruent location rule). They are asked to press the button located directly below the cat on a computer touch 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 these rules are alternated randomly. Scores for the current investigation are calculated based on the percentage of correct responses for each phase of the test (max= 100%). Behavioral inhibition is required on incongruent trials to over-ride the prepotent tendency to respond on the same side as the visual stimulus.

Tests of cerebellar function The CANTAB Simple Reaction Time (SRT) task measures simple reaction time. Participants press a button when a stimulus (a white box) appears on a computer screen. The onset of the stimulus varies between trials. The outcome measure reported here consists of the

Table 2 Arizona cognitive test battery

Domain/test	Description	Primary ability assessed	Score for analysis	Test-retest *r*	Links to brain function	Age range
Benchmark						
KBIT-II verbal subscale (Kaufman and Kaufman 2004)	Points to pictures based on the word or phrase, answers riddles	Verbal comprehension, production	Total subscale raw	.88	–	4–90 years
KBIT-II nonverbal subscale	Semantic or visuo-spatial pattern completion	Problem solving	Total subscale raw	.76	–	4–90 years
Scales of independent behavior—revised (Bruininks et al. 1997)	Parent-report of everyday skills	Adaptive behavior	Standard score	.98	–	Infancy to 80+ years
CANTAB spatial span	Touching of boxes in order changing color on the screen, similar to CORSI span	Immediate memory for spatial-temporal sequences	Span	.64 (Lowe and Rabbitt 1998)	–	4 years and over
Prefrontal						
Modified dots task	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	Percent correct trials	*NA*	Activates prefrontal cortex in children in fMRI studies (Davidson et al. 2006)	4 years to late adolescence
CANTAB IED	Forced-choice discrimination task with change in relevant dimension	Set-shifting	Errors per stage (ln transformed)	.70 (Lowe and Rabbitt 1998)	Impaired in populations with frontal deficits (e.g., autism, Ozonoff et al. 2004). Deficits are ameliorated by dopaminergic medication (Strauss et al. 2006)	4 years and over
Hippocampal						
CANTAB Paired Associates	Recall for hidden abstract patterns and associated locations	Spatial associative memory	Errors to success, number trials completed on first view	.87 (average trials to success, Lowe and Rabbitt 1998)	Differentiates between patients with AD and controls with 98% accuracy 18 months prior to a formal diagnosis (Swainson et al. 2001)	4 years and over
Virtual computer-generated arena	Navigation of a virtual arena (via joystick) to find a fixed hidden target	Spatial memory	Percent time searching target quadrant	*NA*	Patients with hippocampal damage impaired (Skelton et al. 2000)	5 years and over
Cerebellar						
Finger sequencing task (Edgin and Nadel unpublished paradigm)	Sequences generated by tapping a number of fingers (1,2,3,4) to a lever in succession	Motor sequencing	Correct sequences, total taps	0.87, 0.91	Finger sequencing activates cerebellum (Desmond et al. 1997)	4 years and over

NEPSY visuomotor precision (ages 3–4) (Korkman et al. 1998)	Follows two tracks with a pen	Visuo-motor tracking, hand-eye coordination	Total score generated from completion time and errors	0.81	Visuo-motor tracking utilizing coordinated hand-eye movements activates cerebellum (Miall et al. 2000)	3–4 years
CANTAB simple reaction time (SRT)	Participants press a button in response to a box presented on a screen	Motor response time and attention	Median correct latency	*NA*	Simple motor response and attention tasks activate the cerebellum in fMRI (Allen et al. 1997)	4 years and over

median of the time taken for all correct responses recorded (median latency in ms). Slowing of motor response time is typical with cerebellar dysfunction, and studies have reported slowed reaction times in DS in comparison to MA controls and those with other developmental disabilities, such as autism (Frith and Frith 1974).

Finger sequencing is a skill reliant on cerebellar function in typical adults and children with autism (Desmond et al. 1997; Mostofsky et al 2009). In the current study we utilized two versions of the paradigm, a tabletop version modified from the NEPSY battery (Korkman et al. 1998) as well as a computerized version that we have developed after data collection with the NEPSY task. The table top version includes tapping a number of fingers (1, 2, 3 or 4) to the thumb in succession. The dominant and non-dominant hands are both tested. The tabletop task generates the time taken to complete a set number of sequences as an outcome score. During our development of the ACTB it became apparent that there was a high level of non-compliance on this measure (18.9% of children tested). Therefore, we developed a computerized version of the paradigm. The computerized version involves tapping a lever with either 1, 2, 3 or 4 fingers in sequence in the same manner that one would tap fingers to the thumb in the original paradigm. Both dominant and non-dominant hands are tested. There is a 10 s practice period followed by a 30 s test period for each hand. After each set is complete the subjects 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 data reported here are from the tabletop and computerized version. Between-group comparisons have been completed with the tabletop version. However, we report distribution data from the computerized version as it has received a greater level of subject compliance. Furthermore, 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 and Nadel unpublished data).

The final cerebellar task, the NEPSY visuomotor precision task (age category 3–4 years) involves the subject following a series of 2 tracks, a train and car track, from start to finish using a pen. The errors and completion time are considered together to generate a total score.

IQ and benchmark measures

Kaufman Brief Intelligence Test, Second Edition (KBIT-II) is a brief, individually administered measure of both verbal and

nonverbal intelligence appropriate for individuals from 4 to 90 years old (Kaufman and Kaufman 2004). Standard scores for the KBIT-II have a mean equal to 100, standard deviation of 15.

The CANTAB Spatial Span is a test of immediate spatial memory, modeled after the CORSI span task in which participants copy a sequence of blocks which 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 on verbal short-term memory, with strength in spatial short-term memory tasks (reviewed in Edgin et al. 2010). This task serves as a control measure in the battery.

The Scales of Independent Behavior-Revised (SIB-R) (Bruininks et al. 1997) 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 assessment to establish concurrent validity

In order to establish the validity of the neuropsychological test battery, we administered several informant-report measures of behavior and development, as follows:

Task completion checklist/task-specific behavioral ratings For each individual task experimenters rated the participant's attention to task, cooperation, affect, and anxiety level on a 5 point scale.

Nisonger child behavior rating form—parent version The Nisonger Child Behavior Rating Form (CBRF) (Aman et al. 1996) 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).

Behavior rating inventory of executive function—school age The BRIEF (Gioia et al. 2000) 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 (r=.80–.89 for most scales) (Strauss et al. 2006).

Conners 3 parent-ADHD scales The Conners-3 Parent Rating Scale includes subscales pertaining to symptoms of ADHD, mainly inattention and hyperactivity. Group membership (e.g., General Population, ADHD, Behavior Disorder, Learning Disorder) has been found to significantly affect all of the scale scores, suggesting good discriminative validity (Rzepa et al. 2007).

Statistical analysis

All analyses were conducted with SPSS 16.0. In validating this battery we first detail the descriptive statistics and distribution for each variable (Table 3). To measure the normality of test distributions, we calculated levels of skewness and kurtosis for each measure. Floor effects were calculated by determining the percentage of individuals receiving the lowest possible score in total sample in which the test was administered, or in some instances, the percentage of individuals failing to meet acceptable criteria for task performance (i.e., a threshold for the total number of correct trials). Due to multiple comparisons, the alpha for correlational analyses was set at p=0.01. Between-group differences in the MA control group and group with DS were assessed with t-test for normally distributed outcomes and Mann-Whitney U for non-normal outcomes (Table 4). To demonstrate validity, correlations were calculated between each measure and the other cognitive measures (Table 5) and parent report of behavior and development (Table 6). Differences across tasks in experimenter ratings of behavior and cooperation were examined with a paired sample t-test. To test age-related effects, we examined correlations between raw scores on each measure, total number of floor effects and age (Fig. 1). Test-retest data (Table 7) was analyzed with intraclass correlation and paired sample t-tests. The factors influencing battery performance (e.g., total measure floor effects) were examined using multiple linear regression with the factors of IQ, age, assessment location (home vs. lab), total family income, and ethnic group.

Results

Distribution

We first examined the distribution and floor effect of each test for the entire sample with DS (n=74). We analyzed the percentage of individuals unable to complete the task, the number of individuals at floor and the skewness and kurtosis values for each measure (Table 3).

Only four individuals were unable to complete the entire ACTB. Three of these four had a clinical diagnosis of

Table 3 Distribution data for each battery measure in the group with Down syndrome

Measure	n[a]	% not completed	% floor	Mean	SD	Range	Skewness	Kurtosis
Background and benchmark								
SIB-R adaptive behavior standard score	70	5.4	10.0	36.50	25.28	2–89	0.18	−0.89
KBIT-II verbal score	70	5.4	0.0	22.53	11.95	2–59	0.66	0.73
KBIT-II non-verbal score	70	5.4	5.7	11.89	5.64	0–27	−0.23	0.24
CANTAB spatial span span	66	10.8	33.3	2.30	1.27	1–6	0.67	−0.08
Hippocampal								
CANTAB PAL first trials correct	71	4.1	14.1	7.42	6.01	0–22	0.42	−0.80
Computer generated arena % time in the target quadrant	63	14.9	22.2	24.05	20.22	0–77	0.54	−0.26
Prefrontal								
CANTAB ID/ED errors per stage (ln transformed)	67	9.5	14.9	5.45	3.99	1.33–26	−0.85	0.27
Modified dots task inhib. control phase percent correct	65	8.5	29.2	63.59	31.85	0–100	−0.41	−.92
Modified dots task combined phase percent correct	65	8.5	41.5	54.13	18.30	15–100	1.08	1.15
Cerebellar								
CANTAB simple RT median corr. latency (ms)	66	10.8	25.8	735.26	321.39	275–1,656	0.79	0.10
NEPSY visuomotor Precision total score	48	9.4	0.0	15.24	4.57	3–21	−0.73	0.10
Finger sequencing mean correct sequences[b] (all trials)	11	9.1	18.0	230.56	51.70	144–321	0.30	0.67

[a] n varies based on the introduction of the measure, total original $n=74$, [b] based on the computerized version of the task

Table 4 Between-group differences in sample with DS and MA controls

Measure	DS *mean (SD)* ($N=55$)	MA control *mean (SD)* ($N=36$)	*t*	*p*	Effect sizes *d*
Background and benchmark					
KBIT-II verbal score	26.40 (10.33)	27.39 (5.49)	−0.59	0.60	–
KBIT-II non-verbal score	13.66 (4.51)	13.97 (3.44)	−0.36	0.72	–
CANTAB Spatial Span span	2.58 (1.26)	2.88 (1.07)	−1.17	0.25	–
Hippocampal					
CANTAB first trials correct	8.87 (5.78)	13.44 (6.54)	−3.05	0.001	0.74
Computer generated arena % time in the target quadrant	26.73 (19.83)	20.69 (21.19)	1.25	0.21	–
Prefrontal					
CANTAB ID/ED errors per stage[a]	5.02 (2.91)	3.86 (1.44)	2.60	0.009	0.51
Modified dots task inhib. control phase percent correct	67.31 (32.46)	75.62 (21.30)	−1.37	0.18	0.30
Modified dots task combined phase percent correct	57.28 (18.43)	66.55 (22.26)	−1.97	0.05	0.45
Cerebellar					
CANTAB simple RT median corr. latency (ms)	678.93 (314.46)	595.88 (142.69)	1.64	0.11	0.34
NEPSY visuomotor precision total score	15.08 (5.34)	14.57 (4.69)	0.41	0.69	–
Tabletop finger sequencing mean latency (all trials) (s)	44.97 (20.36)	33.92 (12.64)	2.62	0.01	0.65

[a] Analyzed with Mann-whitney U to account for the observed deviations in normality

Table 5 Age-adjusted partial correlations among cognitive measures in the sample with DS

Measure	Hippocampal		Prefrontal			Cerebellar	
	CANTAB PAL	Computer generated arena	CANTAB IDED	Dots inh.	Dots comb.	CANTAB simple RT	NEPSY VP
Hippocampal							
CANTAB PAL first trials completed	–	0.31+	−0.32**	0.44***	0.26+	−0.41**	0.49**
Computer generated arena		–	−0.02	0.20	0.01	−0.13	0.42**
Prefrontal							
CANTAB ID/ED errors			–	−0.45***	−0.38**	0.18	−0.27
Modified dots task inhib. control phase percent correct				–	0.46***	−0.38**	0.55***
Modified dots task combined phase percent correct					–	−0.40**	0.33+
Cerebellar							
CANTAB simple RT median corr. latency (ms)						–	−0.21
NEPSY visuomotor Precision total score							–

Finger sequencing was not compared to the other measures as too few subjects were tested on the computerized version to validate

+ trend at $p<0.05$, **$p<0.01$, *** $p<0.001$

autism and 1 had significant visual impairment. The values for non-completion range from 4.1% on the CANTAB Paired Associates Learning task to 14.9% on the computer generated arena (Table 3). These values included individuals who completed no portion of the task, either due to severe behavioral difficulties or, in 3/74 cases, to computer error.

Next we considered the number of individuals who completed the task, but demonstrated floor levels of performance. Our estimates of floor performance were conservative and it should be noted that even when a subject performed at floor, several of these tests generate a range of measures that are resistant to floor effects, such as number of errors or completion time. Percent of individuals

Table 6 Correlations between IQ, age, parent report of behavior and the cognitive measures in the sample with DS

Measure	KBIT total raw score	Age	SIBR SS	BRIEF Scale	Nisonger scales[b]
Hippocampal					
CANTAB PAL first trials correct	0.55**	0.21	0.32**	WM −0.51***	SA 0.61**
				IN −0.40**	Hyper −0.56**
Computer generated arena	0.40**	0.31+	0.20	WM −0.43**	
Prefrontal					
CANTAB ID/ED errors	0.50**	−0.24+	−0.22	WM 0.46** IN 0.54***	Hyper 0.61**
Modified dots task inhib. control phase percent correct	0.41**	0.29+	0.33**	WM −0.35** IN −0.38**	Hyper −0.56**
Modified dots task combined phase percent correct	0.55**	0.29+	0.27+		
Cerebellar					
CANTAB simple RT median corr. latency (ms)	−0.38**	−0.12	−0.28+		
NEPSY visuomotor precision total score	0.55**	0.54**	0.13		
Finger sequencing[a]					

WM BRIEF working memory, *BRIEF IN* inhibit, *SA* social adaptive scale of the Nisonger scales, *Hyper* Hyperactivity scale of the Nisonger

[a] too few subjects tested on the computerized version to validate

[b] no significant correlations were found with the Conners-III at $p<0.01$

+ trend at $p<0.05$, **$p<0.01$, *** $p<0.001$

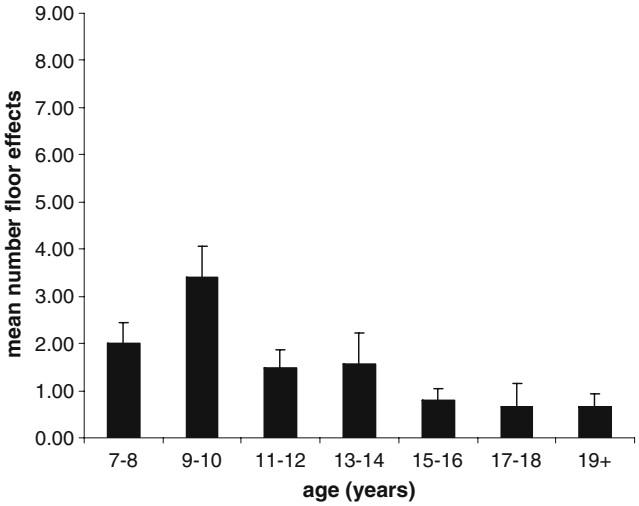

Fig. 1 Mean floor performance on ACTB cognitive measures across age in the group with DS[1]. 1: Maximum score=9; There was a significant negative correlation between total floor effects and age (r=−0.38, p=0.001)

with floor levels of performance (shown in Table 3) was calculated based on the number achieving the lowest possible scores in the total sample. Specifically for each measure, the floor was equivalent to: 1) attaining the lowest standard score on the SIB-R, 2) attaining a raw score of zero on the KBIT-II subtests, 3) attaining no spatial span, 4) by not correctly completing any of the first trials on PAL, 5) spending 0% time in the target quadrant on the virtual water maze task, 6) completing no stages on the ID/ED, 7) detecting less than chance (50%) on any phase of the Modified Dots task, 8) completing less than 70% correct trials on the CANTAB Simple reaction time, suggesting that task performance alone was too difficult, 9) attaining the lowest possible total score on the NEPSY visuomotor precision task and 10) completing only the 1 finger phase of the finger sequencing task. In general, levels of floor performance were low, with 58% (7/12) of the measures in the battery yielding floor performance in less than 15% of the sample. The highest rates occurred on the most advanced stage of the modified dots task (41.5%). Even when minor floor effects were present, most test outcomes had a normal distribution.

To determine the extent to which floor performance was affected by age, we examined the mean number of floor effects occurring across ages (Fig. 1, maximum score=9). The total score includes all the tests detailed in Table 3 that were administered to the child. Two tests had no floor effects. For this analysis, the participants in the sample with DS were split into the following age groups: 7–8 (n=11); 9–10 (n=10), 11–12 (n=15), 13–14 (n=7), 15–16 (n=10), 17–18 (n=6) and 19–38 years (n=15). There was a significant negative correlation with age (r=−.38, p=0.001). The figure shows that floor effects peaked around 9–10 years and were significantly lower after this age.

To define the extent to which each measure displayed a normal distribution, we calculated levels of skewness and kurtosis. Most measures produced scores that were normally distributed, generating values of skewness and kurtosis between −1 and 1. There were two exceptions. The ID/ED task variables (errors per stage) generated scores with high

Table 7 Test retest reliability estimates over 1.5 years in a subset of the sample with DS (n=10)

Measure	Test-retest ICC	Mean difference across time	Mean difference p value
Background and benchmark			
KBIT-II vocabulary raw score	0.93	−0.30	0.82
KBIT-II riddles raw score	0.88	0.00	1.00
KBIT-II matrices raw score	0.79	−1.10	0.50
CANTAB spatial span span	0.94	0.50	0.10
Hippocampal			
CANTAB first trials correct	0.78	0.60	0.69
Computer generated arena[a]	–	–	–
Prefrontal			
CANTAB ID/ED errors	0.81	−1.80	0.39
Modified dots task inhib. control phase percent correct	0.79	0.11	0.22
Modified dots task combined phase percent correct	0.60	−0.01	0.91
Cerebellar			
CANTAB simple RT median corr. latency (ms)	0.27	134.06	0.33
NEPSY visuomotor precision total score[a]	–	–	–
Finger sequencing[a]	–	–	–

ICC Intraclass correlation

[a] Too few subjects received both sessions to measure reliability. In the case of the c–g arena the versions changed from baseline to retest

levels of skewness and kurtosis (skewness=3.80, kurtosis=17.13). After log normal transformation these values fit a normal distribution (skewness=−0.85, kurtosis=0.27). The Modified Dots task also had mild deviations from normality (skewness=1.08, kurtosis=1.15), most likely due to the higher level of floor effects on this measure. However, given the minor nature of these deviations we chose to use this variable without transformation. Therefore, overall the levels of skewness and kurtosis for the battery measures were low and consistent with the expectations for a normal distribution.

Sensitivity to detect between-group effects

We used KBIT-II results to establish a MA control sample of typically developing children for comparison with the DS sample. The ability of each measure to detect differences in relation to the control group was assessed (Table 4). Several measures, including at least one from each domain, showed statistically significant differences between DS and MA groups at $p=0.01$, including the CANTAB Paired Associates Learning task, the CANTAB ID/ED and the tabletop finger sequencing test. The combined phase of the modified dots task was marginally significant at $p=0.05$.

Relation between measures and detection of individual differences

In addition to detecting between group differences, detecting differences within individuals with DS is also important. To determine relationships between measures, we examined the age-adjusted partial correlations among the individually administered neuropsychological measures in the battery (Table 5). There were significant correlations among many measures, with the prefrontal domain showing correlations among the tests in that domain ($p<0.01$). The NEPSY Visuomotor Precision task did not correlate with the CANTAB SRT, and the CANTAB PAL and the cg-arena were only marginally correlated ($p<0.05$).

To establish concurrent validity of the test battery measures, we correlated each neuropsychological measure with general IQ, participants' ages, and caregiver ratings of behavior and development including the SIB-R, (BRIEF school-age version, Nisonger CBRF, and Conners 3 ADHD indices (Table 6). K-BIT II total raw score correlated significantly with every neuropsychological measure ($p<0.01$). Participants' ages also related significantly to the NEPSY visuomotor precision task ($p<0.01$), and non-significant trends ($p<0.05$) were found between age and all prefrontal measures as well as the percent time in quadrant for the c–g arena. The SIB-R adaptive behavior standard score related significantly with the PAL first trials correct ($p<0.01$) and the number correct in the inhibitory control phase of the modified Dots task ($p<0.01$).

Regarding behavioral variables, parent ratings of more impairments on BRIEF Working Memory (WM) scale related to significantly fewer correct trials on the inhibitory control phase of the modified Dots task, fewer CANTAB PAL first trials correct, and less time spent in the target quadrant on the c–g arena ($p<0.01$). Additionally, the BRIEF Inhibit scale related to the CANTAB PAL first trials correct and the inhibitory control phase of the modified Dots task ($p<0.01$). These correlations remained significant after controlling for age and KBIT-II IQ.

A subset of the sample ($n=22$) received the Conners-III ADHD scale and the Nisonger scales of maladaptive behavior. There were no significant correlations with the Conners-III. The Nisonger CBRF adaptive social scale was significantly correlated with the number of first trials completed on the CANTAB PAL. Finally, the hyperactivity scale of the Nisonger was significantly correlated with the CANTAB PAL first trials completed, errors on the ID/ED task, and number correct on the inhibitory phase of the Modified Dots task ($p<0.01$).

In addition to parent-report, experimenters used a five-point scale to provide ratings of attention and compliance on each task in the ACTB for a sub-group of 22 individuals with DS. Performance levels were relatively high across the battery measures, with most tests displaying mean ratings of over 4. The overall mean level of attention was 3.96±0.90, and the mean cooperation level was 4.13±0.85. Using paired measures t-tests, the only test that showed significantly lower than average mean attention was the tabletop finger sequencing test ($M=3.47$, $SD=0.35$, $t(14)=2.75$, $p=0.016$). In the sample of 11 children we have tested so far on the computerized finger sequencing paradigm, the mean level of attention was 4.22 (SD=0.83), which may indicate an improvement over the tabletop version.

Preliminary test-retest reliability

Table 7 shows the results of test-retest reliability on the battery in a subset of 10 participants (mean age=12.50, SD 2.52, 5 males, 5 females) who were retested over an interval of 1.55 years (SD=0.44) on some subtests of the battery (see Table 7). Test-retest reliability was calculated using intraclass correlation (ICC) as this analysis is robust for small samples and eliminates problems in reliability analysis from the use of Pearson's r (Weir 2005). With the exception of the CANTAB SRT, the tests show high levels of test-retest correlation across this interval for most measures and no significant change in mean performance across the 1.55 year interval.

Application across contexts

Another important property of a cognitive test battery is that it should be applicable across contexts. Here, 30/74

participants (41%) were tested in their homes versus a lab environment. In a regression examining the effects of testing location, income level and ethnicity on floor levels of performance after controlling for IQ and age, we obtained a significant model $F(5, 57)=2.59$, $p=0.04$, accounting for 12% of the variance. Age was significantly related to total floor effects ($p=0.004$, $\beta=-.41$), and there was a trend for a relationship with IQ ($p=0.07$, $\beta=-.24$). However, there was no effect based on administering the battery at home vs. lab environment ($p=0.81$), based on total income ($p=0.79$), or based on ethnic background ($p=0.77$). Furthermore, the acquisition of 74 individuals with DS across the three testing sites is proof-in-principle of the efficacy of cross-site implementation of these tests. Therefore, these assessments appear to be robust for use in multiple testing contexts.

Discussion

To address the measurement challenges involved in outcome studies in individuals with Down syndrome (DS), we have developed the ACTB and tested the validity of these tests in a large, diverse sample collected across three testing sites. The ACTB is an assembly of tests designed specifically for known cognitive deficits in the population of people with DS. In designing the ACTB, we balanced the need to represent major cognitive processes known to be affected in DS with the practical aspects of testing a large population under time constraints. We chose measures that allowed for variation in outcome scores and minimized floor effects. We tested the sensitivity, specificity and preliminary reliability of the measures. We assessed the robustness of the tests across a range of ages and in the face of difficulties in attention and cooperation and across differing testing contexts. The bulk of the data suggests that the current collection of measures is well-suited for outcome studies in this syndrome.

Breadth of the ACTB

The ACTB primarily involves nonverbal tests of prefrontal, hippocampal and cerebellar dysfunction. In addition to the ACTB we included benchmark measures, such as the KBIT-II IQ and the SIB-R scale of adaptive behavior. Given past evidence for spared performance in immediate spatial memory (reviewed in Pennington et al. 2003), the CANTAB spatial span was included as a control measure. Alongside the battery we administered some parent-report measures of behavior and development, including the BRIEF scale of everyday executive function, the Conners-III hyperactivity and inattention indices and the Nisonger CBRF. These measures were correlated with battery measures and provide important information regarding the child's level of attention and behavioral difficulties in everyday life. The battery is comprehensive regarding nonverbal tests, but does not address deficits in language, which are also an important aspect of the cognitive profile. However, memory performance (i.e., particularly verbal short-term memory) is strongly related to language outcomes in this population. Given the extensive history of use of the digit span task, and findings of reasonably low levels of floor performance on this task (Edgin et al. 2010), this measure could be a useful complement to the ACTB.

Measure distribution

The majority of the tests in the ACTB demonstrate minimal participant loss and relatively low floor performance levels. Even when floor effects were evident, these effects typically did not disturb the normal distribution of the data. The presence of a normal distribution allows for measures to be useful for intervention studies due to the range of scores that can be achieved if skills improve. However, some of the tests do show a substantial number of participants who were untestable on the measure or at floor. While this issue might be considered a problem for the detection of cognitive decline, floor levels of performance were substantially lower in adults in our sample than in younger children (i.e., less than one test at floor per person on average). Therefore, in the population in which we require assessments of decline, the floor effects were acceptable and at lower levels than floor effects found in other investigations. For instance, Hon et al. (1998) reported that one-third of their sample was untestable on the Rivermead Behavioural Memory Test (Children's Version).

Also, our assessments of the floor effect were calculated using the most conservative measures possible. We have focused on set measures of the tasks to determine "floor" levels of performance (e.g., the attainment of at least 70% correct trials on the CANTAB SRT). The majority of measures also generate a range of variables that provide further information, such as reaction time or the types of errors. Similarly, the Dots tasks had a high level of floor performance on the combined phase, but because the test has graded levels of difficulty, starting with high completion levels in nearly all participants, we can also analyze the loss in accuracy as additional demands are added (i.e., it is informative to know that a subject can complete the first two phases, but has difficulty on the last phase).

ACTB sensitivity

Many battery measures showed sensitivity to detect between and within-group differences. Battery measures had correlations with parent-report of behavior and development,

demonstrating their concurrent validity. In particular, the inhibitory control phase of the Dots task and the CANTAB PAL showed strong relationships with assessments of everyday function. The correlation between the PAL and adaptive behavior in DS replicates the past findings with this measure (Edgin et al. 2010). Battery measures showed consistent relationships with IQ raw scores and some relation with age, particularly on motor measures.

This study revealed between group differences in each cognitive domain, including clear deficits on the CANTAB PAL, the CANTAB IDED and the finger sequencing task. While this large study is one of the first to demonstrate deficits across these domains in a single sample of individuals with DS, not every measure on the ACTB showed a deficit in relation to MA matched controls. As pointed out by Mervis and Klein-Tasman (2004), MA matched controls may not always be the most valid comparison. There are developmental differences in individuals that may not be accounted for by IQ. For instance, the use of a joystick or the use of a pen may be experience-dependent, showing a different developmental trajectory than IQ. The MA group may have had less exposure to these skills than individuals with DS due to their younger average age, masking differences on the NEPSY visuomotor precision task or the computer-generated arena. In fact, the NEPSY and the c–g arena were intercorrelated, suggesting that these measures share skills, most likely reflecting the shared motor component.

Many of the ACTB measures differentiate between DS and control groups, and the battery also provides important measures of within-group variation. Consider, for example, the case of tests assessing prefrontal function. While the current study suggests clear impairment on tests tapping prefrontal function, past studies have not always found these deficits (Pennington et al. 2003). However, in this study prefrontal tests were significantly related to variation in overall cognitive and everyday function in individuals with DS, even after deficits on hippocampal tests were taken into account. This finding suggests that prefrontal function is an important component contributing to variation across the spectrum of ability in DS.

In addition to data suggesting the importance of measurement of prefrontal function, there is a long history documenting the presence of hippocampal deficits in this population. Given the early emergence of Alzheimer's disease in these individuals, this domain is highly important for any outcome study of DS. Several therapeutic agents have been developed that will target memory deficits, and these studies will require validated hippocampal tests. The CANTAB PAL has been consistently impaired across past investigations in this syndrome, and has also been found to be impaired in our development of the ACTB.

However, we did not find impairments on the c–g arena task. Performance on the c–g arena was related to visuomotor precision scores, suggesting that the motor demands on this task may be problematic. This task also demonstrated a higher level of floor effects and non-completion rates. We have recently modified this task and made it easier through the use of clearer spatial cues and a larger target. We are currently developing a new subset of memory tasks that may be applicable across a range of ages (the ACMB-C, the Arizona Contextual Memory Battery for children, Edgin & Nadel) which could complement the ACTB in future studies. Meanwhile, CANTAB PAL provides an excellent measure of memory function for this population.

Turning to the cerebellar tasks, the current findings suggest generally slower motor coordination and response times in DS. Other studies have suggested that there are links between cerebellar-based motor problems and prefrontal function in typical development and the development of children with intellectual disabilities, a finding that is replicated here with the correlations between measures (Diamond 2000; Hartman et al. 2010). Therefore, a substantial deficit in motor performance in this population may relate to the development of other cognitive domains. However, no between-group impairments were found on the NEPSY visuomotor precision task. We hypothesize that the absence of a difference may result from other factors, such as the lack of experience among the MA controls in using a pencil for task completion. Regardless, motor tasks such as the ones contained in the ACTB constitute important assessments of outcome in this population.

Measure specificity and reliability

A number of measures on the ACTB were intercorrelated, with or without controlling for age. A high degree of intercorrelation between cognitive measures is frequently observed in studies of cognition in intellectual disabilities (Detterman and Daniel 1989). For instance, one study found significant correlations between verbal and spatial measures in Williams syndrome, a syndrome that has an opposing profile in these domains (Mervis 1999). Therefore, the specificity of these measures may not be able to be assessed by patterns of correlations alone. The ACTB benefits from targeting measurements to the neurological components underlying task performance. Clearly, fMRI studies will be beneficial to further determine the neural bases of cognitive deficits on these measures. However, given evidence of specificity from correlations with other measures (i.e., inhibitory phase of the dots task is correlated with parent report of inhibition), several of the ACTB tests appear to be more targeted than other available measures. Furthermore, with the exception of the CANTAB SRT, the majority of the tests in this battery also generated very good levels of test-retest reliability over a substantial time period.

Test consistency across different behavioural outcomes, age, and context

These data emphasize that careful control for levels of attention and cooperation is needed in outcome studies of individuals with DS. While inattention or poor compliance cannot always been minimized, concurrent caregiver and experimenter report can help to define any difficulty that may occur during the testing situation or on a regular basis. The ratings included alongside the ACTB are useful for this purpose. Many of these ratings can include multiple informants, such as teachers. On the whole, experimenter ratings of the participant's attention and cooperation levels suggest that the tasks were well-tolerated. The lowest levels were reported for the tabletop finger tapping test, a finding that was remedied by the use of the newly developed computerized version.

Age is an important factor to consider when designing outcome assessments for individuals with DS. Floor performance was highest in younger children, declining after age 10 years. While not applicable in the youngest children, these assessments proved useful here across a wide age-range, spanning nearly 30 years. In this regard, this battery may be useful for studying cognitive decline in DS. Further effort is needed for the development of valid and reliable tests of cognitive outcome in early-mid childhood in this syndrome.

A practical benefit of the ACTB is that it provides consistent results across testing contexts, including home vs. lab visits and ascertainment at multiple sites. We were able to easily implement the testing battery at three sites after standard training. Given the ease of battery implementation and administration, we anticipate few influences by site as testing networks are expanded in the future. The current investigation also suggests that the battery allows for testing across a range of socio-economic and ethnic backgrounds. These characteristics are particularly important since molecular genetics and intervention studies will require the acquisition of hundreds of participants across multiple contexts. One caveat is that our current assessment of the effects of ethnicity and any associated cultural variation was limited to groups living in a westernized culture in the US. Since studies have suggested that non-western cultures may be differently advantaged or disadvantaged on neuropsychological tests (Rosselli and Ardila 2003) more testing would be required to use this battery in a non-western setting.

Conclusions

The ACTB offers a robust assessment of function and dysfunction in a number of areas of critical concern in DS, providing a useful tool for future outcome studies in this syndrome. The standardized platform allows for comparison of results across laboratories, and will be useful for the comparison of data across various outcome studies. Given that the tests are predominantly nonverbal, the battery may also provide links with basic research into specific brain functions across species, serving as a bridge to speed the translation of drug trials in animal models to clinical trials in people with Down syndrome.

Acknowledgements We thank the families that made this work possible. This study was supported in part by grants from the Down Syndrome Research and Treatment Foundation, DSRTF (to R.H.R., L.N. and J.O.E.), the Anna and John Sie Foundation (to R.H.R. and L.N.), the National Down Syndrome Society Charles Epstein Award (to J.O.E.), the Arizona Alzheimer's Research Consortium (to L.N.), the Jerome Lejeune Foundation (to J.O.E. and L.N.), the University of Arizona Foundation (to L.N.), and the Oregon Clinical and Translational Research Institute (OCTRI) (to C.M.).

References

Alexander GE, Saunders AM, Szczepanik J, Straussburger T, Pietrini P, Dani A, et al. Relation of age and apolipoprotein E to cognitive function in Down syndrome adults. NeuroReport. 1997;8:1835–40.

Allen G, Buxton RB, Wong EC, Courchesne E. Attentional activation of the cerebellum independent of motor involvement. Science. 1997;275:1940–3.

Aman MG, Tassé MJ, Rojahn J, Hammer D. The Nisonger CBRF: a child behavior rating form for children with developmental disabilities. Res Dev Disabil. 1996;17(1):41–57.

Bruininks RK, Woodcock RW, Weatherman RF, Hill BK (1997) *Scales of independent behavior—revised (SIB-R)*. Riverside.

Crnic LS, Pennington BF. Down syndrome: neuropsychology and animal models. Progr Infancy Res. 2000;1:69–111.

Davidson MC, Amso D, Anderson LC, Diamond A. Development of cognitive control and executive functions from 4 to 13 years: evidence from manipulations of memory, inhibition, and task switching. Neuropsychologia. 2006;44(11):2037–78.

Desmond JE, Gabrieli JD, Wagner AD, Ginier BL, Glover GH. Lobular patterns of cerebellar activation in verbal working-memory and finger-tapping tasks as revealed by functional MRI. J Neurosci. 1997;17:9675–85.

Detterman DK, Daniel MH. Correlations of mental tests with each other and with cognitive variables are highest for low IQ groups. Intelligence. 1989;13:349–59.

Diamond A. Close interrelation of motor development and cognitive development and of the cerebellum and prefrontal cortex. Child Dev. 2000;71(1):44–56.

Edgin JO, Pennington BF. Spatial cognition in autism spectrum disorders: superior, impaired, or just intact? J Autism Dev Disord. 2005;35(6):729–45.

Edgin JO, Pennington BF, Mervis CB. Neuropsychological components of intellectual disability: the contributions of immediate, working, and associative memory. J Intellect Disabil Res. 2010;54(5):406–17.

Evenhuis HM. Evaluation of a screening instrument for dementia in ageing mentally retarded persons. J Intellect Disabil Res. 1992;36 (4):337–47.

Fernandez F, Morishita W, Zuniga E, Nguyen J, Blank M, Malenka RC, et al. Pharmacotheraphy for cognitive impairment in a mouse model of down syndrome. Nat Neurosci. 2007;10:411–3.

Frith U, Frith CD. Specific motor disabilities in Down's syndrome. J Child Psychol Psychiatry. 1974;15(4):293–301.

Gioia GA, Isquith PK, Guy SC, Kenworthy L. Behavior rating inventory of executive function. Odessa: Psychological Assessment Resources; 2000.

Green MF, Nuechterlein KH, Gold JM, Barch DM, Cohen J, et al. Approaching a consensus cognitive battery for clinical trials in schizophrenia: the NIMH-MATRICS conference to select cognitive domains and test criteria. Biol Psychiatry. 2004;56:301–7.

Hartman E, Houwen S, Scherder E, Visscher C. On the relationship between motor performance and executive functioning in children with intellectual disabilities. J Intellect Disabil Res. 2010;54(5):468–77.

Hon J, Huppert FA, Holland AJ, Watson P. The value of the Rivermead behavioural memory test (Children's Version) in an epidemiological study of older adults with Down syndrome. Br J Clin Psychol. 1998;37(1):15–29.

Kaufman AS, Kaufman NL (2004) *Kaufmann brief intelligence test*. 2nd ed. Pearson Assessments.

Korkman M, Kirk U, Kemp SL. NEPSY: A developmental neuropsychological assessment. San Antonio: The Psychological Corporation; 1998.

Lowe C, Rabbitt P. Test/re-test reliability of the CANTAB and ISPOCD neuropsychological batteries: theoretical and practical issues. Cambridge neuropsychological test automated battery. Neuropsychologia. 1998;36:915.

Luciana M. Practitioner review: computerized assessment of neuropsychological function in children: clinical and research applications of the Cambridge Neuropsychological Testing Automated Battery (CANTAB). J Child Psychol Psychiatry. 2003;44 (5):649–63.

Luciana M, Nelson CA. Assessment of neuropsychological function through use of the Cambridge neuropsychological testing automated battery: performance in 4- to 12-year-old children. Dev Neuropsychol. 2002;22(3):595–624.

Luciana M, Lindeke L, Georgieff M, Mills M, Nelson CA. Neurobehavioral evidence for working-memory deficits in school-aged children with histories of prematurity. Dev Med Child Neurol. 1999;41(8):521–33.

Mervis CB. The Williams Syndrome cognitive profile: strengths, weaknesses, and interrelations among auditory short-term memory, language, and visuospatial constructive cognition. In: Winograd E, Fivush R, Hirst W, editors. Ecological approaches to cognition: essays in honor of Ulric Neisser. Mahwah: Erlbaum; 1999. p. 193–227.

Mervis C, Klein-Tasman B. Methodological issues in group-matching designs: α levels for control variable comparisons and measurement characteristics of control and target variables. J Autism Dev Disord. 2004;34(1):7–17.

Miall RC, Imamizu H, Miyauchi S. Activation of the cerebellum in co-ordinated eye and hand tracking movements: an fMRI study. Exp Brain Res. 2000;135(1):22–33.

Morris R. Developments of a water-maze procedure for studying spatial learning in the rat. J Neurosci Meth. 1984;11(1):47–60.

Mostofsky SH, Powell SK, Simmonds DJ, Goldberg MC, Caffo B, Pekar JJ. Decreased connectivity and cerebellar activity in autism during motor task performance. Brain. 2009;132(9):2413–25.

Nadel L. Down's syndrome: a genetic disorder in biobehavioral perspective. Genes Brain Behav. 2003;2(3):156–66.

Ozonoff S, Cook I, Coon H, et al. Performance on Cambridge neuropsychological test automated battery subtests sensitive to frontal lobe function in people with autistic disorder: evidence from the collaborative programs of excellence in autism network. J Autism Dev Disord. 2004;34(2):139–50.

Papassotiropoulos A, Stephan DA, Huentelman MJ, et al. Common *KIBRA* alleles are associated with human memory performance. Science. 2006;314(5798):475–8.

Pennington BF. How neuropsychology informs our understanding of developmental disorders. J Child Psychol Psychiatry. 2009;50(1–2):72–8.

Pennington BF, Moon J, Edgin J, Stedron J, Nadel L. The neuropsychology of Down syndrome: evidence of hippocampal dysfunction. Child Dev. 2003;74(1):75–93.

Reeves RH, Garner CC. A year of unprecedented progress in Down syndrome basic research. Ment Retard Dev Disabil Res Rev. 2007;13(3):215–20.

Roper RJ, Baxter LL, Saran NG, Klinedinst DK, Beachy PA, Reeves RH. Defective cerebellar response to mitogenic Hedgehog signaling in Down syndrome mice. Proc Natl Acad Sci USA. 2006;103:1452–6.

Rosselli M, Ardila A. The impact of culture and education on nonverbal neuropsychological measurements: a critical review. Brain Cogn. 2003;52:326–33.

Rzepa S, Conners CK, Gallant S, Pitkanen J, Sitarenios G, Marocco ML (2007) Development and psychometric properties of the Conners 3 short forms. Poster session presented at the meeting of the American Psychological Association, San Francisco.

Salehi A, Delcroix JD, Belichenko PV, Zhan K, Wu C, et al. Increased App expression in a mouse model of Down's syndrome disrupts NGF transport and causes cholinergic neuron degeneration. Neuron. 2006;51:29–42.

Salehi A, Faizi M, Colas D, Valletta J, Laguna J, et al. Restoration of norepinephrine modulated contextual memory in a mouse model of Down syndrome. Science Translational Medicine. 2009;1 (7):7–17.

Singer JJ, MacGregor AJ, Cherkas LF, Spector TD. Genetic influences on cognitive function using the Cambridge neuropsychological test automated battery. Intelligence. 2006;34(5):421–8.

Skelton RW, Bukach CM, Laurance HE, Thomas KGF, Jacobs JW. Humans with traumatic brain injuries show place-learning deficits in computer-generated virtual space. J Clin Exp Neuropsychol. 2000;22(2):157.

Strauss E, Sherman E, Spreen O. A compendium of neuropsychological tests. Administration, norms and commentary. New York: Oxford University Press; 2006.

Swainson R, Hodges J, Galton C, Semple J, Michael A, Dunn BD, et al. Early detection and differential diagnosis of Alzheimer's disease and depression with neuropsychological tasks. Dement Geriatr Cogn Disord. 2001;12:265.

Teipel SJ, Hampel H. Neuroanatomy of Down syndrome in vivo: a model of preclinical Alzheimer's disease. Behav Genet. 2006;36 (3):405–15.

Thomas KGF, Hsu M, Laurance HE, Nadel L, Jacobs WW. Place learning in virtual space III: Investigation of spatial navigation training procedures and their application to fMRI and clinical neuropsychology. Behav Res Meth Instrum Comput. 2001;33 (1):21.

Vicari S, Carlesimo GA. Short-term memory deficits are not uniform in Down and Williams syndromes. Neuropsychol Rev. 2006;16:87–94.

Visu-Petra L, Benga O, Tincas I, Miclea M. Visual-spatial processing in children and adolescents with Down's syndrome: a computerized assessment of memory skills. J Intellect Disabil Res. 2007;51(12):942–52.

Waber DP, De Moor C, Forbes PW, Almli CR, Botteron KN, Leonard G, et al. The NIH MRI study of normal brain development: performance of a population based sample of healthy children aged 6 to 18 years on a neuropsychological battery. J Int Neuropsychol Soc. 2007;13:729–46.

Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res. 2005;19:231–40.