ARTICLE OPEN ACCESS

Validation of the NIH Toolbox Cognitive Battery in intellectual disability

Rebecca H. Shields, MS, Aaron J. Kaat, PhD, Forrest J. McKenzie, BS, Andrea Drayton, BS, Stephanie M. Sansone, PhD, Jeanine Coleman, PhD, Claire Michalak, BS, Karen Riley, PhD, Elizabeth Berry-Kravis, MD, PhD, Richard C. Gershon, PhD, Keith F. Widaman, PhD, and David Hessl, PhD

Correspondence
Dr. Hessl
drhessl@ucdavis.edu

Neurology® 2020;94:e1229-e1240. doi:10.1212/WNL.0000000000009131

Abstract

Objective

To advance the science of cognitive outcome measurement for individuals with intellectual disability (ID), we established administration guidelines and evaluated the psychometric properties of the NIH-Toolbox Cognitive Battery (NIHTB-CB) for use in clinical research.

Methods

We assessed feasibility, test-retest reliability, and convergent validity of the NIHTB-CB (measuring executive function, processing speed, memory, and language) by assessing 242 individuals with fragile X syndrome (FXS), Down syndrome (DS), and other ID, ages 6 through 25 years, with retesting completed after 1 month. To facilitate accessibility and measurement accuracy, we developed accommodations and standard assessment guidelines, documented in an e-manual. Finally, we assessed the sensitivity of the battery to expected syndrome-specific cognitive phenotypes.

Results

Above a mental age of 5.0 years, all tests had excellent feasibility. More varied feasibility across tests was seen between mental ages of 3 and 4 years. Reliability and convergent validity ranged from moderate to strong. Each test and the Crystallized and Fluid Composite scores correlated moderately to strongly with IQ, and the Crystallized Composite had modest correlations with adaptive behavior. The NIHTB-CB showed known-groups validity by detecting expected executive function deficits in FXS and a receptive language deficit in DS.

Conclusion

The NIHTB-CB is a reliable and valid test battery for children and young adults with ID with a mental age of ≈5 years and above. Adaptations for very low-functioning or younger children with ID are needed for some subtests to expand the developmental range of the battery. Studies examining sensitivity to developmental and treatment changes are now warranted.

RELATED ARTICLE

Editorial
Finding a common path to the assessment of persons with intellectual development disorders
Page 507

From the University of California Davis Medical Center (R.H.S.); Human Development Graduate Group (R.H.S.), University of California Davis, Sacramento; Feinberg School of Medicine (A.J.K., R.C.G.), Northwestern University, Chicago, IL; MIND Institute and Department of Psychiatry and Behavioral Sciences (F.J.M., A.D., S.M.S., D.H.), University of California Davis Medical Center, Sacramento; Morgridge College of Education (J.C., K.R.), University of Denver, CO; Departments of Pediatrics (C.M., E.B.-K.) Neurological Sciences (E.B.-K.), and Biochemistry (E.B.-K.), Rush University Medical Center, Chicago, IL; and Graduate School of Education (K.F.W.), University of California, Riverside.

Go to Neurology.org/N for full disclosures. Funding information and disclosures deemed relevant by the authors, if any, are provided at the end of the article.

The Article Processing Charge was funded by authors.

Glossary

CAT = computerized adaptive testing; **DCCS** = Dimensional Change Card Sort; **DEXT** = Developmental Extension; **DS** = Down syndrome; **EF** = executive function; **Flanker** = Flanker Inhibitory Control and Attention; **FM** = Forward Memory; **FXS** = fragile X syndrome; **FSIQ** = full-scale IQ; **ICC** = intraclass correlation; **ID** = intellectual disability; **LS** = List Sorting Working Memory; **NEPSY-In** = NEPSY Inhibition subtest; **NIHTB-CB** = NIH Toolbox Cognitive Battery; **OID** = ID of other or unknown cause; **OR** = Oral Reading Recognition; **PC** = Pattern Comparison Processing Speed; **PSM** = Picture Sequence Memory; **PPVT-4** = Peabody Picture Vocabulary Test, 4th edition; **PVT** = Picture Vocabulary; **SB-5** = Stanford-Binet, 5th edition; **VABS-3** = Vineland Adaptive Behavior Scales, 3rd edition.

Approximately 2.0% of the global population has an intellectual disability (ID), an understudied condition with lifelong effects on academic, vocational, and personal functioning.[1] As the etiologies and mechanisms underlying specific forms of ID are discovered, targeted treatments and human clinical trials soon follow in the translational process,[2,3] raising the potential for medical remediation of disability. The preclinical development of promising targeted treatments for ID-associated disorders has not been followed by successes in human trials. Unfortunately, testing accessibility issues, pervasive floor effects, and lack of consensus on acceptable cognitive endpoints have been obstacles in the field. Although other barriers such as limitations of animal models are involved, it is apparent that the IDs continue to lag behind other neurologic or psychiatric conditions on scalable, psychometrically supported, and broadly accepted endpoints.[4,5]

The NIH Toolbox Cognitive Battery[6–8] (NIHTB-CB), an iPad-based battery of brief memory, executive function (EF), processing speed, and language tests, was developed within the NIH Blueprint for Neuroscience Research. The NIHTB-CB has the potential to provide a highly standardized, objective, and scalable tool for use across laboratories and clinical trial sites. As an extension of our pilot work,[9] the present study reflects progress made over a 4-year period to empirically validate and refine the NIHTB-CB for ID (aim 1), including standardized administration guidelines required for this challenging population, with the goal of supporting its use as a set of outcome measures for clinical trials and other clinical research. The second aim was to measure the sensitivity of the battery to detect known cognitive phenotypes in 2 ID-associated syndromes that are a focus of translational research: Down syndrome[5] (DS) and fragile X syndrome[10] (FXS). On the basis of prior research, we hypothesized EF deficits in FXS and DS and episodic memory deficits in DS (relative to a heterogeneous other-ID group), a relative language strength in FXS, and no group differences on visual processing speed.

Methods

Standard protocol approvals, registrations, and patient consents

Institutional Review Board approval was obtained at each site (University of California Davis MIND Institute, University of Denver, and Rush University Medical Center) before study initiation. Written consent was obtained from each guardian-participant pair according to Institutional Review Board requirements.

Participants

Because of the study aim to measure the sensitivity of the battery to syndrome-specific cognitive phenotypes, 3 groups were recruited. Two ID-associated syndromes were chosen: DS (affecting ≈1 in 700)[11] and FXS (affecting ≈1 in 7,000 males and 1 in 11,000 females).[12] Individuals with ID of other or unknown cause (OID) were also recruited to evaluate the NIHTB-CB within a more heterogeneous group and to serve as comparison to FXS and DS. Eligible participants met the following criteria: chronologic age of 6 through 25 years; full-scale IQ (FSIQ) <80 on the Stanford-Binet, 5th edition (SB-5),[13] mental age of at least 3.0 years on the SB-5 (in concordance with the lowest chronologic age limit of the NIHTB-CB), adaptive behavior deficits as measured by the Vineland Adaptive Behavior Scales, 3rd edition Comprehensive Interview[14] (VABS-3), speech of at least short phrases, English as first language, and stable medication and intervention regimen for 6 weeks before enrollment. Exclusion criteria were uncorrected vision impairment, uncontrolled seizures, motor impairment affecting touchscreen use, and a history of head trauma, brain infection, or stroke.

In all, 288 participants consented to the study. After completion of the SB-5, 45 participants were ineligible: 16 with an FSIQ >80 and 29 with a mental age <3.0 years. One participant discontinued due to behavioral noncompliance. Across sites, 242 participants completed initial neuropsychological testing, with 228 completing retesting of the NIHTB-CB ≈1 month later to examine test reliability. This retest duration was selected to evaluate reliability within a typical time interval used in clinical trials. Participants included 91 with DS, 75 with FXS, and 76 with OID. A subset of 21 participants with OID had a diagnosed ID-associated syndrome; the represented syndromes were 16p11.2 deletion (1), 22q11.2 deletion (1), Bannayan Riley Ruvalcaba (1), cri-du-chat (1), fetal alcohol (5), Floating-Harbor (1), Kleefstra (1), mitochondrial disease (1), mosaic trisomy 8 (1), neurofibromatosis type 1 (2), Phelan-McDermid (1), Potocki-Lupski (4), and Williams (1).

Protocol

The NIHTB-CB and all convergent validity measures (see below) were completed at visit 1 across 2 days. After

completion of the SB-5, the order of remaining assessments was randomized with the exception of the NIHTB-CB, which was the first assessment of day 2. The order of the NIHTB-CB tests was randomized for each participant. At visit 2, the NIHTB-CB was readministered with the same test order within participants.

The NIHTB-CB

The NIHTB-CB is a computerized assessment validated in ages 3 to 85 years in the general population. The battery includes 7 tests: Dimensional Change Card Sort (DCCS), Flanker Inhibitory Control and Attention (Flanker), List Sorting Working Memory (LS), Pattern Comparison Processing Speed (PC), Picture Sequence Memory (PSM), Picture Vocabulary (PVT), and Oral Reading Recognition (OR).[9,15] The NIHTB-CB provided experimental Developmental Extension (DEXT) versions of DCCS and Flanker designed to be more accessible to lower-functioning or very young participants. Because tests have multiple age versions, the participant's mental age derived from the SB-5 was used to select test versions, allowing for a starting point of reasonable difficulty, thereby reducing frustration and improving compliance. The DEXT versions were used for participants in the 3- to 7-year mental age range. For PVT and OR, there is 1 computerized adaptive testing (CAT) version, and the start point is typically based on age (children) or education (adults). Instead, we used the education override feature, entering the grade equivalent of the mental age as the start point.

In addition, PSM has multiple forms available for each test version. Pilot reliability results suggested nonequivalence of forms. To assess PSM reliability, Form A was used at visit 1 and for half of participants at visit 2 (PSM A-A); the other half received Form B at visit 2 (PSM A-B). For LS, pilot studies demonstrated that additional teaching items improved feasibility. For the current study, PowerPoint slides of these teaching items were used before test items on LS Age 3–6. The NIHTB-CB developers then released the LS Age 3–6 Experimental version with these extended instructions during the study, which was subsequently used.

Convergent validity tests

Six tests were preselected as convergent validity measures for the NIHTB-CB. The NEPSY Inhibition subtest[16] (NEPSY-In), iPad version, was used as the convergent measure for DCCS. The NEPSY-In measures cognitive flexibility and inhibitory control. From piloting the NEPSY-In, we found that participants could rarely do the most difficult level (Switching). We thus administered only the Naming and Inhibition portions and created a prorated score indicating the number of correct items per minute. For Flanker, we used the Conners Kiddie Continuous Performance Test 2nd Edition[17], administered on a computer with a spacebar as the response button. The hit reaction time SD was used as the convergent validity variable. For LS convergent validity, the SB-5 verbal working memory raw score was used. PC validity was measured with the Wechsler Preschool and Primary Scale of Intelligence, 4th Edition[18] Bug Search, from the number of correct items per minute. The Leiter International Performance Scale, 3rd Edition[19] Forward Memory (FM) subtest assesses sequential memory span. The raw score was the convergent variable for PSM. The Peabody Picture Vocabulary Test, 4th edition[20] (PPVT-4) measures receptive vocabulary. The raw score from the PPVT-4 iPad version was used for PVT. For OR, the Woodcock Johnson 4th Edition[21] Letter-Word Identification was used, which measures letter recognition and single word reading. Discriminant measures for NIHTB-CB tests were selected out of these measures by choosing a feasible measure of a different construct than the NIHTB-CB test.

Adaptations

To increase feasibility and to improve reliability and validity in ID, we developed a manual of standardized procedures regarding the test environment and NIHTB-CB administration: the "NIH Toolbox Cognitive Battery Supplemental Administrator's Manual for Intellectual and Developmental Disabilities" (e-Manual; hereafter Supplemental Manual, links.lww.com/WNL/B58).[22] Strategies to proactively improve feasibility and to reduce participant stress included using a visual schedule before the visit, a caregiver questionnaire on behaviors and potential reinforcement rewards, and a visual token board of the NIHTB-CB for the participant to check off during testing. Best practices in administering standardized assessments with appropriate accommodations for ID were used.[23] Test-specific guidelines are available in the manual to aid future users of the NIHTB-CB in standard administration and feasibility specifically for the ID population. In addition, the Supplemental Manual includes the Administration Form that we developed to document test environment, behavioral responses, and validity of tests for each participant.

Data cleaning and scoring

After every administration of the NIHTB-CB, the Administration Form was used to record whether each test was considered valid for the participant. All analyses used only valid scores. The most common reasons for invalid scores were an invalid response pattern, refusal, and excessive prompting (3.0%, 0.78%, and 0.72% of all scores, respectively).

Before conducting analyses, we visually inspected all data and bivariate correlations for normality and the presence of outliers. All NIHTB-CB tests were examined for floor or ceiling issues. Only 2 tests had such issues. At visit 1, 21 individuals received a score at the floor on LS Age 3–6, 33 received a score at the ceiling on PSM Age 3–4, and 4 received a score at the floor on PSM Age 5–6. After a thorough review of administration details and reliability and validity analyses, LS floored scores were kept in the analyses. Because PSM has multiple age versions and these participants likely should have received a harder or easier version, PSM scores at floor or ceiling were excluded.

Data analysis

SAS version 9.4 (SAS Institute Inc, Cary, NC) and R version 3.6.0 (R Foundation for Statistical Computing, Vienna, Austria) were used for analyses. Visit 1 data were used for

feasibility and validity analyses. For validity and reliability, NIHTB-CB raw scores were used: computed score on DCCS, Flanker, and PC; theta score on PSM, PVT, and OR; raw score on LS and Flanker DEXT; and percent correct on DCCS DEXT. Because DEXT scores are currently on a different scale than standard Flanker and DCCS scores, DEXT results are presented separately. For test-retest reliability, single-score intraclass correlations (ICCs) were used. The Cohen *d* was used to evaluate potential practice effects, with paired-sample *t* tests to measure the significance of change. Convergent, discriminant, and ecologic validities were measured with Pearson correlations.

Our prior work on IQ measurement demonstrated the utility of deviation-based scoring to deal with problematic floor effects in ID.[24] We used this method in the current study in place of NIHTB-CB standard scores to circumvent imposed floored scores (e.g., the current NIHTB-CB winsorizes age-corrected standard scores at 54). We created *z* scores by transforming participant raw scores on each NIHTB-CB test using normative means and SDs for their chronologic age band. The *z* scores were used to create deviation-based composites following the previously defined criteria.[25] For a Crystallized Composite, 1 of 2 valid scores on PVT and OR was required; for a Fluid Composite, 4 of 5 valid scores on Flanker, DCCS, LS, PC, and PSM were required. The Crystallized and Fluid *z* scores were used to create a Cognitive Function Composite, used in ecologic validity analyses. Deviation scores were also created for FSIQ on the SB-5[24] and were used in FSIQ analyses. Group comparisons to examine known-groups validity were assessed with a 2-way mixed-model analysis of variance on NIHTB-CB *z* scores. Significant results were followed up with the Tukey honest significant difference tests to examine group differences.

Data availability

On request to the corresponding author, anonymized data are available to share.

Results

Descriptive statistics

Table 1 presents descriptive statistics by diagnostic group and overall. Groups did not differ significantly by chronologic age ($F_{2,238}$ = 0.91, p = 0.41) or by VABS-3 Adaptive Behavior Composite ($F_{2,227}$ = 1.50, p = 0.23). However, FSIQ differed significantly by group ($F_{2,238}$ = 31.6, p < 0.001), with FSIQ higher in OID than in both DS [$t(238)$ = 7.57, p < 0.001] and FXS [$t(238)$ = 6.05, p < 0.001]. FSIQ did not significantly differ between DS and FXS [$t(238)$ = 1.23, p = 0.22]. Similarly, mental age was significantly different by group ($F_{2,238}$ = 25.3, p < 0.001), with mental age higher in OID than in both DS [$t(238)$ = 6.76, p < 0.001] and FXS [$t(238)$ = 5.40, p < 0.001]. Mental age did not differ significantly between DS and FXS [$t(238)$ = 1.10, p = 0.27]. Figure 1A shows the distribution of the NIHTB-CB Cognitive Function Composite age-adjusted standard scores of the sample without the current imposed floor, illustrating the variability of the sample below this floor and the benefit of deviation-based composites (used in all analyses).

Feasibility

Feasibility data are provided in table 2 as the percentage of participants with valid scores on each test. Feasibility overall was similar to the normative 3- to 15-year-old sample,[26] with DCCS, PC, and LS having slightly lower feasibility and Flanker, PSM, PVT, and OR having similar or higher feasibility rates than the normative sample. Even down to a mental age of 3 years, PSM, PVT, and OR feasibility was very good. The feasibility of the remaining tests improved particularly at 5 years. All tests were feasible for nearly every participant with a mental age of ≥6 years.

Test-retest reliability

Test-retest reliability was assessed after ≈1 month (mean = 31.7 days, SD = 6.3 days) (table 3). ICCs on each test were moderate to strong, with the exception of DCCS DEXT and PSM A-A, which were in the high 0.40s. All composites had strong reliability. Three tests had small but significant visit 1 to visit 2 effect sizes, reflecting a modest increase in performance: Flanker, PC, and the PSM A-B group. However, the PSM A-B increase likely reflects nonequivalent forms rather than a practice effect because the PSM A-A group had no practice effect. The Fluid and Cognitive Function composites also had small but significant increases, suggestive of small practice effects. Within groups, reliability coefficients were mostly moderate to strong; some DEXT and PSM reliabilities in small sample sizes were not significant: Flanker DEXT in OID (ICC = 0.46, p = 0.14) and both PSM groups in DS (A-A: ICC = 0.24, p = 0.15; A-B: ICC = 0.33, p = 0.05). In FXS, DCCS reliability (ICC = 0.41) was notably lower than in the total sample (ICC = 0.71), but FXS Flanker reliability (ICC = 0.84) was stronger than in the total sample (ICC = 0.74).

Construct validity

Convergent and discriminant validity results are presented in table 4. Convergent correlations ranged from moderate to strong, with the exception of Flanker DEXT (n = 29, r = −0.26, p = 0.17). Group results generally reflect validity similar to that of the overall results. Of note, PSM was only weakly correlated with Leiter-FM in DS (r = 0.33, p = 0.01) and in OID (r = 0.32, p = 0.01).

In DCCS, for a subgroup of participants below a raw score of 1.88, there appeared to be no association with the NEPSY-In; in this subgroup, DCCS score was not significantly correlated with NEPSY-In score (r = 0.29, p = 0.16). This DCCS score represents participants who did not pass the introductory switching portion of the test. When this subgroup was removed, validity improved (r = 0.57, p < 0.001).

Ecologic validity

Table 5 provides the ecologic validity of NIHTB-CB tests and composites. The composites each had moderate to strong correlations with FSIQ (figure 1B), as did all NIHTB-CB test scores other than Flanker DEXT. VABS-3 Adaptive Behavior Composite had small but significant correlations with several tests (Flanker, PC, PSM, PVT, and OR) and with the Crystallized and Cognitive Function composites, with better performance associated with higher levels of adaptive behavior.

Table 1 Participant descriptive information

	Total (n = 242)		Down syndrome (n = 91)		Fragile X syndrome (n = 75)		Other ID (n = 76)	
Race, n								
American Indian/Alaska Native	3		1		1		1	
Asian	6		2		2		2	
Black	24		4		8		12	
Native Hawaiian/Pacific Islander	3		1		1		1	
White	171		70		60		41	
>1	26		10		2		14	
Ethnicity, % Hispanic or Latino	19.9		18.7		8.0		32.9	
Sex, % male	59.3		45.1		73.3		63.2	
Primary caregiver 4-y degree, %	61.8		64.8		62.7		56.6	
ASD diagnosis (parent report), %	35.3 (8 unknown)		6.6 (0 unknown)		52.0 (6 unknown)		52.6 (1 unknown)	
	M	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
Chronologic age, y	15.71	5.15	15.88	5.17	16.16	4.92	15.05	5.35
Mental age (SB-5), y	5.20	1.53	4.67	1.20	4.91	1.35	6.12	1.64
Full-scale IQ (SB-5)	53.75	15.87	47.65	12.90	50.48	15.13	64.24	14.69
Vineland-3 ABC score	52.59	17.11	54.93	16.23	50.24	18.63	51.93	16.54
DCCS computed score	3.81	2.51	3.33	2.24	3.20	2.26	4.68	2.71
Flanker computed score	4.92	2.36	4.54	2.23	4.46	2.44	5.71	2.28
LS raw score	6.30	4.41	4.30	3.63	6.02	3.63	8.34	4.75
PC computed score	33.07	15.93	25.92	13.74	33.10	14.08	39.84	16.74
PSM theta score	−1.58	0.90	−1.86	0.76	−1.74	0.87	−1.09	0.88
PVT theta score	−2.91	2.80	−3.79	2.73	−2.70	2.57	−2.09	2.83
OR theta score	−5.84	4.96	−6.62	5.29	−6.21	4.53	−4.60	4.78

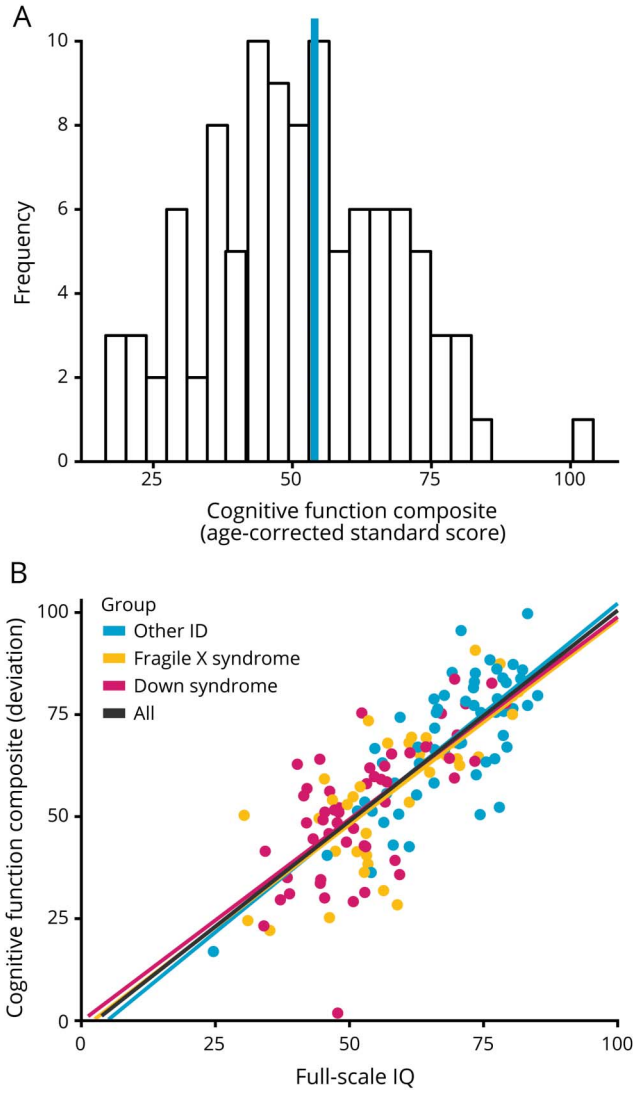
Abbreviations: ABC = Adaptive Behavior Composite; ASD = autism spectrum disorder; DCCS = Dimensional Change Card Sort; Flanker = Flanker Inhibitory Control and Attention; ID = intellectual disability; LS = List Sorting Working Memory; M = mean; OR = Oral Reading and Recognition; PC = Pattern Comparison Processing Speed; PSM = Picture Sequence Memory; PVT = Picture Vocabulary; SB-5 = Stanford-Binet, 5th edition.

Syndrome-specific comparisons (known-groups validity)

To examine the specificity of the NIHTB-CB to detect syndrome-specific performance, a 2-way mixed-model analysis of variance was conducted on NIHTB-CB test z scores with group as a between-participants factor and NIHTB-CB test as a within-participants factor; we also examined their interaction (figure 2). IQ was included as a repeated-measures varying covariate, and an IQ-by-test interaction term was included to allow the effect of IQ to vary by test. Because groups differed on FSIQ, covarying IQ aimed to clarify whether group results reflect phenotype-specific impairments or if they simply reflect globally poorer performance due to overall level of cognitive functioning. To avoid overcontrolling for the domain of interest (because NIHTB-CB domains overlap with components of IQ), verbal IQ was used as the covariate for the fluid NIHTB-CB test outcomes (DCCS, Flanker, LS, PC, and PSM), and nonverbal IQ was used as the covariate for crystallized NIHTB-CB tests (PVT and OR). To obtain effect sizes for pairwise comparisons, the Cohen d was calculated from the estimated marginal means from the model to account for the effects of IQ.

There was a main effect of group on NIHTB-CB z scores ($F_{2,238} = 4.90$, $p = 0.008$), as well as a main effect of test on z scores ($F_{6,1,067} = 17.26$, $p < 0.001$). These were qualified by

Figure 1 NIHTB-CB cognitive function composite score distribution and association with FSIQ



(A) Histogram showing the true distribution of cognitive function composite scores in the full sample of individuals with intellectual disability (ID). The NIH Toolbox Cognitive Battery (NIHTB-CB) current floor (winsorized at 54; vertical blue line) excluded more than half of the present sample from more accurate measurement below that level. (B) Association between the deviation-based cognitive function composite and deviation-based full-scale IQ (FSIQ) by group.

a significant group × test interaction ($F_{12,1,067} = 3.68$, $p < 0.001$) and a significant IQ × test interaction ($F_{6,1,067} = 6.72$, $p < 0.001$). Follow up Tukey tests showed that FXS performed worse on DCCS than OID [$t(238) = 4.02$, $p < 0.001$, $d = 0.52$] and worse than DS [$t(238) = -3.04$, $p = 0.007$, $d = 0.39$]. On Flanker, FXS performed worse than OID [$t(238) = 3.45$, $p = 0.002$, $d = 0.45$] and worse than DS [$t(238) = -2.85$, $p = 0.01$, $d = 0.37$]. These 2 EF test results supported the hypothesized EF impairment in FXS, although the DS impairment relative to OID was not supported. On PVT, FXS performed better than DS, fitting with expected language strength in FXS [$t(238) = -2.77$, $p = 0.02$, $d = 0.36$]. On PC, DS showed a poorer performance than OID, which approached significance [$t(238) = 2.24$, $p = 0.07$, $d = 0.29$]. There were no significant group differences on LS, PSM, or OR.

Discussion

This study provides the first comprehensive examination of the psychometric properties and feasibility of the NIHTB-CB for individuals with ID, with an initial focus on 2 of the most common genetic causes with robust translational research programs: FXS and DS. Overall, the NIHTB-CB has demonstrated strong potential for use as an objective, standardized outcome measure that can be confidently used in ID trials with participants with a mental age of 5 years or higher. Results of the study demonstrate very strong psychometrics for the Crystallized reasoning tests (PVT and OR) and good to excellent performance of Fluid reasoning tests, with more variation across FXS, DS, and OID groups for some measures (e.g., strong reliability for Flanker in FXS compared to DS, and vice versa for DCCS). Indeed, the Fluid Composite appeared to have more consistently strong reliability across conditions and a solid convergent association with FSIQ. Thus, it may be a good candidate outcome measure for studies seeking to examine broad nonverbal cognitive changes for individuals with a mental age of ≥5 years. Below a mental age of 5 years, feasibility was more variable across tests, indicating the need for further adaptations, scoring algorithms on developmental extensions, or new tests targeting these lower-functioning individuals.

The Supplemental Manual was compiled after hundreds of administrations of the NIHTB-CB to individuals with ID, and the study results on feasibility, reliability, and validity support its use. We encourage researchers and examiners planning to use the NIHTB-CB to follow these guidelines for the ID population.

Group comparisons demonstrated that the NIHTB-CB is sensitive to substantial EF deficits among individuals with ID in that all 3 groups performed relatively poorly on Flanker and DCCS. In particular, participants with FXS showed weakness in inhibitory control and attention and cognitive flexibility, in excess of their general cognitive level and compared to controls with other forms of ID. This aligns with previous research showing that boys with FXS are impaired in inhibitory control, set shifting, and planning relative to mental age–matched controls[27] and that boys with FXS have impairments in inhibition and attention relative to mental age–matched controls and relative to children with DS.[28] The hypothesized EF weakness in DS compared to OID (to a lesser extent than FXS) was not found; however, there have been some mixed results on EF in children with DS compared to children with other IDs.[29,30] The low-scoring subgroup on DCCS are those who failed introduction to the switching portion of the test; for participants who cannot perform switching at the earliest level, the test may be less sensitive to variation in EF. Removing this subgroup from analyses strengthened the convergent validity correlation. This suggests that these

Table 2 At visit 1, proportion of participants who received a valid score[a]

	Total, n (%)	Down syndrome, n (%)	Fragile X syndrome, n (%)	Other ID, n (%)
Flanker	195 (80.6)	78 (86.7)	50 (66.7)	67 (88.2)
Flanker (with DEXT)[b]	239 (98.8)	89 (98.9)	75 (98.7)	75 (98.7)
DCCS	152 (63.6)	55 (61.1)	40 (54.1)	57 (76.0)
DCCS (with DEXT)[b]	232 (97.1)	87 (96.7)	72 (97.3)	73 (97.3)
LS	164 (68.9)	57 (64.8)	45 (60.0)	62 (82.7)
PC	179 (75.2)	59 (64.8)	58 (80.1)	62 (82.7)
PSM	220 (92.1)	85 (94.4)	65 (86.7)	70 (93.3)
PVT	237 (98.3)	88 (97.8)	73 (97.3)	76 (100.0)
OR	233 (97.9)	87 (97.8)	72 (97.3)	74 (98.7)

Abbreviations: DCCS = Dimensional Change Card Sort; DEXT = Developmental Extension; Flanker = Flanker Inhibitory Control and Attention; ID = intellectual disability; LS = List Sorting Working Memory; OR = Oral Reading and Recognition; PC = Pattern Comparison Processing Speed; PSM = Picture Sequence Memory; PVT = Picture Vocabulary.
[a] Technical difficulties and administration errors are excluded.
[b] On Flanker, 27.4% of the sample needed the DEXT portion. On DCCS, 53.1% needed DEXT.

individuals' level or lack of cognitive flexibility may not be captured by the introductory switching portion of DCCS. An extension of DCCS that does capture this construct in very young or low-functioning persons would have much value. The lower reliability of DCCS in the FXS group may also reflect this limitation in that participants with FXS overall scored very low on this test. Because anxiety and hyperactivity are common in FXS, it is possible that individual state interacts with the task complexity and cognitive flexibility to result in more variable performance over time in this group. When participants needed Flanker DEXT or DCCS DEXT, they were almost always able to perform the tests; however, a few issues remain to be worked out regarding these experimental versions, notably the ability to interpret these scores relative to the standard Flanker and DCCS scores. This study highlighted other concerns with the DEXT measures (e.g., difficulty may be ordered incorrectly, or portions are overly burdensome). Our results provide clear evidence that DEXT levels are necessary (and extremely feasible), especially in those with FXS and DS, but that further refinements and modifications are necessary.

To improve feasibility of LS, extra instructions and practice items were developed and used. The LS Age 3–6 experimental version with these additions is now available. Feasibility did improve after our initial pilot studies; however, the test remains challenging for this population, and some participants pass practice but get no test sequences completely correct. The limited feasibility and floored or low variability in raw scores suggest that a lower range of LS is necessary. A potential approach may be to give some credit on early sequences that are partially recalled or items recalled out of sequence because the construct of working memory builds on more basic short-term memory (e.g., SB-5 Working Memory index and Wechsler Working Memory indices).[13,31]

Both language tests, PVT and OR, had excellent performance in our samples of individuals with ID. Both demonstrated strong reliability and clear domain specificity with a much higher convergent than discriminant correlation. The NIHTB-CB was sensitive to expected language characteristics in that DS was impaired relative to FXS on PVT.[32] These tests have an advantage over other language measures such as the PPVT-4 in that PVT and OR are brief (≈3 minutes each) but accurate, owing to CAT; the ability to obtain results with a brief assessment is especially important in individuals with frequent behavior or attention issues such as in those with ID.

Episodic memory is relevant to clinical trials, particularly for DS, in which memory impairments are well documented,[33] and FXS, in which memory of sequential information is especially impaired.[34] It is important to emphasize that despite these known weaknesses, PSM was among the highest of the test scores in each group (figure 2), suggesting that individuals may have compensatory strategies for performing well, perhaps such as use of the contextual information in the stories. Reliability was moderate and lower than that of the normative 3- 15-year-old study, although improved from our pilot study. The significant effect size in the PSM A-B group suggests nonequivalence of Forms A and B. Notably, in DS, ICCs were small and nonsignificant, suggesting that in its present form, PSM appears contraindicated as a separate outcome measure for this population.

The high rate of ceiling scores on PSM 3–4 suggests that mental age may not be a good indicator of start point on PSM in ID, at least at this mental age level. It is also possible that this age version is too easy in comparison to PSM scores on older versions; perhaps a person's score on 1 PSM version is not equivalent to that person's score on another. Development of a CAT PSM test with the full range of version

Table 3 Test-retest reliability and examination of practice effects

	Total			Down syndrome			Fragile X syndrome			Other ID		
	No.	ICC (95% CI)	Cohen *d*	No.	ICC (95% CI)	Cohen *d*	No.	ICC (95% CI)	Cohen *d*	No.	ICC (95% CI)	Cohen *d*
Flanker	144	0.74 (0.65–0.81)	0.09[a]	56	0.63 (0.44–0.76)	0.20	37	0.84 (0.70–0.91)	−0.11	51	0.69 (0.51–0.81)	0.15
Flanker DEXT	36	0.60 (0.33–0.77)	−0.01	8	0.73 (0.18–0.94)	−0.33	23	0.62 (0.28–0.82)	−0.05	5	0.46 (−0.37–0.92)	0.56
DCCS	97	0.71 (0.60–0.80)	0.13	32	0.78 (0.58–0.88)	0.18	23	0.41 (0.01–0.69)	0.24	42	0.82 (0.68–0.90)	0.03
DCCS DEXT	73	0.49 (0.30–0.65)	−0.16	32	0.46 (0.14–0.69)	−0.17	27	0.50 (0.16–0.74)	−0.13	14	0.53 (0.03–0.82)	−0.21
LS	113	0.74 (0.64–0.81)	0.14	37	0.75 (0.56–0.86)	0.24	32	0.65 (0.40–0.81)	−0.11	43	0.69 (0.49–0.82)	0.22
PC	133	0.77 (0.62–0.85)	0.29[b]	45	0.74 (0.48–0.86)	0.39[b]	40	0.71 (0.50–0.84)	0.21[a]	48	0.75 (0.54–0.86)	0.35[c]
PSM A-A	60	0.47 (0.25–0.64)	0.04	22	0.24 (−0.22–0.60)	−0.14	15	0.54 (0.07–0.82)	0.27	23	0.40 (0.01–0.69)	0.11
PSM A-B	60	0.55 (0.34–0.71)	0.29[a]	22	0.33 (−0.06–0.65)	0.22	18	0.54 (0.14–0.80)	0.25	20	0.40 (−0.03–0.71)	0.34
PVT	186	0.85 (0.81–0.89)	0.03	69	0.87 (0.80–0.92)	0.08	57	0.79 (0.66–0.87)	0.01	60	0.86 (0.78–0.92)	−0.02
OR	183	0.96 (0.95–0.97)	0.02	70	0.95 (0.92–0.97)	0.01	56	0.96 (0.93–0.98)	−0.01	57	0.98 (0.97–0.99)	0.04
FC	97	0.83 (0.67–0.91)	0.33[b]	28	0.84 (0.52–0.93)	0.40[b]	27	0.79 (0.56–0.90)	0.34[a]	42	0.77 (0.55–0.88)	0.33[c]
CC	191	0.93 (0.91–0.95)	0.00	73	0.92 (0.87–0.95)	0.02	58	0.94 (0.90–0.96)	−0.03	60	0.91 (0.85–0.94)	−0.00
CFC	97	0.92 (0.85–0.96)	0.22[b]	28	0.93 (0.79–0.97)	0.23[c]	27	0.91 (0.79–0.96)	0.27[a]	42	0.89 (0.78–0.94)	0.19[a]

Abbreviations: A-A = Form A at visit 1 and Form A at visit 2; A-B = Form A at visit 1 and Form B at visit 2; CC = Crystallized Composite; CFC = Cognitive Function Composite; CI = confidence interval; DCCS = Dimensional Change Card Sort; DEXT = Developmental Extension; FC = Fluid Composite; Flanker = Flanker Inhibitory Control and Attention; ICC = intraclass correlation; ID = intellectual disability; LS = List Sorting Working Memory; OR = Oral Reading and Recognition; PC = Pattern Comparison Processing Speed; PSM = Picture Sequence Memory; PVT = Picture Vocabulary.
[a] $p < 0.05$.
[b] $p < 0.001$.
[c] $p < 0.01$.

difficulties would likely simplify the testing process and yield more comparable and reliable scores.

PC showed adequate feasibility overall and good feasibility in FXS and OID, with sufficient feasibility at a mental age of ≥4 years. While the ICC was excellent, there was a small but significant practice effect, although smaller than that found in the normative sample.[35] The most common reason for lack of feasibility on PC was an invalid alternating response pattern, especially common in DS. The task has an inherent challenge of understanding "same" and "different" while mapping this choice onto smiley face and frown face options. For participants without invalid response patterns, PC performs well in ID. On the basis of the feasibility challenges, we developed a new processing speed task, Speeded Matching, now available as an experimental version in the app. The task is to select the animal face among 3 foils that matches a target image. Future psychometric studies will provide more information about its performance.

The Fluid, Crystallized, and Cognitive Function composites demonstrated reliability results similar to those of the normative age 3 to 6 sample, with small practice effects in Fluid and Cognitive Function composites (although smaller than in the normative sample). Each composite was well correlated with FSIQ. Although not all participants were able to receive a composite (due to missing valid scores on some tests), when complete, the composites appear to perform well. The deviation method used to create these composites has a clear advantage over the current age-corrected standard scores, on which more than half of our sample obtained scores at the lower limit, currently set at 54. We are conducting analyses to identify the best option for composite scores below this floor (deviation approach vs extension of existing age-adjusted

Table 4 Convergent and discriminant validity of NIHTB-CB tests (Pearson *r*)

	Total				Down syndrome				Fragile X syndrome				Other ID			
	No.	Conv. Validity	No.	Disc. Validity	No.	Conv. Validity	No.	Disc. Validity	No.	Conv. Validity	No.	Disc. Validity	No.	Conv. Validity	No.	Disc. Validity
Flanker	144	−0.52[a]	193	0.53[a]	52	−0.56[a]	78	0.48[a]	36	−0.44[b]	48	0.49[a]	56	−0.44[a]	67	0.56[a]
Flanker DEXT	29	−0.26	61	0.36[b]	7	0.04	15	0.51	15	−0.31	32	0.50[b]	7	−0.23	14	0.34
DCCS[c]	109	0.48[a]	151	0.46[a]	34	0.55[a]	55	0.56[a]	30	0.33	39	0.07	45	0.36[d]	58	0.54[a]
DCCS DEXT	43	0.42[b]	113	0.37[a]	16	0.34	48	0.45[b]	11	0.33	40	0.43[b]	16	0.48	25	0.11
LS	165	0.65[a]	164	0.49[a]	57	0.54[a]	57	0.47[a]	46	0.50[a]	45	0.38[d]	62	0.66[a]	62	0.52[a]
PC	171	0.66[a]	178	0.45[a]	54	0.60[a]	59	0.58[a]	56	0.48[a]	57	0.42[b]	61	0.69[a]	62	0.36[b]
PSM	175	0.47[a]	179	0.50[a]	69	0.34[b]	71	0.50[a]	49	0.64[a]	51	0.52[a]	57	0.33[d]	58	0.33[d]
PVT	235	0.83[a]	232	0.47[a]	88	0.75[a]	86	0.54[a]	71	0.85[a]	70	0.50[a]	76	0.88[a]	76	0.53[a]
OR	230	0.92[a]	227	0.58[a]	86	0.92[a]	84	0.62[a]	70	0.89[a]	69	0.61[a]	74	0.95[a]	74	0.59[a]
FC	151	0.60[a]	152	0.61[a]	53	0.49[a]	53	0.43[b]	40	0.47[b]	40	0.57[a]	58	0.47[a]	76	0.71[a]
CC	237	0.75[a]	237	0.68[a]	89	0.68[a]	89	0.64[a]	72	0.80[a]	73	0.68[a]	59	0.55[a]	75	0.64[a]

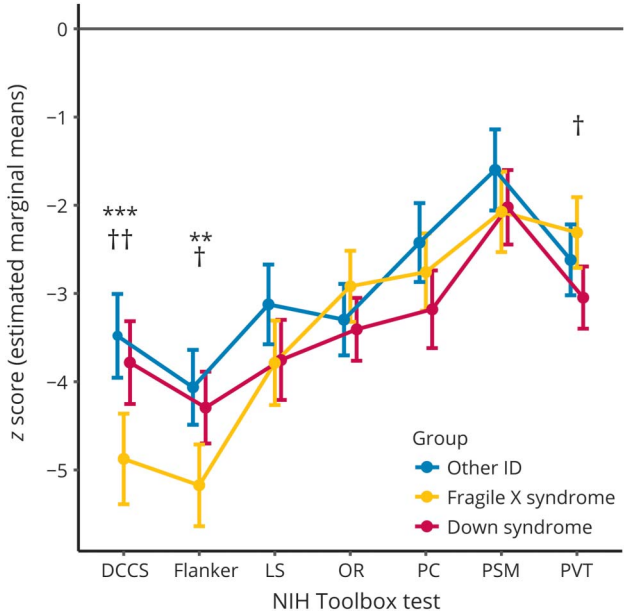
Abbreviations: CC = Crystallized Composite; Conv. = convergent; DCCS = Dimensional Change Card Sort; DEXT = Developmental Extension; Disc. = discriminant; FC = Fluid Composite; Flanker = Flanker Inhibitory Control and Attention; ID = intellectual disability; LS = List Sorting Working Memory; NEPSY-In = NEPSY Inhibition subtest; NIHTB-CB = NIH Toolbox Cognitive Battery; OR = Oral Reading and Recognition; PC = Pattern Comparison Processing Speed; PSM = Picture Sequence Memory; PVT = Picture Vocabulary.
Convergent, discriminant measure for each test: Flanker, Conners Kiddie Continuous Performance Test, 2nd Edition (KCPT2), Woodcock Johnson 4th Edition Letter-Word Identification (WJ-LW); Flanker DEXT, KCPT2, WJ-LW; DCCS, NEPSY-In, WJ-LW; DCCS DEXT, NEPSY-In, WJ-LW; LS, Stanford-Binet, 5th edition (SB-5) Verbal Working Memory, WJ-LW; PC, Wechsler Preschool and Primary Scale of Intelligence, 4th Edition Bug Search, WJ-LW; PSM, Leiter International Performance Scale, 3rd Edition Forward Memory (Leiter-FM), WJ-LW; PVT, Peabody Picture Vocabulary Test, 4th edition, Leiter-FM; OR, WJ-LW, Leiter-FM; Fluid Composite, SB-5 Fluid Reasoning IQ, SB-5 Verbal IQ; Crystallized Composite, SB-5 Verbal IQ, SB-5 Fluid Reasoning IQ.
[a] $p < 0.001$.
[b] $p < 0.01$.
[c] When the DCCS low subgroup (n = 26) was removed, in fragile X syndrome, the convergent correlation was significant (n = 21, $r = 0.49$, $p = 0.02$).
[d] $p < 0.05$.

Table 5 Ecologic validity (Pearson *r*)

	Chronologic age	Mental age	FSIQ	VABS-3 ABC
Flanker	0.25[a]	0.58[a]	0.50[a]	0.18[b]
Flanker DEXT	0.25	0.37[c]	0.18	0.10
DCCS	0.30[a]	0.63[a]	0.56[a]	0.14
DCCS DEXT	0.14	0.66[a]	0.49[a]	0.17
LS	0.20[b]	0.70[a]	0.66[a]	0.16[b]
PC	0.10	0.59[a]	0.57[a]	0.19[b]
PSM	0.04	0.58[a]	0.58[a]	0.27[a]
PVT	0.35[a]	0.76[a]	0.74[a]	0.39[a]
OR	0.17[c]	0.68[a]	0.69[a]	0.39[a]
FC	—	—	0.66[a]	0.15
CC	—	—	0.76[a]	0.44[a]
CFC	—	—	0.78[a]	0.25[c]

Abbreviations: CC = Crystallized Composite; CFC = Cognitive Function Composite; DCCS = Dimensional Change Card Sort; DEXT = Developmental Extension; FC = Fluid Composite; Flanker = Flanker Inhibitory Control and Attention; FSIQ = full-scale IQ; LS = List Sorting Working Memory; OR = Oral Reading and Recognition; PC = Pattern Comparison Processing Speed; PSM = Picture Sequence Memory; PVT = Picture Vocabulary; VABS-3 ABC = Vineland-3 Adaptive Behavior Composite.
[a] $p < 0.001$.
[b] $p < 0.05$.
[c] $p < 0.01$.

Figure 2 Profile plot of NIH Toolbox Cognitive Battery test *z* scores by group



The *z* scores on each test (representing group performance relative to the general population average performance) are shown, derived from the mixed-model analysis of variance, adjusted for IQ. A *z* score of 0 (horizontal line at top) represents the average performance in the general population normative sample. The *z* scores <0 represent the number of SDs below the general population average for the chronologic age band. Error bars represent 95% confidence intervals. *Comparison between fragile X syndrome (FXS) and intellectual disability (ID) of other or unknown cause. †Comparison between FXS and Down syndrome. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; †$p < 0.05$; ††$p < 0.01$; †††$p < 0.001$. DCCS = Dimensional Change Card Sort; LS = List Sorting Working Memory; OR = Oral Reading and Recognition; PC = Pattern Comparison Processing Speed; PSM = Picture Sequence Memory; PVT = Picture Vocabulary.

standard scores). These findings provide further evidence that test developers should consider and address the range and sensitivity of tests and scores for individuals with moderate to severe ID.[24,36]

This study has some important limitations that warrant consideration. Construct validity challenges are inherent with the ID population in that fully adequate convergent validity measures are not always available or they present their own feasibility and psychometric limitations. The lack of clear discriminant validity in most measures is another challenge. While discriminant correlations are generally desired to be markedly lower than the convergent correlations, in early development, domains of cognition (especially EFs) are thought to be unidimensional, with increasing differentiation of constructs occurring through early adulthood.[37] In ID, there is likely even less differentiation between domains than in typically developing children. Our discriminant validity results are similar to normative 3- to 6-year-old results.[38] Therefore, the Fluid Composite may be a good outcome measure choice on the basis of its psychometric performance and some limitations of subtest construct differentiation in this population. The study was also limited by sample size in evaluations of group-specific results. Future work with larger samples should provide more clarity about reliability and validity within individual ID subgroups.

The NIHTB-CB is a promising outcome measure for ID clinical trials and for many types of nonintervention observational studies. Although not originally intended for clinical use or for special education purposes, ongoing and future research may be done to explore such applications[23] such as in the school psychology setting for accurate and feasible assessment of students with IDs. In addition, the Supplemental Manual developed from this study provides key guidance to examiners and researchers working with ID populations; the procedures on administration, test environment, fidelity, and scoring not only improve examiner familiarity and comfort but, more important, extend accessibility of the NIHTB-CB to more cognitively impaired or behaviorally challenged individuals.

Besides evaluating the NIHTB-CB as an appropriate assessment for ID in general, the present results demonstrate the sensitivity of the battery to known syndrome-specific cognitive phenotypes, such as the impairment of EFs in FXS relative to other ID and to DS. A critical remaining question is the degree to which the battery is sensitive to change, especially to effects of intervention. As an initial test of sensitivity to change, we are currently collecting longitudinal data from

study participants to explore natural developmental changes within each NIHTB-CB test and the composites relative to measures already established as change sensitive such as the SB-5 and Vineland. The present results warrant the next step of evaluating the NIHTB-CB in ID for individuals down to a mental age of 5 years to demonstrate the treatment-specific sensitivity of the battery and to determine the degree to which measure gains reflect functional improvements in daily life. Below a mental age of 5 years, the NIHTB-CB performs more variably, and adaptations to the lower test ranges or scoring adaptations are needed for some measurement domains. Studies of the performance of the battery in older adults with ID are needed, especially focusing on those experiencing cognitive decline or dementia. Overall, the present validation results represent an important step toward providing an objective, scalable, and standardized method for successfully measuring cognition and tracking cognitive changes in ID.

Acknowledgment

The authors extend their appreciation to the families who gave their time and effort to participate in the study and to Leonard Abbeduto, LeAnn Baer, Ruth McClure Barnes, Kyle Bersted, Mikayla Brown, Ana Candelaria, Erin Carmody, Darian Crowley, Suzanne Delap, Randi Hagerman, Anne Hoffmann, Londi Howard, Paige Landau, Caroline Leonczyk, Michael Nelson, Jacklyn Perales, Lacey Pomerantz, Shanelle Rodriguez, Melanie Rothfuss, Ryan Shickman, Andrea Schneider, Haleigh Scott, Laurel Snider, Rachel Teune, Talia Thompson, Denny Tran, and Jamie Woods for their contributions to the study.

Study funding

Funding provided by the National Institute of Child Health and Human Development (R01HD076189), Health and Human Services Administration of Developmental Disabilities (90DD0596), the MIND Institute Intellectual and Developmental Disabilities Research Center (U54 HD079125), and the National Center for Advancing Translational Sciences, NIH, through grant UL1 TR000002.

Disclosure

R. Shields, A. Kaat, F. McKenzie, A. Drayton, S. Sansone, J. Coleman, and C. Michalak report no disclosures. K. Riley has received compensation for consulting to Novartis regarding FXS clinical trials. E. Berry-Kravis has received funding (all funding including consulting goes to Rush University Medical Center) from Seaside Therapeutics, Novartis, Roche, Alcobra, Neuren, Cydan, Fulcrum, GW, Neurotrope, Marinus, Acadia, Zynerba, BioMarin, Ovid, Acadia, Yamo, Ionis, Lumos, Ultragenyx, Vtesse, Sucampo, and Mallinkcrodt Pharmaceuticals to consult on trial design or development strategies and/or to conduct clinical trials in FXS or other neurogenetic disorders, and from Asuragen Inc to develop testing standards and to do validation testing for *FMR1* and *SMN* testing. R. Gershon and K. Widaman report no disclosures. D. Hessl has received funding for consulting from Novartis, Roche, Zynerba, Autifony, and Ovid pharmaceutical companies regarding FXS clinical trials. Go to Neurology.org/N for full disclosures.

Publication history

Received by *Neurology* August 14, 2019. Accepted in final form October 31, 2019.

Appendix Authors

Name	Location	Contribution
Rebecca H. Shields, MS	MIND Institute, Sacramento, CA	Authored the manuscript, performed the statistical analysis, and coordinated the study
Aaron J. Kaat, PhD	Northwestern University, Chicago, IL	Directed NIHTB-CB activities for the study, advised on protocol and analysis, and authored portions of the manuscript
Forrest J. McKenzie, BS	MIND Institute, Sacramento, CA	Authored portions of the manuscript
Andrea Drayton, BS	MIND Institute, Sacramento, CA	Authored portions of the manuscript
Stephanie M. Sansone, PhD	MIND Institute, Sacramento, CA	Assisted in developing protocol and coordinated the study
Jeanine Coleman, PhD	University of Denver, CO	Coordinated assessments at the University of Denver and critically reviewed the manuscript
Claire Michalak, BS	Rush University Medical Center, Chicago, IL	Coordinated assessments and administered NIHTB-CB at Rush University
Richard C. Gershon, PhD	Northwestern University, Chicago, IL	PI and director of the NIH Toolbox and critically reviewed the manuscript
Karen Riley, PhD	University of Denver, CO	PI at the University of Denver and critically reviewed the manuscript
Elizabeth Berry-Kravis, MD, PhD	Rush University Medical Center, Chicago, IL	PI at Rush University and critically reviewed the manuscript
Keith F. Widaman, PhD	University of California, Riverside	Supervised analysis plan, performed the statistical analysis, and critically reviewed the manuscript
David Hessl, PhD	MIND Institute, Sacramento, CA	Designed the study, obtained funding, directed the multisite study, and authored portions of the manuscript

References

1. Olusanya BO, Davis AC, Wertlieb D, et al. Developmental disabilities among children younger than 5 years in 195 countries and territories, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Glob Health 2018;6:e1100–e1121.
2. Lee AW, Ventola P, Budimirovic D, Berry-Kravis E, Visootsak J. Clinical development of targeted fragile X syndrome treatments: an industry perspective. Brain Sci 2018;8:E214.
3. Picker JD, Walsh CA. New innovations: therapeutic opportunities for intellectual disabilities. Ann Neurol 2013;74:382–390.
4. Budimirovic DB, Berry-Kravis E, Erickson CA, et al. Updated report on tools to measure outcomes of clinical trials in fragile X syndrome. J Neurodev Disord 2017;9:14.
5. Esbensen AJ, Hooper SR, Fidler D, et al. Outcome measures for clinical trials in Down syndrome. Am J Intellect Dev Disabil 2017;122:247–281.

6. Gershon RC, Wagster MV, Hendrie HC, Fox NA, Cook KF, Nowinski CJ. NIH Toolbox for assessment of neurological and behavioral function. Neurology 2013;80:S2–S6.
7. Weintraub S, Bauer PJ, Zelazo PD, et al. I, NIH toolbox cognition battery (CB): introduction and pediatric data. Monogr Soc Res Child Dev 2013;78:1–15.
8. Casaletto KB, Umlauf A, Beaumont J, et al. Demographically corrected normative standards for the English version of the NIH Toolbox Cognition Battery. J Int Neuropsychol Soc 2015;21:378–391.
9. Hessl D, Sansone SM, Berry-Kravis E, et al. The NIH Toolbox Cognitive Battery for intellectual disabilities: three preliminary studies and future directions. J Neurodev Disord 2016;8:35.
10. Berry-Kravis E, Lindemann L, Jonch AE, et al. Drug development for neurodevelopmental disorders: lessons learned from fragile X syndrome. Nat Rev Drug Discov 2018;17:280–299.
11. Parker SE, Mai CT, Canfield MA, et al. Updated national birth prevalence estimates for selected birth defects in the United States, 2004-2006. Birth Defects Res A Clin Mol Teratol 2010;88:1008–1016.
12. Hunter J, Rivero-Arias O, Angelov A, Kim E, Fotheringham I, Leal J. Epidemiology of fragile X syndrome: a systematic review and meta-analysis. Am J Med Genet A 2014; 164A:1648–1658.
13. Roid G, Miller L. Stanford-Binet Intelligence Scales, 5th ed. San Antonio: Pearson; 2003.
14. Sparrow S, Cicchetti D, Saulnier C. Vineland Adaptive Behavior Scales, 3rd ed. San Antonio: Pearson; 2016.
15. Weintraub S, Dikmen SS, Heaton RK, et al. Cognition assessment using the NIH Toolbox. Neurology 2013;80:S54–S64.
16. Korkman M, Kirk U, Kemp S. NEPSY-II: A Developmental Neuropsychological Assessment. San Antonio: Pearson; 2007.
17. Conners C. Conners Kiddie Continuous Performance Test, 2nd ed. Toronto: Multi-Health Systems Inc.; 2015.
18. Wechsler D. Wechsler Preschool and Primary Scale of Intelligence, 4th ed. San Antonio: Pearson; 2012.
19. Roid G, Miller L, Pomplum M, Koch C. Leiter International Performance Scale, 3rd ed. Wood Dale: Stoelting Co; 2013.
20. Dunn L, Dunn D. Peabody Picture Vocabulary Test. 4th ed. San Antonio: Pearson; 2007.
21. Woodcock R, Johnston M. Woodcock-Johnson Tests of Achievement, 4th ed. Itasca: Riverside Publishing; 2014.
22. McKenzie FJ, Drayton A, Shields R, et al. NIH Toolbox Cognitive Battery supplemental administrator's manual for intellectual and developmental disabilities [online]. Available at: nihtoolbox.desk.com/customer/portal/articles/2981718. Accessed September 30, 2019.
23. Thompson T, Coleman JM, Riley K, et al. Standardized assessment accommodations for individuals with intellectual disability. Contemp Sch Psychol 2018;22:443–457.
24. Sansone SM, Schneider A, Bickel E, Berry-Kravis E, Prescott C, Hessl D. Improving IQ measurement in intellectual disabilities using true deviation from population norms. J Neurodev Disord 2014;6:16.
25. Akshoomoff N, Beaumont JL, Bauer PJ, et al. VIII. NIH Toolbox Cognition Battery (CB): composite scores of crystallized, fluid, and overall cognition. Monogr Soc Res Child Dev 2013;78:119–132.
26. Zelazo P, Bauer P. National Institutes of Health Toolbox cognition battery (NIH Toolbox CB): validation for children between 3 and 15 years. Monogr Soc Res Child Dev 2013;78:1–172.
27. Hooper SR, Hatton D, Sideris J, et al. Executive functions in young males with fragile X syndrome in comparison to mental age-matched controls: baseline findings from a longitudinal study. Neuropsychology 2008;22:36–47.
28. Wilding J, Cornish K, Munir F. Further delineation of the executive deficit in males with fragile-X syndrome. Neuropsychologia 2002;40:1343–1349.
29. Pennington BF, Moon J, Edgin J, Stedron J, Nadel L. The neuropsychology of Down syndrome: evidence for hippocampal dysfunction. Child Dev 2003;74:75–93.
30. Vicari S, Bellucci S, Carlesimo GA. Implicit and explicit memory: a functional dissociation in persons with Down syndrome. Neuropsychologia 2000;38:240–251.
31. Wechsler D. Weschler Intelligence Scale for Children, 5th ed. Bloomington: Pearson; 2014.
32. Abbeduto L, Murphy MM, Cawthon SW, et al. Receptive language skills of adolescents and young adults with down or fragile X syndrome. Am J Ment Retard 2003; 108:149–160.
33. Godfrey M, Lee NR. Memory profiles in Down syndrome across development: a review of memory abilities through the lifespan. J Neurodev Disord 2018;10:5.
34. Dykens EM, Hodapp RM, Leckman JF. Strengths and weaknesses in the intellectual functioning of males with fragile X syndrome. Am J Ment Defic 1987;92:234–236.
35. Carlozzi NE, Tulsky DS, Kail RV, Beaumont JL. VI, NIH Toolbox Cognition Battery (CB): measuring processing speed. Monogr Soc Res Child Dev 2013;78:88–102.
36. Hessl D, Nguyen DV, Green C, et al. A solution to limitations of cognitive testing in children with intellectual disabilities: the case of fragile X syndrome. J Neurodev Disord 2009;1:33–45.
37. Wiebe SA, Espy KA, Charak D. Using confirmatory factor analysis to understand executive control in preschool children: I, latent structure. Dev Psychol 2008;44: 575–587.
38. Mungas D, Widaman K, Zelazo PD, et al. VII, NIH Toolbox Cognition Battery (CB): factor structure for 3 to 15 year olds. Monogr Soc Res Child Dev 2013;78:103–118.