American Journal on Intellectual and Developmental Disabilities

NIH Toolbox Cognition Battery feasibility in individuals with Williams syndrome

--Manuscript Draft--

Manuscript Number:	AJIDD-D-21-00073R1
Article Type:	Research Report
Keywords:	Williams syndrome; Intellectual Disability; executive function; cognition; NIH Toolbox
Corresponding Author:	Emma Elizabeth Condy, Ph.D. National Institute of Mental Health and Neurosciences: National Institute of Mental Health and Neuro Sciences Bethesda, MD UNITED STATES
First Author:	Emma E Condy, Ph.D.
Order of Authors:	Emma E Condy, Ph.D.
	Lindsey Becker, B.A.
	Cristan Farmer, Ph.D.
	Aaron J Kaat, Ph.D.
	Colby Chlebowski, Ph.D.
	Beth A Kozel, M.D., Ph.D.
	Audrey Thurm, Ph.D.
Manuscript Region of Origin:	UNITED STATES
Abstract:	The NIH Toolbox Cognition Battery (NIHTB-CB) was developed for epidemiological and longitudinal studies across a wide age span. Such a tool may be useful for intervention trials in conditions characterized by intellectual disability (ID), such as Williams syndrome (WS). Three NIHTB-CB tasks, including two executive functioning (Flanker, Dimensional Change Card Sort) and one episodic memory (Picture Sequence Memory) task, were given to 47 individuals with WS, ages 4 to 50, to evaluate feasibility (i.e., proportion of valid administrations) in this population. Findings indicated that NIHTB-CB tests showed good feasibility. Flanker and DCCS age-corrected scores were negatively correlated with age and showed floor effects, indicating these scores may not be useful for quantifying performance on these NIHTB-CB tests in ID.

Abstract

The NIH Toolbox Cognition Battery (NIHTB-CB) was developed for epidemiological and longitudinal studies across a wide age span. Such a tool may be useful for intervention trials in conditions characterized by intellectual disability (ID), such as Williams syndrome (WS). Three NIHTB-CB tasks, including two executive functioning (Flanker, Dimensional Change Card Sort) and one episodic memory (Picture Sequence Memory) task, were given to 47 individuals with WS, ages 4 to 50, to evaluate feasibility (i.e., proportion of valid administrations) in this population. Findings indicated that NIHTB-CB tests showed good feasibility. Flanker and DCCS age-corrected scores were negatively correlated with age and showed floor effects, indicating these scores may not be useful for quantifying performance on these NIHTB-CB tests in ID.

NIH Toolbox Cognition Battery feasibility in individuals with Williams syndrome

Williams syndrome (WS) is a medical condition caused by a microdeletion in chromosome 7q11.23 (OMIM # 194050) and is associated with a variety of medical complications (e.g., cardiovascular, digestive, visual issues), in addition to a high prevalence of intellectual disability, such that 75% of the population have an IQ<70 (ID; Donnai & Karmiloff-Smith, 2000; Kozel et al., 2021). Specific behavioral and psychiatric problems have also been highlighted in WS, with attention deficit hyperactivity disorder a common comorbid psychiatric diagnosis (Leyfer et al., 2006); manifestations include issues with attentional control (Breckenridge et al., 2013) and attentional disengagement (Greer et al., 2017; Lense et al., 2011). The attentional problems seen in WS relate to executive functioning (EF), a set of cognitive skills that allow for the planning and control of behavior. Studies examining whether EF, as measured through performance-based assays (e.g., the Wisconsin Card Sort Task, the Flanker task), is related to other meaningful clinical outcomes in ID are limited, in part due to a lack of appropriate EF measures in this population.

Cognitive Function in Williams syndrome and ID

Performance-based measures of EF are traditionally lab-based tasks, including those that require maintenance and shifting between various rules (e.g., Flanker, Digit Span, Wisconsin Card Sort). These tasks have often been study-specific, lacking standardization in administration modality (e.g., computerized versus paper tests), scoring, and other task parameters (e.g., stimulus timing, trial number), thus requiring the use of control groups when assessing clinical populations of interest. Without normative data, it is challenging for clinicians to contextualize task performance and make conclusions regarding an individual's or even a group's skill level (Lee et al., 2016).

Within WS, the Cambridge Neuropsychological Test Automated Battery, a performance-based EF test, was used to demonstrate impaired EF in individuals aged 11-29 with WS, relative to verbal-ability matched controls. Impairments were observed in shifting, working memory, and planning ability (Rhodes et al., 2010), indicating EF deficits in WS may exceed those expected based on verbal ability. An exaggerated deficit in EF relative to other types of cognitive skills is important to consider in future therapeutic trials in ID, as genetic, pharmacological, and behavioral therapies for conditions associated with ID are under development (Berry-Kravis et al., 2006; de la Torre et al., 2016; Hessl et al., 2019; Protic et al., 2019; Spiridigliozzi et al., 2016). However, many studies operationalize EF as subscales from IQ tests, parent reports, or other standardized assessments that are time consuming or not intended to measure change over time, which threatens the validity of their findings. The utility of performance-based EF tasks in this arena is evidenced by their use in the developmental literature (Carlson, 2005) and, with proper standardization, the structure of performance-based EF tasks can lend them to computerized formats which allow for sensitivity to subtle changes (such as millisecond-level precision) over time (Best & Miller, 2010). It is evident that alternative methods for measuring EF, including existing and relatively short computer-based standardized measures such as the NIH Toolbox, may be valuable and feasible in people with ID (Hessl et al., 2016), especially if attentional issues may make longer tests difficult to complete.

The NIH Toolbox

The NIH Toolbox for the Assessment of Neurological and Behavioral Function was conceptualized by the NIH Blueprint for Neuroscience Research (www.neuroscienceblueprint.nih.gov) to create measures that would fulfill the needs of large epidemiological studies, longitudinal studies, and clinical trials and to provide a common metric

assessing neurological and behavioral constructs. Administered through an iPad, the NIH Toolbox batteries were developed with the following goals in mind: 1. measures are brief and easy to administer, reducing the burden on participants and researchers in large cohort studies; 2. measures are continuous across the 3 to 85 age range, making them uniquely fitted for use in longitudinal studies; 3. measures are normed based on age, as well as demographics (education, sex assigned at birth, and race/ethnicity for Fully Corrected T-Scores), providing norm-referenced scores for performance on the tasks; and 4. measures are sensitive to change over time, allowing repeated administration in multi-visit studies and clinical trials (Gershon et al., 2013).

The NIH Toolbox Cognition Battery (NIHTB-CB) consists of seven tasks measuring different domains of cognitive function, including EF, memory, processing speed, and language (Weintraub et al., 2014; Zelazo et al., 2013). The battery was normed in a demographically diverse sample of 4,859 participants ages 3 to 85, with eligibility criteria including the capability of following test instructions and adequate visual and auditory functioning (Beaumont et al., 2013). Four of the seven cognitive tests can be consistently administered across all ages. The Dimensional Card Change Sort Test and the Flanker Inhibitory Control and Attention Test measure aspects of EF, specifically cognitive flexibility (shifting) and inhibitory control/attention (inhibiting) respectively. The Picture Sequence Memory Test measures episodic memory (Gershon et al., 2013), an aspect of long-term memory that evidence thus far indicates may be relatively preserved in ID (Lifshitz et al., 2011). And lastly, the Picture Vocabulary Test measures receptive vocabulary using a computer adaptive format (Gershon et al., 2014). Due to the large age span for these tests, these are viable candidates for longitudinal studies.

Use of the NIH Toolbox in ID

Though the NIH Toolbox was developed to assess individuals across the "range of normal function," interest in expanding use of the NIHTB-CB in individuals with ID has been growing, including its use as a potentially meaningful measure of change in treatment trials. However, a particular challenge for measuring change in ID is the potential for floor effects (Hessl et al., 2009; Sansone et al., 2014), which may lead to a mischaracterization of a potentially meaningful change over time in individuals with ID (Bishop et al., 2015; Thurm et al., 2020). Additionally, many assessments are not appropriate for repeated administrations due to practice effects. The NIHTB-CB, with norms in a wide age range, computerized adaptive format, and standardized administration could potentially aid in characterizing the cognitive profile of individuals with ID. Efforts to validate the measure in individuals with ID indicate good feasibility in the moderate ID range and a mental age above 5 years (Shields et al., 2020) and convergent validity comparable to the original NIHTB-CB norming study (Hessl et al., 2016). A floor effect was noted for age-corrected standard scoring of the measures, which was remedied through a rescoring procedure. However, the NIHTB-CB also provides uncorrected standard scores, serving as a metric more akin to a person ability score, that measure absolute performance and are argued to be a more appropriate metric in ID (Farmer et al., 2020). Additionally, a recent study using the NIHTB-CB in individuals with autism spectrum disorder found administration was feasible, though test completion rates were lower in those with below average IQ compared to average (83% vs. 96% for Flanker, 61% vs. 94% for DCCS, and 48% vs. 69% for PSMT; Jones et al., 2021). Performance was associated with full scale IQ and exhibited an effect of diagnosis on certain subscales (Flanker and Pattern Comparison Processing Speed), indicating its potential for uncovering specific cognitive deficits between populations (Jones et al., 2021). Replicating these findings in additional genetic disorders associated with ID will be helpful in

establishing the utility of the NIHTB-CB for use as an outcome measure in future intervention work.

Aims and Hypotheses

The aims of the present study were to assess the feasibility of administering the core NIHTB-CB tasks (Flanker Inhibitory Control and Attention Test (Flanker), Dimensional Change Card Sort Test (DCCS), and Picture Sequence Memory Test (PSMT)) and to investigate whether performance on these measures was related to other cognitive variables in individuals with WS and mild-to-moderate ID. Although the Picture Vocabulary Test is part of the core cognitive battery, the feasibility of similar measures of receptive vocabulary has already been established (Rossi & Giacheti, 2017). Therefore, we prioritized the domains for which feasibility data are lacking, specifically EF and memory. Based on previous studies showing that people with ID can complete the NIHTB-CB (Hessl et al., 2009; Shields et al., 2020), we expect the proportion of valid administrations to be comparable to the initial Toolbox validation studies (83.7% Flanker, 79.8% DCCS, 98.1% PSMT; Zelazo & Bauer, 2013). Additionally, to explore the validity of the NIHTB-CB tasks in WS, scores from the NIHTB-CB will be correlated with nonverbal IQ, a measure of fluid cognition which should be positively correlated with EF ability (Blair, 2006; Zelazo et al., 2013), and not participant age. Furthermore, since individuals with WS have been shown to exhibit deficits in EF, we expect poorer performance on the Flanker and DCCS compared to the PSMT.

Method

Participants

This study included 47 participants with WS that attempted the three NIHTB-CB tests. The sample ranged in age from 4-50 years (*Mdn*=20.18, *IQR*=10.46-27.12) and consisted of 25

males and 22 females. The sample consisted of white (82.98%), multiple race (8.51%), and native Hawaiian/Pacific Islander (2.13%) individuals, with 6.38% not reported. Median FSIQ of the sample was 58 (IQR = 52-65.75; 1 missing) and NVIQ was 60 (IQR= 55.25-65; 1 missing), with 80% of the sample (37/46) showing FSIQ estimates below 70.

Procedures

Participants were referred from an ongoing natural history study of WS and WS-like conditions at the NIH ([REDACTED]). All 47 participants in the present analyses have molecular testing showing, at minimum, the equivalent of fluorescent in situ hybridization (FISH) positivity for WS and were enrolled in a study focused on their neurodevelopment. Of those, 44 of 47 participants had additional deletion size testing, 43 of whom show the typical WS deletion size. The remaining individual with deletion size testing had a slightly larger deletion. Consent was obtained by participants or caregivers/guardians, depending on status as a minor and cognitive capacity, into an [REDACTED] IRB approved protocol ([REDACTED]). The NIHTB-CB was administered during an assessment that included a battery of cognitive tests, including IQ testing commensurate with age, selected from the Wechsler Adult Intelligence Scale (WAIS (Wechsler, 2008)), Wechsler Intelligence Scale for Children (WISC (Wechsler, 2014)), Wechsler Preschool & Primary Scale of Intelligence (WPPSI (Wechsler, 2012)), or Kaufman Brief Intelligence Test (KBIT (Kaufman & Kaufman, 2004)). The NIHTB-CB was administered after the IQ testing.

The NIH Toolbox Cognition Battery

The NIHTB-CB consists of specific subtests in the NIH Toolbox that measure cognition, administered through an iPad. Three NIHTB-CB tests were administered: two tests of EF (the Flanker (EF-inhibitory control and attention) and the Dimensional Change Card Sort (DCCS;

EF-cognitive flexibility) (Zelazo et al., 2013)), and the Picture Sequence Memory Test (PSMT; episodic memory) (Bauer et al., 2013). The NIHTB-CB tasks were administered in the following order: the Flanker, DCCS, and PSMT. The experimental version of the Speeded Matching test was also administered after these tests for some participants; however, scores from this test were not analyzed in the present study as the test is still under development.

Per the NIH Toolbox manual, subtest version was determined based on their chronological age. Specifically, the Flanker and DCCS both contain versions for 3-7, 8-11, and 12+ years old and the PSMT contains versions for 3-4, 5-6, 7, and 8+ years old. These versions are designed to assess the same construct, but use slightly different stimuli, instructions, and prompts throughout the test (e.g., while the 3-7 and 8-11 version of the DCCS contain both an auditory and written prompt for the words "SHAPE" and "COLOR" between trials, the 12+ version only contains the written prompt). Tests on the NIHTB-CB were then administered following the standard administration procedures, with some accommodations implemented as needed, based on standards used for testing of people with ID (Thompson et al., 2018), including allowance for minimal redirecting and prompting (e.g., examiner saying "look here" to draw the individual's attention back to the iPad), even on the DCCS for which the manual states that prompting is not allowed. An administration form was filled out for each participant that included notes on unanticipated interruptions in testing and participant compliance and engagement. Participants who were unable to complete the task for reasons unrelated to their compliance (i.e., interruptions in testing, technical difficulties) were removed from the attempted administration total before calculating the proportion of valid administrations of each task. Participants' scores were considered invalid if they did not pass the standard practice trials for a

task, in which case the task was unscorable, or were unable to complete the task for other reasons (i.e., participant refusal, inattention).

The NIHTB-CB tasks produce both uncorrected standard scores and age-corrected standard scores, with means of 100 and standard deviation of 15. Uncorrected standard scores index performance relative to all individuals in the normative sample (ages 3-85 years old), and age-corrected standard scores index performance relative to same-aged peers. The lowest possible age-corrected standard score is 54 as they were winsorized (i.e., extreme values are limited through a transformation) in a Toolbox update in 2018, presumably because extreme scores have lower reliability. The implementation of this "floor" score meant that tests administered before 2018 could have achieved scores below the set floor of 54 but with current scoring would be set to 54 (Shields et al., 2020). With this detail in mind, scores for the age-corrected standard score (SS) were classified as at the floor if they were ≤ 54 in our sample.

Statistical Analysis

The proportion of valid administrations of each task was calculated and compared to the initial NIHTB-CB validation study (Zelazo & Bauer, 2013) at the task level, to determine feasibility of each subtest. In addition, age-corrected SS were examined on each task to determine whether the range of scores in our WS participants was constrained by the winsorization of these scores, creating a floor effect. Finally, correlations between the NIHTB-CB age-corrected SS and age and nonverbal IQ (NVIQ) were examined to assess convergent validity of the Toolbox tasks. Scores were then compared through a one-way ANOVA to determine whether there were differences in performance across subtests in a sample with WS. Analyses were conducted in R Studio (version 3.6.3; R Core Team, 2020)

Results

Rates of successful administrations

Of the 47 participants who attempted testing, 79% (37/47) had valid administrations of all three NIHTB-CB tasks included for analysis. Successful completion of each task individually was also high, with the lowest rate of success on the Flanker (Table 1). The NIHTB-CB Flanker subtest was attempted with 47 participants. Of these, one was considered invalid due to technical difficulties, and one had an interruption during the task and thus were removed from further analysis. Of the remaining 45 participants, 40 (89%) were able to complete the task and receive scores through the NIHTB-CB. Those who were unable to complete the task either could not pass the practice (n = 4) or did not engage with any NIHTB-CB tasks (n = 1). The median FSIQ for those who completed the Flanker task was 59 (*IQR*=52-67); the median FSIQ for the five participants who were unable to complete the task was 52 (*range:* 50-64). The NIHTB-CB DCCS subtest was attempted with 47 participants. Of these, 42 (89%) were able to complete the task and receive scores through the NIHTB-CB. Those who did not complete the DCCS either could not pass the practice (n = 4) or did not engage with any NIHTB-CB tasks (n = 1). The median FSIQ of those who completed the DCCS was 58 (*IQR*=52-65.75); the median FSIQ for the five who were unable to complete the task was 61 (*range:* 46-82). The NIHTB-CB PSMT subtest was attempted with 47 participants with WS. Of these, 42 (89%) were able to complete the task and receive scores through the NIHTB-CB and had a median FSIQ of 59 (*IQR*=52-65.75); the median FSIQ for the five who were unable to complete the task was 55 (*range:* 46-82). Those who did not complete the PSMT either could not pass the practice (n = 4) or did not engage with any NIHTB-CB tasks (n = 1).

Performance and rates of floor scores

Most individuals with WS who completed the tasks had an age-corrected SS that was at least two standard deviations below the mean (≤ 70) on the Flanker (83%; 33/40) and DCCS (76%; 32/42); only 14% (6/42) scored in this lower range on the PSMT. No WS participants scored at the floor on the PSMT; however, 30% (12/40) of Flanker scores were at the floor, as were 21% (9/42) of the DCCS scores.

With approximately one-third of the sample scoring at the floor of age-corrected SS, we explored the role of chronological age (Figure 1 a, b). A scatterplot of age-corrected SS and age shows a floor effect in the age-corrected SS of individuals ages 20-50 on the Flanker and DCCS, evidenced by the settling of age-corrected SS points around 54 compared to the variation in the uncorrected SS.

Comparing performance across NIHTB-CB tests

Uncorrected SS on the Flanker, DCCS, and PSMT were compared within individuals with WS who completed all three tests (n=37) using a nonparametric one-way repeated measures ANOVA. The uncorrected SS was used because the floor effect in the age-corrected scores may artificially minimize differences in performance between the tests, and because within-subject analysis does not require adjustment for age. A Friedman test indicated a difference in performance across these tests, χ^2 (2) = 28.32, p<0.001 (Figure 2). Follow-up Wilcox Signed-Rank tests indicated this difference was driven by differences between the Flanker (*Mdn* = 65, *IQR* = 57-68) and PSMT (*Mdn* = 82, *IQR* = 79-88) (V = 19, p < 0.001) and DCCS (*Mdn* = 56, *IQR* = 51-78) and PSMT (V = 55, p < 0.001); however, the Flanker and DCCS scores did not differ (V = 299.5, p = 0.98).

Convergent validity

NVIQ did not correlate with age-corrected SS on any of the NIHTB-CB tasks (Table 2). There was a large negative relationship between age-corrected SS on the Flanker and age (r_s= -0.82, p<0.01), and a moderate negative relationship between age-corrected SS on the DCCS and age (r_s=-0.45, p<0.01). These correlations indicate that older individuals with WS are receiving lower age-corrected SS on the Flanker and DCCS. Additionally, there was a positive correlation between age-corrected SS on the Flanker and DCCS (r_s=0.58, p<0.01), but none with the age-corrected SS on the PSMT (r_s=0.07, p=0.68).

Discussion

NIHTB-CB scores on the Flanker, DCCS, and PSMT were examined in a sample of individuals with WS that was representative of the overall WS population as 80% had estimated IQs below 70. Overall, the NIHTB-CB tasks showed good feasibility for use in people with WS, with 89% able to complete each task, and the majority able to complete all three tasks (79%). These rates of successful administrations for each task are comparable to the initial validation study of the Toolbox in typically developing children and to those previously reported in individuals with ID (Shields et al., 2020), though higher on the DCCS in the present study (90% vs. 64% in the previous study of individuals with ID). However, in a small sample of individuals with ID and autism spectrum disorder (n = 23) (Jones et al., 2021), rates of successful administration on the PSMT appear notably lower (48%) than in the individuals with WS in the present study (89%). It is possible that the discrepancy is due in part to the differing age ranges of these studies (3-17 in Jones et al. (2021) versus 4-50 in the present study); however, no notable relationship between age and PSMT performance or patterns around age of non-completers was observed in the present study. This discrepancy, in addition to the relatively strong performance on the PSMT for the WS cohort, indicates that it may be worth exploring

differences in the episodic memory task compared to the other tasks across individuals with neurodevelopmental disorders in future studies.

Supporting our hypothesis that people with WS (who generally have ID) exhibit deficits in EF, we observed much higher rates of impairment relative to age-based expectations on the Flanker and DCCS, with 83% (33/40) and 76% (32/42) scoring at least two standard deviations below the mean using the age-corrected standard scores respectively. Performance was better on the episodic memory test (PSMT) than the EF tests (Flanker & DCCS). The discrepant performance between these subtests appears consistent with previous findings using the NIHTB-CB in ID (Hessl et al., 2016).

Uncorrected versus age-corrected standard scores in the NIHTB-CB

Age-corrected standard scores on the NIHTB-CB provide a measure of performance relative to same aged peers, whereas the uncorrected standard score compares an individual's performance to the entire normative sample, regardless of other characteristics, namely age. However, due to the slower rate of increasing cognitive skills in people with ID, compared to their same aged peers, age-corrected standard scores often decrease over time in individuals with ID even though their ability may have actually improved (Farmer et al., 2020). The winsorized floor of the age-corrected standard scores is one point more extreme than three standard deviations from the mean, an unusually narrow scorable range relative to IQ tests, which usually extend to at least four standard deviations. This results in a substantial floor effect, where participants of varying levels of ability are assigned the same score (Figure 1a, b). Further exploration of these scores through scatterplots show that the age-corrected SS floor effect seems to be occurring in the older individuals ages 20-50. Alternatively, NIHTB-CB uncorrected SS minimizes an artificial floor. The NIHTB-CB uncorrected SS approximates a "person ability

score", which indexes an individual's level of functioning relative to the entire normative sample and could potentially be more sensitive to change over time. For developmental constructs, the variability in performance within a narrow age range will be less than that of a sample with a wider age range. Age-corrected standard scores reduce the variability in performance of the reference group, and individuals with ID often have performance that is outside the range observed in their chronological-age peers. This results in floor scores, and unless performance jumps into the range observed in their chronological-aged peers, age-corrected standard scores will remain stagnant over time even though performance may improve. Comparing that performance to a wider age range, such as that used for the uncorrected SS, reduces the frequency of floor scores, and allows for within-subject comparison. The use of uncorrected SS must be undertaken with caution, however, as they are inherently confounded with age. That said, age-corrected SS on the Flanker and DCCS were also confounded with age in this sample. For these reasons, uncorrected SS may be more appropriate for use in a population with ID, particularly if longitudinal studies are being conducted. Furthermore, use of uncorrected SS on these NIHTB-CB tasks should be strongly considered for individuals with ID in the 20-50-year-old range due to the floor effects that seemed to be particularly problematic in this age group. Alternatively, previous studies using the NIHTB-CB in ID have used a z-score deviation scoring approach (Hessl et al., 2009; Shields et al., 2020). It is worth investigating whether this approach yields better information than simply using the uncorrected SS provided by the NIHTB-CB.

Convergent validity concerns

The relationships between NVIQ and the NIHTB-CB tasks were explored to determine whether scores on these tasks show convergent validity. As a measure of fluid intelligence, it is expected that NVIQ correlates with EF ability, such as the Flanker and DCCS (Blair, 2006;

Zelazo et al., 2013). However, NVIQ was not related to age-corrected SS on any of the NIHTB-CB tasks. Instead, the strongest correlations were negative associations with age on the Flanker and the DCCS, which raises concerns about age as a major confound with age-corrected SS. This is common in individuals with ID because often their rate of skill development does not keep up with those of their same-aged peers, resulting in lower scores as they age even though their ability level may have remained stable or even progressed at a slower rate than their peers. Though a correlation between uncorrected SS and age is also likely, this may be appropriate when looking to index ability. Indeed, due to the ongoing development of EF from early childhood into early adulthood, a positive relationship between EF scores and age is to be expected when indexing an ability that develops over an extended period of the lifespan.

Considerations for NIHTB-CB measurement in WS and ID more broadly

In the present study, individuals with WS were retained in the analyses if they were able to complete the practice trials and did not experience technical difficulties or administration disruptions that interfered with administration. The majority of participants who did not complete tasks were unable to pass the practice. Further, administration notes from those who did not pass the practice indicated that attentional issues appeared to be the most common cause, which is unsurprising given the attentional issues widely reported in WS (Breckenridge et al., 2013; Greer et al., 2017; Leyfer et al., 2006). For this reason, it is possible that scores on the tasks could be slightly elevated as participants who were unable to receive scores may have attained low scores; however, the rates of incomplete tests were comparable to those in the initial validation studies, indicating the issue is not necessarily specific to WS or ID.

While further study is required in other ID samples to provide confirmation of how feasibility may differ in other subpopulations of individuals with ID, recent validation efforts of

the NIHTB-CB in ID have generated a manual with administration guidelines in individuals with ID (McKenzie et al., 2019) that uses a similar approach to that taken in the current study of allowing for prompting and redirection, during the task. This is generally important in order to make appropriate testing accommodations for participants with ID (Thompson et al., 2018), although it is noted that administration accommodations such as prompting and redirecting could potentially affect the Flanker and DCCS as their scoring incorporates reaction time into their calculation. However, reaction time is only accounted for in the scoring of these tasks when an accuracy of ≥ 80% is achieved, and thus wouldn't affect scores for those below this threshold. Further investigation of the item-level data and computed scores could potentially reveal whether that is the case in populations with ID. Regardless, the alternative of not using accommodations, and thus lowering feasibility of this measure in people with ID, is problematic as well.

Another consideration when administering the NIHTB-CB to individuals with ID is the difference in instructional modalities for the different versions for each age range. The present study based the task version administered on chronological age, unlike other studies looking at the Toolbox in ID, which have used mental age as estimated through IQ testing. However, in the ages 12+ version of the Flanker and DCCS, there are no longer audio prompts that accompany the word prompts that appear on the screen between each trial ("MIDDLE" and "SHAPE"/"COLOR", respectively). This becomes problematic in an ID population, as it is possible that they may not have the reading ability required for the written prompts. Future studies should carefully consider their participant's reading level when administering different forms of these tasks in participants with ID.

Limitations

A potential limitation of the present study is the use of standard NIH Toolbox procedures of task version selection based on chronological age, which allowed for comparison to the normative sample. This is a potential limitation because previous studies examining the NIHTB-CB in ID used mental age to determine which subtest version was administered (Hessl et al., 2016; Shields et al., 2020). Administration based on mental age may be helpful in ameliorating the concerns regarding reading level previously mentioned but may be untenable because it requires administering a simultaneous IQ test. Based on our findings, selecting tasks based on chronological age did not appear to interfere with feasibility of administration and may be a viable option for those that are unable to assess mental age prior to NIHTB-CB testing.

In addition, the current sample comprised a wide age range. There was limited administration of the NIHTB-CB in certain age ranges, such as the 3–7-year-old versions of the tasks, limiting conclusions about the different task types based on age range. With few in the 3–7-year-old age group for the Flanker and DCCS, it was difficult to investigate the relatively higher age-corrected standard scores in this age range compared to the 12+ age range on the Flanker and DCCS (Figure 1).

Finally, the current study did not include information on a variety of other factors that may be related to differences in completion rates or performance, including co-occurring medical problems, psychiatric diagnoses, and medications.

Conclusion

The current study found good feasibility for use of several subtests from the NIH Toolbox in a wide age-range of people with Williams syndrome, though raise concerns about the validity of using age-corrected SS in individuals with ID. Performance on the PSMT was significantly better than performance on the Flanker and DCCS, which could be related to the relatively

preserved episodic memory seen in WS. Further research focusing on better scoring norms for those scoring in the lower extremes of the Toolbox and a more careful examination of which scores (i.e., age-corrected or uncorrected SS) are appropriate for indexing the abilities of individuals with ID is warranted.

References

Bauer, P. J., Dikmen, S. S., Heaton, R. K., Mungas, D., Slotkin, J., & Beaumont, J. L. (2013, Aug). Iii. Nih toolbox cognition battery (cb): Measuring episodic memory. *Monographs of the Society for Research in Child Development*, *78*(4), 34-48. https://doi.org/10.1111/mono.12033

Beaumont, J. L., Havlik, R., Cook, K. F., Hays, R. D., Wallner-Allen, K., Korper, S. P., Lai, J. S., Nord, C., Zill, N., Choi, S., Yost, K. J., Ustsinovich, V., Brouwers, P., Hoffman, H. J., & Gershon, R. (2013, Mar 12). Norming plans for the nih toolbox. *Neurology*, *80*(11 Suppl 3), S87-92. https://doi.org/10.1212/WNL.0b013e3182872e70

Berry-Kravis, E., Krause, S. E., Block, S. S., Guter, S., Wuu, J., Leurgans, S., Decle, P., Potanos, K., Cook, E., Salt, J., Maino, D., Weinberg, D., Lara, R., Jardini, T., Cogswell, J., Johnson, S. A., & Hagerman, R. (2006, Oct). Effect of cx516, an ampa-modulating compound, on cognition and behavior in fragile x syndrome: A controlled trial. *Journal of Child and Adolescent Psychopharmacology*, *16*(5), 525-540. https://doi.org/DOI 10.1089/cap.2006.16.525

Best, J. R., & Miller, P. H. (2010, Nov-Dec). A developmental perspective on executive function. *Child Dev*, *81*(6), 1641-1660. https://doi.org/10.1111/j.1467-8624.2010.01499.x

Bishop, S. L., Farmer, C., & Thurm, A. (2015, Apr). Measurement of nonverbal iq in autism spectrum disorder: Scores in young adulthood compared to early childhood. *J Autism Dev Disord, 45*(4), 966-974. https://doi.org/10.1007/s10803-014-2250-3

Blair, C. (2006, Apr). How similar are fluid cognition and general intelligence? A developmental neuroscience perspective on fluid cognition as an aspect of human cognitive ability. *Behavioral and Brain Sciences, 29*(2), 109-+. <Go to ISI>://WOS:000237215900001

Breckenridge, K., Braddick, O., Anker, S., Woodhouse, M., & Atkinson, J. (2013, Jun). Attention in williams syndrome and down's syndrome: Performance on the new early childhood attention battery. *British Journal of Developmental Psychology, 31*(2), 257-269. https://doi.org/10.1111/bjdp.12003

Carlson, S. M. (2005). Developmentally sensitive measures of executive function in preschool children. *Dev Neuropsychol, 28*(2), 595-616. https://doi.org/10.1207/s15326942dn2802_3

de la Torre, R., de Sola, S., Hernandez, G., Farre, M., Pujol, J., Rodriguez, J., Espadaler, J. M., Langohr, K., Cuenca-Royo, A., Principe, A., Xicota, L., Janel, N., Catuara-Solarz, S., Sanchez-Benavides, G., Blehaut, H., Duenas-Espin, I., del Hoyo, L., Benejam, B., Blanco-Hinojo, L., Videla, S., Fito, M., Delabar, J. M., Dierssen, M., & Grp, T. S. (2016, Jul). Safety and efficacy of cognitive training plus epigallocatechin-3-gallate in young adults with down's syndrome (tesdad): A double-blind, randomised, placebo-controlled,

phase 2 trial. *Lancet Neurology, 15*(8), 801-810. https://doi.org/Doi 10.1016/S1474-4422(16)30034-5

Donnai, D., & Karmiloff-Smith, A. (2000, Sum). Williams syndrome: From genotype through to the cognitive phenotype. *American Journal of Medical Genetics, 97*(2), 164-171. https://doi.org/Doi 10.1002/1096-8628(200022)97:2<164::Aid-Ajmg8>3.0.Co;2-F

Farmer, C. A., Kaat, A. J., Thurm, A., Anselm, I., Akshoomoff, N., Bennett, A., Berry, L., Bruchey, A., Barshop, B. A., Berry-Kravis, E., Bianconi, S., Cecil, K. M., Davis, R. J., Ficicioglu, C., Porter, F. D., Wainer, A., Goin-Kochel, R. P., Leonczyk, C., Guthrie, W., Koeberl, D., Love-Nichols, J., Mamak, E., Mercimek-Andrews, S., Thomas, R. P., Spiridigliozzi, G. A., Sullivan, N., Sutton, V. R., Udhnani, M. D., Waisbren, S. E., & Miller, J. S. (2020, Nov 1). Person ability scores as an alternative to norm-referenced scores as outcome measures in studies of neurodevelopmental disorders. *Am J Intellect Dev Disabil, 125*(6), 475-480. https://doi.org/10.1352/1944-7558-125.6.475

Gershon, R. C., Cook, K. F., Mungas, D., Manly, J. J., Slotkin, J., Beaumont, J. L., & Weintraub, S. (2014, Jul). Language measures of the nih toolbox cognition battery. *J Int Neuropsychol Soc, 20*(6), 642-651. https://doi.org/10.1017/S1355617714000411

Gershon, R. C., Wagster, M. V., Hendrie, H. C., Fox, N. A., Cook, K. F., & Nowinski, C. J. (2013). Nih toolbox for assessment of neurological and behavioral function. *Neurology, 80*(11 Suppl 3), S2-6. https://doi.org/10.1212/WNL.0b013e3182872e5f

Greer, J. M. H., Hamilton, C., McMullon, M. E. G., Riby, D. M., & Riby, L. M. (2017, May 8). An event related potential study of inhibitory and attentional control in williams syndrome adults (vol 2, e0170180, 2017). *Plos One*, *12*(5). https://doi.org/ARTN e0177587
10.1371/journal.pone.0177587

Hessl, D., Nguyen, D. V., Green, C., Chavez, A., Tassone, F., Hagerman, R. J., Senturk, D., Schneider, A., Lightbody, A., Reiss, A. L., & Hall, S. (2009, Mar). A solution to limitations of cognitive testing in children with intellectual disabilities: The case of fragile x syndrome. *J Neurodev Disord*, *1*(1), 33-45. https://doi.org/10.1007/s11689-008-9001-8

Hessl, D., Sansone, S. M., Berry-Kravis, E., Riley, K., Widaman, K. F., Abbeduto, L., Schneider, A., Coleman, J., Oaklander, D., Rhodes, K. C., & Gershon, R. C. (2016, Sep 6). The nih toolbox cognitive battery for intellectual disabilities: Three preliminary studies and future directions. *Journal of Neurodevelopmental Disorders*, *8*. https://doi.org/ARTN 35
10.1186/s11689-016-9167-4

Hessl, D., Schweitzer, J. B., Nguyen, D. V., McLennan, Y. A., Johnston, C., Shickman, R., & Chen, Y. J. (2019, Apr 15). Cognitive training for children and adolescents with fragile x syndrome: A randomized controlled trial of cogmed. *Journal of Neurodevelopmental Disorders*, *11*. https://doi.org/ARTN 4

10.1186/s11689-019-9264-2

Jones, D. R., Dallman, A., Harrop, C., Whitten, A., Pritchett, J., Lecavalier, L., Bodfish, J. W., & Boyd, B. A. (2021, Mar 24). Evaluating the feasibility of the nih toolbox cognition battery for autistic children and adolescents. *J Autism Dev Disord*. https://doi.org/10.1007/s10803-021-04965-2

Kaufman, A., & Kaufman, N. (2004). *Kaufman brief intelligence test (kbit-2).* Pearson.

Kozel, B. A., Barak, B., Kim, C. A., Mervis, C. B., Osborne, L. R., Porter, M., & Pober, B. R. (2021, Jun 17). Williams syndrome. *Nat Rev Dis Primers, 7*(1), 42. https://doi.org/10.1038/s41572-021-00276-z

Lee, N. R., Maiman, M., & Godfrey, M. (2016). *What can neuropsychology teach us about intellectual disability?: Searching for commonalities in the memory and executive function profiles associated with down, williams, and fragile x syndromes.* (Vol. 51). Academic Press.

Lense, M. D., Key, A. P., & Dykens, E. M. (2011, Nov). Attentional disengagement in adults with williams syndrome. *Brain and Cognition, 77*(2), 201-207. https://doi.org/10.1016/j.bandc.2011.08.008

Leyfer, O. T., Woodruff-Borden, J., Klein-Tasman, B. P., Fricke, J. S., & Mervis, C. B. (2006, Sep 5). Prevalence of psychiatric disorders in 4 to 16-year-olds with williams syndrome. *American Journal of Medical Genetics Part B-Neuropsychiatric Genetics, 141b*(6), 615-622. https://doi.org/10.1002/ajmg.b.30344

Lifshitz, H., Shtein, S., Weiss, I., & Vakil, E. (2011). Meta-analysis of explicit memory studies in populations with intellectual disability. *European Journal of Special Needs Education 26*, 93 - 111.

McKenzie, F. J., Drayton, A., Shields, R., Kaat, A. J., Coleman, J., Thompson, T., Sansone, S., Riley, K., Berry-Kravis, E., Gershon, R., & Hessl, D. (2019). National institutes of health toolbox cognitive battery supplemental administrator's manual for intellectual and developmental disabilities a guide on administration and scoring standards.

Protic, D., Salcedo-Arellano, M. J., Dy, J. B., Potter, L. A., & Hagerman, R. J. (2019). New targeted treatments for fragile x syndrome. *Current Pediatric Reviews, 15*(4), 251-258. https://doi.org/10.2174/1573396315666190625110748

Rhodes, S. M., Riby, D. M., Park, J., Fraser, E., & Campbell, L. E. (2010, Apr). Executive neuropsychological functioning in individuals with williams syndrome. *Neuropsychologia, 48*(5), 1216-1226. https://doi.org/10.1016/j.neuropsychologia.2009.12.021

Rossi, N. F., & Giacheti, C. M. (2017, Jul). Association between speech-language, general cognitive functioning and behaviour problems in individuals with williams syndrome. *J Intellect Disabil Res, 61*(7), 707-718. https://doi.org/10.1111/jir.12388

Sansone, S. M., Schneider, A., Bickel, E., Berry-Kravis, E., Prescott, C., & Hessl, D. (2014). Improving iq measurement in intellectual disabilities using true deviation from population norms. *Journal of Neurodevelopmental Disorders, 6*(1), 16. https://doi.org/10.1186/1866-1955-6-16

Shields, R. H., Kaat, A. J., McKenzie, F. J., Drayton, A., Sansone, S. M., Coleman, J., Michalak, C., Riley, K., Berry-Kravis, E., Gershon, R. C., Widaman, K. F., & Hessl, D. (2020). Validation of the nih toolbox cognitive battery in intellectual disability. *Neurology, 94*(12), E1229-E1240. https://doi.org/10.1212/Wnl.0000000000009131

Spiridigliozzi, G. A., Hart, S. J., Heller, J. H., Schneider, H. E., Baker, J. A., Weadon, C., Capone, G. T., & Kishnani, P. S. (2016, Jun). Safety and efficacy of rivastigmine in children with down syndrome: A double blind placebo controlled trial. *American Journal of Medical Genetics Part A, 170*(6), 1545-1555. https://doi.org/10.1002/ajmg.a.37650

Team, R. C. (2020). *R: A language and environment for statistical computing.* https://www.R-project.org/

Thompson, T., Coleman, J. M., Riley, K., Snider, L. A., Howard, L. J., Sansone, S. M., & Hessl, D. (2018). Standardized assessment accommodations for individuals with intellectual disability. *Contemp Sch Psychol, 22*(4), 443-457. https://doi.org/10.1007/s40688-018-0171-4

Thurm, A., Kelleher, B., & Wheeler, A. (2020). Outcome measures for core symptoms of intellectual disability: State of the field. *American Journal on Intellectual and Developmental Disabilities, 125*(6), 418-433. https://doi.org/10.1352/1944-7558-125.6.418

Wechsler, D. (2008). *Wechsler adult intelligence scale*. Psychological Corporation.

Wechsler, D. (2012). *Wechsler preschool and primary scale of intelligence-fourth edition.* The Psychological Corporation.

Wechsler, D. (2014). *Wisc-v: Technical and interpretive manual.* Pearson.

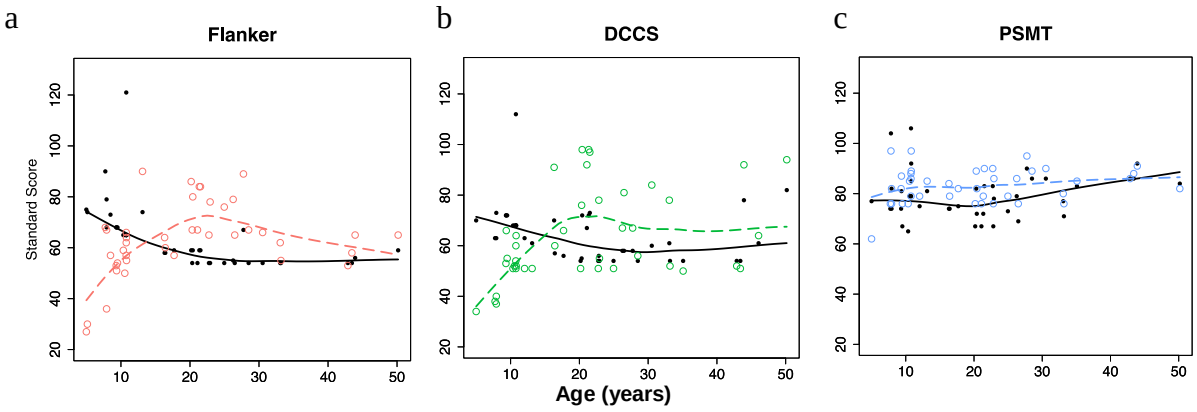
Weintraub, S., Dikmen, S. S., Heaton, R. K., Tulsky, D. S., Zelazo, P. D., Slotkin, J., Carlozzi, N. E., Bauer, P. J., Wallner-Allen, K., Fox, N., Havlik, R., Beaumont, J. L., Mungas, D., Manly, J. J., Moy, C., Conway, K., Edwards, E., Nowinski, C. J., & Gershon, R. (2014, Jul). The cognition battery of the nih toolbox for assessment of neurological and behavioral function: Validation in an adult sample. *J Int Neuropsychol Soc, 20*(6), 567-578. https://doi.org/10.1017/S1355617714000320

Zelazo, P. D., Anderson, J. E., Richler, J., Wallner-Allen, K., Beaumont, J. L., & Weintraub, S. (2013, Aug). Ii. Nih toolbox cognition battery (cb): Measuring executive function and attention. *Monographs of the Society for Research in Child Development, 78*(4), 16-33. https://doi.org/10.1111/mono.12032

Zelazo, P. D., & Bauer, P. J. (2013). National institutes of health toolbox cognition battery (nih toolbox cb): Validation for children between 3 and 15 years. *Monographs of the Society for Research in Child Development, 78*.

Figure 1

Scatterplots of age-corrected and uncorrected on the Flanker, DCCS, and PSMT by age.



Note. Age-corrected SS are shown by black filled dots and the black solid lowess line; uncorrected SS are shown by the unfilled dots and dotted lowess line. DCCS=Dimensional Change Card Sort; PSMT=Picture Sequence Memory Test.

Figure 2

NIHTB-CB test scores across WS individuals who completed all three tests (n=37).

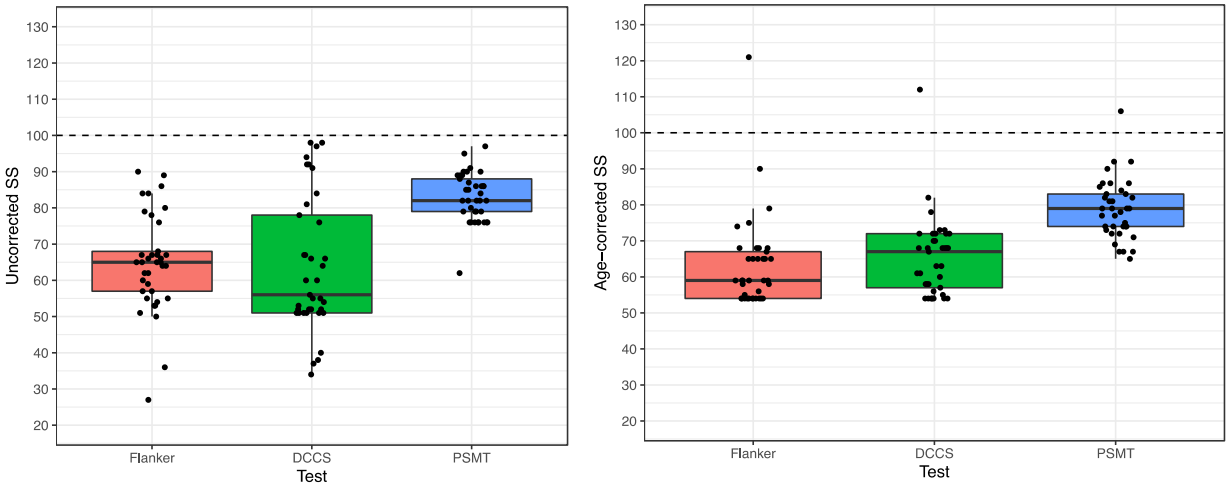


Table 1

Counts of successful administrations within tasks on the NIHTB-CB in WS (N=47).

	Completed above floor	Completed at floor	Did not complete
Flanker	28	12	5*
DCCS	33	9	5
PSMT	42	0	5

Note. DCCS=Dimensional Change Card Sort; PSMT=Picture Sequence Memory Test.

*Two additional participants were removed from the Flanker analysis due to administration issues unrelated to participant compliance.

Table 2
Spearman correlations (r_s) between age-corrected SS on the NIHTB-CB tasks and age and NVIQ.

	Age	NVIQ	Flanker Age-corrected SS	DCCS Age-corrected SS	PMST Age-corrected SS
Age	—				
NVIQ	0.06 (0.68) [-0.26, 0.39]	—-			
Flanker Age-corrected SS	**-0.82 (<0.01) [-0.94, -0.70]**	0.01 (0.93) [-0.35, 0.38]	—		
DCCS Age-corrected SS	**-0.45 (<0.01) [-0.78, -0.12]**	0.27 (0.08) [-0.08, 0.61]	**0.58 (<0.01) [0.34, 0.82]**	—	
PSMT Age-corrected SS	0.20 (0.21) [-0.14, 0.54]	-0.01 (0.94) [-0.35, 0.32]	0.05 (0.79) [-0.31, 0.40]	0.07 (0.68) [-0.30, 0.43]	—

Note. P-values indicated in parentheses and 95% confidence intervals in brackets. Confidence intervals were derived using the *spearmanCI* package in R.