Published in final edited form as:

Res Dev Disabil. 2022 January ; 120: 104147. doi:10.1016/j.ridd.2021.104147.

Short-term memory outcome measures: Psychometric evaluation and performance in youth with Down syndrome

Emily K. Schworer[1], **Kellie Voth**[1], **Emily K. Hoffman**[1], **Anna J. Esbensen**[1,2]

[1.] Division of Developmental and Behavioral Pediatrics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

[2.] Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA

Abstract

Background.—Improving short-term memory (STM) performance for individuals with Down syndrome (DS) has been a target of recent clinical trials. Validation of STM outcome measures is essential for research rigor in trials among children and adolescents with DS. *Aims.* The current study investigated the psychometric properties of four direct STM assessments and one everyday memory parent form.

Methods and Procedures.—Measures were administered to a sample of 74 youth with DS at two visits, two weeks apart. Overall cognitive abilities were also assessed.

Outcomes and Results.—The OMQ-PF had good feasibility and distribution of scores, but floor effects were prominent for direct measures. Test-retest reliability was poor to moderate for all measures and practice effects were problematic for the NEPSY-II List Memory and DAS-II Recall of Objects subtests. Commonalities in responses were observed, including primacy/recency effects, and some STM scores were correlated with overall cognitive abilities.

Conclusions and Implications.—The OMQ-PF met most study criteria, but no direct measure met sufficient criteria to be strongly recommended for future clinical trials. Because higher cognitive abilities were related to assessment completion, STM measures may require adaptation for use in broader samples of youth with DS across all levels of cognitive ability.

Keywords

Down syndrome; short-term memory; measurement; clinical trials; children

1. Introduction

Down syndrome (DS) is a neurogenetic syndrome marked by the triplication of chromosome 21. Individuals with DS have various cognitive strengths and challenges that often manifest in a distinct neurological profile (Chapman & Hesketh, 2000; Frenkel &

Correspondence concerning this article should be addressed to Emily K. Schworer, Cincinnati Children's Hospital Medical Center, Division of Developmental and Behavioral Pediatrics, 3430 Burnet Avenue, MLC 4002, Cincinnati, OH 45229. Phone: 513-636-3880 emily.schworer@cchmc.org.

The authors have no conflicts of interest to disclose.

Bourdin, 2009; Silverman, 2007). One well-characterized cognitive challenge in DS is memory consolidation, with the short-term memory (STM) component particularly affected throughout the lifespan (Godfrey & Lee, 2018; Jarrold et al., 2008). STM, also known as immediate memory, is the limited capacity for storage of a small amount of information for a brief time (Cowan, 2008). Cognitive functioning, including STM capacity, is related to broader challenges with academics and employment outcomes in those with and without intellectual disabilities (Bull et al., 2008; Daunhauer et al., 2020; Su et al., 2008; Tomaszewski et al., 2018). Because of structural differences in brain development in individuals with DS, the impact STM has on cognitive functioning, and the relation between STM and critical adaptive outcomes, this cognitive skill is a current target for interventions in DS (Conners et al., 2008; Jarrold et al., 2009).

STM storage can vary based on task modality (i.e., verbal or visuospatial) and performance on verbal or visuospatial measures differs intra-individually in DS (Cowan, 2008; Lee et al., 2016). Verbal STM is commonly described as a relative weakness throughout the lifespan in DS (Godfrey & Lee, 2018). Individuals with DS perform consistently lower on verbal STM measures such as digit, word, and sentence span tasks in comparison to mental age-matched typically developing peers and chronological age-matched children with other neurogenetic syndromes such as Williams syndrome (Edgin et al., 2010; Jarrold & Baddeley, 1997; Klein & Mervis, 1999; Seung & Chapman, 2000). Conversely, visuospatial STM is considered a relative strength within DS. Visuospatial STM in children with DS is comparable to children with typical development matched on mental age and comparable or stronger when related to other groups with intellectual disability (Carney et al., 2013; Jarrold & Baddeley, 1997; Kogan et al., 2009). Visuospatial aids have also been found to improve performance on verbal STM tasks in children and adolescents with DS (Duarte et al., 2011). Although there are clear relative strengths in this domain, because of overall intellectual disability, visuospatial STM performance is lower in DS compared to peers with typical development and a similar chronological age (Edgin, 2013; Godfrey & Lee, 2018).

Relative challenges with STM, particularly within the verbal component, further impact other cognitive processes that rely on STM. In fact, lower STM capacity has a cascading effect on working memory and long-term memory, as research shows that challenges with immediate memory prevent sufficient processing and long-term storage of information (Baddeley & Jarrold, 2007; Broadley et al., 1995; Cowan, 2008; Lanfranchi et al., 2004). For example, verbal STM is necessary for new word learning (Jarrold et al., 2009). A novel word must first be stored in STM to develop a long-term representation of that word (Baddeley et al., 1998; Gathercole, 2006; Jarrold et al., 2009). Such difficulties with new word learning may impact performance in more general developmental domains, reinforcing the need for intervention programs in DS targeting STM (Baddeley & Jarrold, 2007; Jarrold et al., 2009). Additionally, there is interaction among STM and other memory systems, as evidenced by primacy and recency effects (Capitani et al., 1992). Primacy effects imply working memory or longer-term stores are used compared to recency effects that are supported by STM (Capitani et al., 1992). Previous studies found recency but not primacy effects in adolescents and adults with DS compared to typically developing children and, therefore, attention to serial position effects may be important for interpreting STM performance in DS (Purser & Jarrold, 2005; Vicari et al., 2004).

Both behavioral and pharmaceutical cognitive interventions are in development for children and adolescents with DS, with varying degrees of preliminary evidence. Several studies involving school-aged children with DS show initial evidence that memory training is effective in improving STM skills and related cognitive processes (Bennett et al., 2013; Broadley & MacDonald, 1993; Conners et al., 2001; Conners et al., 2008; Laws et al., 1996). These memory trainings involved implementation of rehearsal and categorization strategies by study staff (Broadley & MacDonald, 1993; Laws et al., 1996), parents (Conners et al., 2001; Conners et al., 2008), or computerized systems monitored by a teaching assistant (Bennett et al., 2013). Modest improvements were shown in proximal memory outcomes. In one study, children with higher language and working memory skills benefited most from training, suggesting differential results depending on skills at study entry (Conners et al., 2008). Pharmaceuticals are another potential treatment for memory-related challenges in individuals with DS (Hart et al., 2017). Thus far, pharmaceutical trials have found little to no effect on improving cognitive abilities in this population (Hart et al., 2017; Kishnani et al., 2010), however, safety has been established and improvements in memory have been observed in a pilot study with a small sample size (Heller, Spiridigliozzi, Crissman, Sullivan, et al., 2006). Further, anecdotal reports from clinical trials noted distinct improvement in functional outcomes (Moyer, 2021; Roche, 2016). Together with the mix of promising preliminary evidence, the lack of clinical trial findings may be due to measurement limitations and a paucity of STM outcome measures validated in children and adolescents with DS.

The limited number of appropriate cognitive outcome measures is a major concern surrounding DS-focused behavioral and clinical medication trials (Esbensen et al., 2017; Heller, Spiridigliozzi, Crissman, Sullivan-Saarela, et al., 2006). Because standardized measures were normed with typically developing populations, these assessments are at risk for producing floor effects when used in studies of youth with DS, which hinders sensitivity to track performance and evaluate intervention results (Esbensen et al., 2017; Heller et al., 2006). There have been several studies aimed to address these measurement challenges by psychometrically evaluating cognitive tests in groups with DS and determining what measures are suitable for the population (d'Ardhuy et al., 2015; Edgin et al., 2017; Edgin et al., 2010; Schworer, Esbensen, et al., 2021). In previous reports, a parent memory questionnaire (Observer Memory Questionnaire-Parent Form), a spatial memory task (Cambridge Neuropsychological Test Automated Battery Paired Associate Learning), and list learning subtest (Repeatable Battery for the Assessment of Neuropsychological Status List Learning) showed minimal floor effects (4–14%) in individuals 7–38 years old (d'Ardhuy et al., 2015; Edgin et al., 2017; Edgin et al., 2010; Spanò & Edgin, 2016), whereas forward digit span (Differential Ability Scales-II Digits Forward) showed more moderate floor effects (22%) in children 6–19 years old (Schworer, Esbensen, et al., 2021). Furthermore, moderate to good test-retest was found in each of the studies, signaling that performance was consistent across multiple study visits (d'Ardhuy et al., 2015; Edgin et al., 2010; Schworer, Esbensen, et al., 2021). Although this work provides options for measurement of STM in youth with DS, additional measures require validation to meet the needs of future clinical trials.

1.1 Current study

Ultimately, behavioral and pharmaceutical STM interventions require psychometrically sound outcome measures to assess memory performance in DS (Heller et. al, 2006; Costa, 2011; Costa & Scott-McKean, 2013). Although STM interventions have been implemented in this population, previous studies have been limited by a lack of appropriate endpoints to accurately measure change in cognitive abilities, including STM, for children and adolescents with DS. The current study investigated a set of STM subtests from three clinical assessments determined to be promising for use in DS: the Developmental Neuropsychological Assessment, second edition (NEPSY-II), Differential Ability Scales, second edition (DAS-II), and Children's Memory Scale (CMS) (Esbensen et al., 2017). A computerized measure designed for research, the Cambridge Neuropsychological Test Automated Battery Paired Associate Learning (CANTAB PAL), and parent questionnaire, the Observer Memory Questionnaire-Parent Form (OMQ-PF), were also evaluated. Feasibility and score distributions were assessed to determine the number of participants with DS who could complete each measure and the variability in observed scores. Psychometric properties (i.e., test-retest, practice effects, and convergent validity) were evaluated to verify that the measures were psychometrically sound for use in clinical trials in DS. Participant performance was examined to determine if there were commonalities in responses across participants, such as primacy/recency effects or acquiescence. Finally, post hoc sensitivity and specificity were calculated for subtests with low feasibility (< 80%). Understanding what overall cognitive abilities are required to complete the STM measures is critical for making recommendations regarding inclusion criteria for clinical trials and preparing appropriate adaptations to outcome measures in future studies.

2. Material and methods

2.1 Participants

This study included 74 children and adolescents with DS 6 to 19 years old (M = 12.76, SD = 3.22). Age was normally distributed and there was no significant skewness (−0.54) or kurtosis (−0.55). Average IQ was 48.70, SD = 4.75 and average Vineland-3 Adaptive Behavior Composite was 68.28, SD = 11.06. Participants were primarily White (88%) and not Hispanic (93%). There were approximately equal numbers of males and females included (41 male/33 female). Participants were seen as part of a larger longitudinal study on cognitive outcome measures in DS.

2.2 Procedure

The Streamlined, Multisite, Accelerated Resources for Trials (SMART) IRB platform approved all study procedures. To participate, English was required as the family's primary language. The child needed to have an estimated developmental level of approximately 3 years old according to parent report, to support child participation in the broader study. Nonverbal or minimally verbal participants were included in the study. Caregivers confirmed DS diagnosis. Two sites recruited participants and study information was distributed through local DS clinics and DS associations. Participants visited the clinic or laboratory space at two time points within a two-week interval. A large battery of cognitive measures was administered, including measures of intelligence and memory pertinent to the current study.

The order of the study measures was randomized in blocks to prevent systematic differences in performance based on participant attention span. However, the NEPSY-II List Memory and DAS-II Recall of Objects were in the same block and therefore administered in the same order (DAS-II Recall of Objects followed by NEPSY-II List Memory). Computerized tasks, including the CANTAB PAL, were administered after all standardized measures. Testing sessions lasted approximately 1.5 – 2.5 hours. Breaks were provided to participants when needed throughout the visit to prevent fatigue. Only participants who completed both Time 1 and Time 2 were included in analyses.

2.3 Measures

2.3.1 Short-term memory measures.—STM measures were selected from standardized cognitive assessments typically used in clinical practice or considered promising by prior working groups (Esbensen et al., 2017). Participants' chronological age was generally within the standard administration guidelines. Measures were not administered if participants were younger than the normative range of the measure. Participants older than the age range of the standardized measures were included, as their developmental level was within the normative range. In these cases, scaled scores were determined using the highest chronological age available. Standard administration was used unless otherwise specified below.

2.3.1.1 Developmental Neuropsychological Assessment, second edition (NEPSY-II; Korkman et al., 2007) List Memory.: The NEPSY-II List Memory assesses short-term recall of verbal information and is normed for children 7 – 12 years old. Participants were read a list of 15 common words at a rate of one per second and then asked to immediately recall the words. This procedure was repeated for a total of five trials. List Memory was not administered to the 6-year-olds in the study (n = 3) and therefore the total sample size for this task was slightly smaller (n = 71). Total number of correct words (raw scores) and scaled scores ($M = 10$, $SD = 3$) are reported.

2.3.1.2 Differential Ability Scales, second edition (DAS-II; Elliott, 2007) Recall of Objects.: The DAS-II Recall of Objects measures short-term recall of verbal information with a visual support and is normed for children ages 4 – 17 years. A grid of 20 pictures (i.e., visual support) that corresponded with list words were labeled for the participant in Trial 1. Next, the visual support was removed, and participants were asked to immediately recall the words. In the second and third trials, participants were shown the visual support for 20 seconds, but were not read the words before being asked to recall them. Total number of correct words (raw scores) and T-scores ($M = 50$, $SD = 10$) are reported.

2.3.1.3 Children's Memory Scale (CMS; Cohen, 1997) Dot Locations.: The CMS Dot Locations subtest assesses visuospatial STM and is normed for children 5 – 16 years old. For participants 5 – 8 years old, a 3×4 blank grid and 6 chips were provided. Participants 9 years and older received a 4×4 blank grid and 8 chips. An array of blue dots (6 or 8 respectively) was shown to participants for 5 seconds. The pictured array was then removed, and the participant was required to place their chips on the blank grid in the locations from the pictured array. This procedure was completed on three consecutive trials. Next, a novel

array of red dots was displayed, and the participant was asked to recall the new pattern. This trial was not included in scoring. Finally, a 1-second reminder was shown of the first array and participants were asked to generate the first dot array. The brief reminder viewing of the first array was a deviation from standard CMS Dot Locations administration procedures, given that children did not understand the task demands of this short delay task in pilot testing. Raw scores are reported as a percentage correct because of the differences in grid size corresponding with chronological age. Percentage correct and scaled scores are reported for the first three learning trials and the short delay trial. Combined total percentage correct and scaled scores ($M = 10$, $SD = 3$), which include both learning and delay trials, are also presented.

2.3.1.4. Cambridge Neuropsychological Test Automated Battery Paired Associate Learning (CANTAB PAL).: The CANTAB PAL is a cognitive assessment of visuospatial STM and is administered using an iPad. The measure has been determined to be feasible in preliminary studies of children, adolescents, and adults with DS (Edgin et al., 2017; Edgin et al., 2010) and considered appropriate for individuals with DS (Esbensen et al., 2017). Boxes are displayed on the perimeter of the screen and are opened in a randomized order. One or more of the boxes contains a colorful abstract picture. The pictures are then displayed in the middle of the screen, one at a time, and the participant must select the box in which the picture was originally located. The number of pictures increases incrementally, starting at two pictures and ending with twelve on the most difficult trials. The number of CANTAB PAL First Attempt Memory (number of times the correct response is selected on their first attempt), Mean Errors to Success (mean number of attempts needed to complete the stage successfully), Total Errors Adjusted (incorrect selections adjusted for trials they did not reach), and Number of Patterns Reached (number of patterns on the participant's last problem) were used in analyses.

2.3.1.5. Observer Memory Questionnaire Parent Form (OMQ-PF; Gonzalez et al., 2008).: The OMQ-PF is a parent rating form of child memory for typically developing children ages 5 – 16 years and has been previously validated in children and adults with DS (d'Ardhuy et al., 2015; Spanò & Edgin, 2016). The questionnaire consists of 27 items that describe everyday memory in home and school contexts. Ratings are on a scale of strongly agree (1) to strongly disagree (5) or never (1) to always (5). Fifteen of the items are reverse scored and total scores range from 27 to 135, with higher scores indicating better memory. The total score was used in analyses.

2.3.2 Adaptive behavior.

Vineland Adaptive Behavior Scale, Third Edition (VABS-3; Sparrow et al., 2016).: Caregivers completed the VABS-3 at Time 1. The VABS-3 measures adaptive social, daily living, and communication skills, which together create the Adaptive Behavior Composite (ABC). The VABS-3 ABC ($M = 100$, $SD = 15$) was analyzed to determine the relation between STM measures and adaptive behavior.

2.3.3 Cognitive abilities.

Stanford Binet, fifth edition (SB-5; Roid, 2003).: The abbreviated battery IQ (ABIQ) was used to describe overall IQ in the sample. The SB-5 ABIQ is a standardized measure of cognitive ability, which includes nonverbal and verbal subtests. The SB-5 has high reliability (Roid, 2003) and correlations between the ABIQ and full-scale IQ are strong in clinical samples (Twomey et al., 2018). Deviation scores were used in this study to eliminate floor effects (deviation scoring procedures described in Sansone et al., 2014). The ABIQ deviation scores are an estimate of the full-scale *z* deviation scores. Both ABIQ deviation scores and subtest *z* scores (nonverbal Fluid Reasoning and verbal Knowledge) were used to compare performance on overall cognition to STM measures. Scores were normally distributed and there was no significant skewness (0.27) or kurtosis (0.79) for ABIQ deviation scores. There was a small correlation between age and ABIQ deviation scores ($r = .29$, $p = .01$).

2.4 Analysis plan

First, the feasibility and score distributions were assessed for the standardized STM measures in individuals 6 – 19 years old with DS. Feasibility was defined as the percentage of participants who provided responses (correct or incorrect) for the memory measures at Time 1 and Time 2. Before analyses were started, criterion for feasibility was set to ≥ 80% and has been a previous benchmark for feasibility in studies evaluating measures in groups with intellectual disability and DS (Hessl et al., 2016; Schworer, Esbensen, et al., 2021; Schworer, Hoffman, et al., 2021). Reasons for non-completion were recorded by examiners. Score distributions were also investigated and included descriptions of means, median, range of scores, skewness, and kurtosis. Skewness of less than −1 or greater than 1 and kurtosis of less than −2 or greater than 2 were considered outside the acceptable range. Statistical tests were modified to use nonparametric analyses when appropriate (i.e., Spearman correlations for scaled scores). To examine floor effects, two response options were summed: the number of participants who completed but received the lowest score on a measure and the number of participants unable to complete/generate responses to the measure at Time 1. Floor effects exceeding 20% were deemed problematic.

The next aim of the study involved evaluation of the STM measures' psychometric properties (i.e., test-retest reliability, practice effects, and convergent validity) in DS. Test-retest reliability and practice effects over a two-week interval were examined. Intraclass correlation coefficients (ICC) were used to assess test-retest reliability and characterized as poor (< .50), moderate (.50 – .74), good (.75 – .90), or excellent (> .90; Koo & Li 2016). Good or excellent categories were selected as *a priori* criterion for reliability. Practice effects were evaluated using paired samples *t*-tests. Significant differences between scores at the two time points and effect sizes (Cohen's *d*) larger than 0.20 signaled the presence of practice effects. Convergent validity among the five memory measures at Time 1 was assessed using bivariate Pearson correlations. Descriptive categories included poor (< .50), adequate (.50 – .70) or good (> .70), and the adequate and good categories were selected *a priori* as acceptable for research (Schworer, Esbensen, et al., 2021). Correlations with age, VABS-3, and SB-5 ABIQ were also examined to assess associations between STM measures at Time 1 and broader developmental domains.

The third study aim concerned describing performance at Time 1 on the four STM direct measures to observe potential commonalities in responses among participants with DS, such as primacy/recency effects or acquiescence. Planned analyses differed by measure to match the task demands of each type of assessment. For the NEPSY-II List Memory, the proportion of responses was calculated across the five trials to examine primacy effects, recency effects, and frequency of common participant responses. Item level responses were also examined for the DAS-II Recall of Objects, however, given the added visual component, responses were investigated considering both primacy and recency effects, as well as their location on the visual support page. The CMS Dot Locations was investigated for acquiescence (i.e., responding without considering dot placement options) because no verbal responses were required for this subtest and a score could be obtained from random placement of chips. Examples of acquiescence were the child forming structured rectangle shapes on the grid or placement of chips without looking at the grid. *T*-tests were used to compare performance of participants determined as acquiescing as opposed to participants showing cognitive effort to remember the placement of dots. Finally, score distributions for the level of patterns reached by participants on the CANTAB PAL were examined and the relation with IQ was assessed.

The final aim investigated measures performing below the feasibility criterion. Post hoc sensitivity and specificity analyses were completed for any measure with feasibility below 80%. Sensitivity calculations detail the proportion of participants in the study sample who were correctly identified as able to complete the subtest. Specificity proportions indicate the probability of participants in the study sample that were correctly identified as unable to complete a subtest. Age and IQ conditions were examined to determine sensitivity and specificity for varying benchmarks. Both age (8 and 10 years) and IQ (no restriction, IQ > 40, IQ > 45, and IQ > 50) benchmarks were examined. Selected ages were informed by ages of children with DS in recent clinical trials (Kishnani et al., 2010) and used in previous studies examining sensitivity and specificity of outcome measures (Schworer, Esbensen, et al., 2021).

3. Results

3.1 Study aim 1: Feasibility and score distribution

The DAS-II Recall of Objects and OMQ-PF were the only measures that met criterion for feasibility (81.1%, 94.6%; Table 1). *A priori* criterion for feasibility was not met for NEPSY-II List Memory, CMS Dot Locations, or CANTAB PAL, with less than 80% of participants completing each measure. For the NEPSY-II List Memory, reasons for non-completion were not understanding the task (14.1%), verbal ability (8.5%), noncompliance (3.5%), verbal refusal (1.4%), and only completing at one visit (3.5%). Similar reasons were indicated for the CMS Dot Locations and included not understanding the task (14.9%), noncompliance (9.4%), verbal refusal (4.0%), and only completing at one visit (1.4%). The CANTAB PAL had a variety of reasons for non-completion including not understanding the task (44.5%), behavioral noncompliance (6.8%), verbal refusal (5.4%), child fatigue (6.8%), and technology error (6.8%).

Tables 1 and 2 present information on the score distributions, including mean, median, range, skewness, and kurtosis. When considering raw/ability scores, a range of scores was

observed across measures. Skewness and kurtosis were not problematic for any of the raw/ability scores. However, floor effects were problematic for all direct measures (> 20%). Direct measure scaled scores and T-scores had a more restricted range in the sample and were positively skewed (> 1). Floor effects also exceeded acceptable levels for all subtests' scaled scores (31.1 – 81.7%). No floor effects were observed on the OMQ-PF.

3.2 Study aim 2: Psychometric evaluation

Overall, test-retest reliability for the direct STM measures was poor to moderate (.14 – .69; see Table 2) and no measures met the good to excellent reliability categories for criterion set *a priori*. There was evidence for practice effects for the NEPSY-II List Memory raw and standard scores, DAS-II Recall of Objects ability and T-scores, and CMS Dot Locations Total raw score. Test-retest for the OMQ-PF was moderate (.70) and there were no significant differences between parents' responses at Time 1 and Time 2.

None of the measures demonstrated convergent validity with all other assessments (Table 3). However, some correlations among specific assessments provide evidence for convergent validity. First, the NEPSY-II List Memory and DAS-II Recall of Objects raw scores had a strong correlation (r = .68), indicating convergent validity between these two verbal list memory measures, but not with the visuospatial CMS Dot Locations or CANTAB PAL measures. The second observation of strong correlations was between CMS Dot Locations raw scores and the CANTAB PAL First Attempt Memory and Total Errors Adjusted scores (r = .49 – .54 and −.56 – −.61). There were also correlations among CMS Dot Locations raw scores. While these correlations within CMS Dot Locations raw scores demonstrate internal consistency, rather than convergent validity between different measures, it is noteworthy that the different scores on this measure were consistent in the varying types of administration (three trials in a row vs. short delay with a brief reminder). Finally, the OMQ-PF was not significantly correlated with any of the direct measures of STM (Table 3). Scaled scores were not investigated for convergent validity given the floor effects observed in Aim 1.

There were significant positive correlations between chronological age and the NEPSY-II List Memory, DAS-II Recall of Objects, and CMS Dot Locations raw scores (Table 2; r = .31 – .39). The NEPSY-II List Memory raw/scaled scores, DAS-II Recall of Objects ability/T-scores, and CMS Dot Locations Short Delay raw/scaled score were positively associated with the VABS-3 ABC (Table 2; r = .32 – .44). Unexpectedly, the CANTAB PAL First Attempt Memory score was negatively associated with the VABS-3 ABC (r = −.50). Significant associations with the SB-5 deviation and SB-5 subdomain z scores were also observed for NEPSY-II List Memory raw scores, DAS-II Recall of Objects ability scores, and CMS Dot Locations raw scores (r = .29 – .53). The NEPSY-II List Memory and DAS-II Recall of Objects scaled scores were also both associated with the SB-5 verbal Knowledge domain (r = .29 – .35).

3.3 Study aim 3: Direct measure responses

3.3.1 NEPSY-II List Memory.—Proportions of responses for each list word were calculated across the five NEPSY-II List Memory trials (Figure 1). Recency effects were observed for list words in position 13 – 15, with word 15 at the highest frequency of any

response (62%). The primacy effect was also observed with a relatively high frequency of list word 1 (31%). Mid-list there were also several words above 20%, specifically word 7 and word 10 (both animal words).

3.3.2 DAS-II Recall of Objects.—Similar to NEPSY-II List Memory, proportions of responses were calculated for list words across the three DAS-II Recall of Objects trials (Figure 2). Prominent primacy effects were observed, as word 1 (the first list word read and positioned in the top left corner on the visual support) had the highest frequency in responses (49%). Other high frequency responses were words in position 2 (21%), 4 (26%), 5 (23%), 6 (28%), and 19 (25%). Of the more frequent participant responses, two were animal words (words 5 and 19). Recency effects were not detected, as the frequency list word 20 (the last word) was < 20%. Further, responses were analyzed based on the position of the pictures on the visual support, and response frequency percentages were calculated for rows and columns (Figure 3). The highest frequencies were row 1 and column 1 of the visual support, which correspond with the item level data, with word 1 (included in both row 1 and column 1) at a greater frequency in participant responses compared to other list words. The middle row (row 3) had the lowest frequency of responses.

3.3.3 CMS Dot Locations.—For participants who completed the CMS Dot Locations measure, examiners noted whether participants were acquiescing when making their responses. Two categories were then created, “participants demonstrating acquiescence” and “participants not demonstrating acquiescence.” At Time 1, 30% of participants demonstrated acquiescence in their responses. There was no difference in performance between the two groups for Total raw scores, $t(50) = -0.42$, $p = .68$, $d = 0.13$, Learning raw scores, $t(50) = -0.28$, $p = .78$, $d = 0.09$, or Short-Delay raw scores, $t(50) = -0.24$, $p = .81$, $d = 0.08$. Participants who acquiesced had significantly lower SB-5 ABIQ deviation scores than those who did not, $t(49) = -2.38$, $p = .02$, $d = 0.72$.

3.3.4 CANTAB PAL.—Participants who were able to complete the CANTAB PAL had a normal distribution of scores and performance level ranged from two to twelve pictures. The level reached by participants was two pictures (9.1%), four pictures (31.8%), six pictures (27.3%), eight pictures (18.2%), and twelve pictures (13.6%). Level reached was not significantly correlated with SB-5 ABIQ deviation scores ($r = .21$, $p = .36$).

3.4 Study aim 4: Measures below feasibility criterion

Post hoc sensitivity and specificity probabilities were calculated for the three measures below feasibility criterion: the NEPSY-II List Memory, CMS Dot Locations, and CANTAB PAL. The age benchmarks examined, age 8 and 10, resulted in similar probability ratios for both sensitivity and specificity. Different probabilities were observed based on ABIQ deviation scores. Sensitivity was high and specificity was low for no restriction or lower ABIQ deviation score benchmarks. Conversely, as ABIQ deviation score benchmarks became more restrictive, sensitivity decreased, and specificity increased. Relatively high sensitivity and specificity were observed for NEPSY-II List Memory and CMS Dot at the ABIQ deviation ≥ 30 benchmark. There was no benchmark for the CANTAB PAL that had > 65% probabilities for both sensitivity and specificity.

4. Discussion

This study investigated STM measures from standardized clinical assessments, a computerized measure, and a parent form to determine the outcome measures that would be appropriate for children and adolescents with DS in future clinical trials (see summary in Table 5). One measure, DAS-II Recall of Objects, met feasibility criterion and had convergent validity with the NEPSY-II List Memory, but did not meet other *a priori* psychometric criteria. All examined direct measures had problematic floor effects and test-retest reliability, but practice effects were only problematic for the NEPSY-II List Memory and DAS-II Recall of Objects. The OMQ-PF had good feasibility, but only moderate test-retest reliability and no convergent validity with any of the direct STM assessments. Primacy and recency effects were observed in the list memory tasks and there was no difference in performance on the CMS Dot Locations subtest based on acquiescence. Children with higher cognitive abilities were more likely to complete low feasibility measures (NEPSY-II List Memory, CMS Dot, and CANTAB PAL), but IQ was not associated with performance on the CANTAB PAL. Measures evaluated in the current study should be used with caution or with restricted subgroups of individuals with DS in treatment studies.

4.1 Feasibility and psychometric evaluation

The DAS-II Recall of Objects and OMQ-PF were the only two measures to meet feasibility criterion. Standard administration of the DAS-II Recall of Objects includes a visual support and the higher feasibility of this measure signals that participants benefited from the visual support in remembering list words. The utility of visual supports has been reported in previous studies examining the impact of visual aids on verbal memory task performance in DS (Duarte et al., 2011). Additionally, the parent questionnaire format of the OMQ-PF supported its feasibility. The OMQ-PF was the only measure with acceptable levels of floor effects, which was consistent with prior findings (d'Ardhuy et al., 2015; Spanò & Edgin, 2016) and indicates this measure is promising for measuring STM outcomes for all participants with DS. Having measures without floor effects is critical for determining effects in treatment outcomes. Unfortunately, no direct measures had acceptable levels of floor effects and therefore may have issues capturing a full range of performance and change over time if used in treatment studies without restricted inclusion criteria. Additionally, the present findings regarding feasibility of the CANTAB PAL are inconsistent with other reports of feasibility (Edgin et al., 2017), and although order effects may have impacted performance (see 4.5 Limitations), further evaluation is needed regarding this outcome measure.

No evaluated STM measure met criterion for test-retest reliability which demonstrates that scores are not stable over a two-week testing interval. The majority of measures had moderate test-retest reliability (NEPSY-II List Memory, DAS-II Recall of Objects, CMS Dot Locations Total and Short Delay, CANTAB PAL First Attempt Memory and Total Errors Adjusted and OMQ-PF), indicating that although not at the *a priori* study criterion, these measures were approaching acceptable reliability. The CANTAB PAL test-retest reliability was comparable to that previously reported for children with DS (Edgin et al., 2017). Parent OMQ-PF ratings did not differ over the 2-week testing interval, providing evidence for

stability in measurement. Negligible practice effects were found for the CANTAB PAL First Attempt Memory and a portion of the CMS Dot Locations scores. Results indicating practice effects on word memory lists (NEPSY-II List Memory and DAS-II Recall of Objects) were expected, given the practice effects reported on scaled scores in typically developing populations (Korkman et al., 2007), but suggest a need for multiple versions of list memory measures to avoid practice effects when monitoring clinical trial outcomes across short intervals.

Convergent validity was observed between certain STM measures, but not among all measures. First, convergent validity was observed between the NEPSY-II List Memory and DAS-II Recall of Objects, which was expected given the similarity in word list recall verbal task demands. There was also convergent validity observed between the CMS Dot Locations and CANTAB PAL. These two tasks were similar in that they both were visuospatial tasks. There was no correlation between the evaluated parent questionnaire (OMQ-PF) and any of the direct STM measures. This result is consistent with previous work that showed the OMQ-PF was not significantly correlated with STM digit span measures (Gonzalez et al., 2008). The lack of convergence between parent-report and direct assessments indicates that the OMQ-PF may be tapping longer-term learning or retention, rather than short-term information storage. Although it will be important for future studies to evaluate the association between the OMQ-PF and standardized assessments of long-term memory or working memory, the evidence from the current study does not yet discount the utility of the OMQ-PF as a valid measurement of everyday memory. Rather, results suggest the OMQ-PF does not capture STM performance in youth with DS. It is possible that different types of memory or learning measures would have better convergent validity with the OMQ-PF.

Raw scores on several STM tasks (NEPSY-II List Memory, DAS-II Recall of Objects, and CMS Dot Locations) were positively associated with chronological age and suggest older children and adolescents with DS perform better than younger children on these three measures. Associations with cognition generally corresponded with STM task demands. For example, the NEPSY-II List Memory was significantly associated with SB-5 verbal knowledge, but not fluid reasoning, whereas the DAS-II Recall of Objects was significantly associated with both verbal and nonverbal SB-5 domains. Surprisingly, not all visuospatial STM tasks were significantly associated with the nonverbal fluid reasoning SB-5 domain. There were also moderate correlations observed between STM measures and adaptive behavior. It is plausible that better STM would be associated with better adaptive behavior, as memory likely supports adaptive performance in real-world contexts. In contrast, the CANTAB PAL first attempt memory score was negatively correlated with adaptive behavior. Replication of this finding is needed to confirm this inverse association considering the small sample size that could feasibly complete the PAL. Investigating the relation between the CANTAB PAL and subscales of the VABS-3 will be important for better understanding this association. Finally, the OMQ-PF was not associated with cognitive abilities or adaptive behavior, which deviates from previous reports of a moderate correlation between the OMQ-PF and IQ (Gonzalez et al., 2008).

4.2 Direct measure responses

Commonalities in list memory responses were observed for both the NEPSY-II List Memory and DAS-II Recall of Objects. Participant responses to both measures demonstrated primacy effects. Recency effects were prominent in NEPSY-II List Memory responses but were not as strong for the DAS-II Recall of Objects subtest. Primacy and recency effects may represent an interaction and recruitment of working memory and long term-memory systems required to recall long lists of words and suggests participants are indeed utilizing working memory and longer-term storage to complete word list measures. Responses corresponded with the visual support in the DAS-II Recall of Objects, such that the first row and first column had the greatest proportion of responses. The location of the stimuli impacting STM may be in part due to the spatial organization of pictures. This corresponds with research on augmentative and alternative communication (AAC) that shows that the position of symbols impacts the rate of responses (Wilkinson & McIlvane, 2013). Across the word list subtests, the types of words with high frequency tended to be animals and suggests that the words themselves may impact memory. This finding also raised concerns, as animals are common stimuli across neuropsychology testing and there was overlap on the animal stimuli in assessments that were part of the larger test battery in the study. Clinical trials should be cautious with this type of overlap to avoid priming participants for memory of certain words. There were also overlapping words between the two list memory tasks, which may also have cued participants to remember those words in the administration of the NEPSY-II List Memory, as this subtest was in a randomized block with the DAS-II Recall of Objects and was always administered after it.

Performance on the CMS Dot Locations was not significantly different when comparing participants who acquiesced in their responses and those who did not. This lack of differentiation suggests that children completing the CMS Dot Locations may not achieve meaningful scores, as equal scores can be obtained through effortful actions and through acquiescence. Thus, scores may not reflect visuospatial STM ability. Participants in the acquiescence group did have lower cognitive abilities, indicating that those with higher IQ were less likely to acquiesce in their responses. The CANTAB PAL scores did not correspond with overall cognitive abilities. This may suggest that overall fatigue confounded feasibility, as this measure was consistently administered towards the end of the visit in the broader study. It is also plausible that the STM abilities assessed in the CANTAB PAL are separate from abilities assessed using non-computerized clinical assessments presented here.

4.3 Measures below feasibility criterion

The NEPSY-II List Memory, CMS Dot Locations, and CANTAB PAL were all below feasibility criterion and thus sensitivity and specificity probabilities were calculated for each measure. As expected, higher ABIQ scores raised the probability of correctly identifying participants who could not complete a subtest, but also were not representative of all participants who could complete a subtest. The ABIQ deviation score of 30 gave reasonable sensitivity *and* specificity probabilities for both the NEPSY-II List Memory and CMS Dot Locations. This provides an approximate IQ that could be used as inclusion criteria in trials using these measures. Conversely, the CANTAB PAL did not have a clear benchmark

where sensitivity and specificity were relatively high and therefore, our sample did not show adequate feasibility for any participants, regardless of age or IQ.

4.4 Implications for clinical trials

Table 5 summarizes the met and unmet study criteria for the STM measures. This graphic illustrates the problematic floor effects, feasibility, test-retest reliability, and convergent validity for the STM measures when used with children with DS. The OMQ-PF fared best out of all evaluated measures, and with moderate test-retest reliability (.70) is deemed appropriate for use in clinical trials including youth with DS. The marked floor effects on the list memory subtests (NEPSY-II List Memory and DAS-II Recall of Objects), even using raw/ability scores, demonstrate a potential need for teaching trials on these measures, or options for shorter lists of words for individuals with DS. This would create a better fit between the limited STM capacity of individuals with DS (Godfrey & Lee, 2018; Purser & Jarrold, 2005) and STM task demands. By adjusting the task demands for the population, there is potential for more variability in the lower range of scores and a better likelihood of fewer participants scoring at the floor. The CMS Dot Locations also needs to be monitored closely for acquiescence, as participants were able to respond, but responses were not necessarily meaningful or correlated with cognitive abilities. This task may also benefit from a smaller matrix of dots, so that task demands more closely match the cognitive abilities of participants. Finally, although the CANTAB PAL has demonstrated good psychometrics in previous studies (Edgin et al., 2017), the current study suggests that completing this task at the end of a battery of assessments (ranging in duration from 1.5 – 2.5 hours) is not appropriate for children and adolescents with DS. Taken together, the measurement challenges identified in the current study highlight a key problem with use of traditional standardized clinical STM measures for children with DS or other intellectual disabilities. As evidenced in the current study, the STM capacity of children and adolescents with DS may not match the task demands presented on STM standardized clinical assessments. Further investigations and modification of measures to address these measurement problems are essential for developing outcome measures appropriate for children and adolescents with a range of cognitive abilities.

4.5 Limitations and future directions

This study provides guidance for STM measures appropriate for clinical trials that include children and adolescents with DS but is not without limitations. First, the current study is limited in that the testing interval was a short period. Future work will need to determine if, for example, practice effects are still problematic with a 3- or 6-month testing interval. Additionally, some stimuli were repeated on the NEPSY-II List Memory, DAS-II Recall of Objects, and other assessments in the neuropsychological battery not included in the current study. These words may have been more likely to be remembered by participants. In the current study, these two measures were also always administered in the same order because they were grouped in a randomization block. Future studies should randomize the order of these two measures. Participant fatigue was also a concern and should be considered when evaluating CANTAB PAL results, as this measure was administered near the end of the testing battery. Finally, presented analyses were completed using data from

research participants who volunteered for the study and may not be representative of the full population of children and adolescents with DS.

5.0 Conclusions

The psychometric evaluation of STM measures is important for the valid assessment of outcomes in future clinical trials for youth with DS (Eunice Kennedy Shriver National Institute of Child Health and Human Development, 2015; Esbensen et al., 2017). Findings from the current study support the use of the OMQ-PF, although test-retest reliability was lower than *a priori* criteria, and no convergent validity was observed with direct measures. The DAS-II Recall of Objects demonstrated feasibility, but no other study criteria were met. The NEPSY-II List Memory, CMS Dot Locations, and CANTAB PAL may be appropriate for individuals with DS with higher cognitive abilities. Adaptations to these clinical assessments may be necessary to make them accessible for broader groups of children and adolescents with DS.

Acknowledgments

This manuscript was prepared with support from the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development of the National Institutes of Health (R01 HD093754, Esbensen PI). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This research would not have been possible without the contributions of the participating families and the community support.

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Baddeley A, Gathercole S, & Papagno C. (1998). The phonological loop as a language learning device. Psychological Review, 105, 158–173. 10.1037/0033-295x.105.1.158 [PubMed: 9450375]

Baddeley A, & Jarrold C. (2007). Working memory and Down syndrome. Journal of Intellectual Disability Research, 51(12), 925–931. [PubMed: 17990999]

Bennett SJ, Holmes J, & Buckley S. (2013). Computerized memory training leads to sustained improvement in visuospatial short-term memory skills in children with Down syndrome. Am J Intellect Dev Disabil, 118(3), 179–192. 10.1352/1944-7558-118.3.179 [PubMed: 23734613]

Broadley I, & MacDonald J. (1993). Teaching short term memory skills to children with Down syndrome. Down Syndrome Research and Practice, 1(2), 56–62.

Broadley I, MacDonald J, & Buckley S. (1995). Working memory in children with Down's syndrome. Down Syndrome Research and Practice, 3, 3–8.

Bull R, Espy KA, & Wiebe SA (2008). Short-Term Memory, Working Memory, and Executive Functioning in Preschoolers: Longitudinal Predictors of Mathematical Achievement at Age 7 Years. Developmental Neuropsychology, 33(3), 205–228. 10.1080/87565640801982312 [PubMed: 18473197]

Capitani E, Della Sala S, Logie RH, & Spinnler H. (1992). Recency, primacy, and memory: reappraising and standardising the serial position curve. Cortex, 28(3), 315–342. 10.1016/s0010-9452(13)80143-8 [PubMed: 1395637]

Carney DP, Henry LA, Messer DJ, Danielsson H, Brown JH, & Rönnberg J. (2013). Using developmental trajectories to examine verbal and visuospatial short-term memory development in children and adolescents with Williams and Down syndromes. Research in Developmental Disabilities, 34(10), 3421–3432. [PubMed: 23920025]

Chapman RS, & Hesketh LJ (2000). Behavioral phenotype of individuals with Down syndrome. Mental Retardation and Developmental Disabilities Research Reviews, 6(2), 84–95. [PubMed: 10899801]

Cohen M. (1997). Children's Memory Scale (CMS). Pearson Assessment London.

Conners F, Rosenquist C, & Taylor L. (2001). Memory training for children with Down syndrome. Down Syndrome Research and Practice, 7(1), 25–33.

Conners FA, Rosenquist CJ, Arnett L, Moore M, & Hume LE (2008). Improving memory span in children with Down syndrome. Journal of Intellectual Disability Research, 52(3), 244–255. [PubMed: 18261023]

Cowan N. (2008). What are the differences between long-term, short-term, and working memory? Prog Brain Res, 169, 323–338. 10.1016/s0079-6123(07)00020-9 [PubMed: 18394484]

d'Ardhuy XL, Edgin JO, Bouis C, de Sola S, Goeldner C, Kishnani P, Nöldeke J, Rice S, Sacco S, & Squassante L. (2015). Assessment of cognitive scales to examine memory, executive function and language in individuals with Down syndrome: Implications of a 6-month observational study. Frontiers in Behavioral Neuroscience, 9, 300. [PubMed: 26635554]

Daunhauer LA, Will E, Schworer E, & Fidler DJ (2020). Young students with Down syndrome: Early longitudinal academic achievement and neuropsychological predictors. Journal of Intellectual & Developmental Disability, 45(3), 211–221. 10.3109/13668250.2020.1726016

Eunice Kennedy Shriver National Institute of Child Health and Human Development (2015). Outcome measures for clinical trials in individuals with Down syndrome https://www.nichd.nih.gov/about/meetings/2015/Documents/DS_outcomes_meeting_summary.pdf

Duarte CP, Covre P, Braga AC, & de Macedo EC (2011). Visuospatial support for verbal short-term memory in individuals with Down syndrome. Research in Developmental Disabilities, 32(5), 1918–1923. [PubMed: 21530159]

Edgin JO (2013). Cognition in Down syndrome: a developmental cognitive neuroscience perspective. Wiley Interdiscip Rev Cogn Sci, 4(3), 307–317. 10.1002/wcs.1221 [PubMed: 26304208]

Edgin JO, Anand P, Rosser T, Pierpont EI, Figueroa C, Hamilton D, Huddleston L, Mason G, Spanò G, & Toole L. (2017). The Arizona Cognitive Test Battery for Down Syndrome: Test-Retest Reliability and Practice Effects. American Journal on Intellectual and Developmental Disabilities, 122(3), 215–234. [PubMed: 28452581]

Edgin JO, Mason GM, Allman MJ, Capone GT, DeLeon I, Maslen C, Reeves RH, Sherman SL, & Nadel L. (2010). Development and validation of the Arizona Cognitive Test Battery for Down syndrome. Journal of Neurodevelopmental Disorders, 2(3), 149–164. 10.1007/s11689-010-9054-3 [PubMed: 21274406]

Edgin JO, Pennington BF, & Mervis CB (2010). Neuropsychological components of intellectual disability: the contributions of immediate, working, and associative memory. Journal of Intellectual Disability Research, 54(5), 406–417. 10.1111/j.1365-2788.2010.01278.x [PubMed: 20537047]

Elliott CD (2007). Differential Ability Scales, 2nd edition: Introductory and technical handbook. The Psychological Corporation.

Esbensen A, Hooper SR, Fidler D, Hartley SL, Edgin J, d'Ardhuy XL, Capone G, Conners FA, Mervis CB, & Abbeduto L. (2017). Outcome measures for clinical trials in Down syndrome. American Journal on Intellectual and Developmental Disabilities, 122(3), 247–281. [PubMed: 28452584]

Frenkel S, & Bourdin B. (2009). Verbal, visual, and spatio-sequential short-term memory: assessment of the storage capacities of children and teenagers with down's syndrome. Journal of Intellectual Disability Research, 53(2), 152–160. [PubMed: 19077148]

Gathercole SE (2006). Nonword repetition and word learning: The nature of the relationship. Applied Psycholinguistics, 27(4), 513–543.

Godfrey M, & Lee NR (2018). Memory profiles in Down syndrome across development: a review of memory abilities through the lifespan. Journal of Neurodevelopmental Disorders, 10(1), 1–31. [PubMed: 29329511]

Gonzalez LM, Anderson VA, Wood SJ, Mitchell LA, Heinrich L, & Harvey AS (2008). The Observer Memory Questionnaire—Parent Form: Introducing a new measure of everyday memory

for children. Journal of the International Neuropsychological Society, 14(2), 337–342. 10.1017/S135561770808020X [PubMed: 18282331]

Hart SJ, Visootsak J, Tamburri P, Phuong P, Baumer N, Hernandez MC, Skotko BG, Ochoa-Lubinoff C, Liogier D'Ardhuy X, Kishnani PS, & Spiridigliozzi GA (2017). Pharmacological interventions to improve cognition and adaptive functioning in Down syndrome: Strides to date. Am J Med Genet A, 173(11), 3029–3041. 10.1002/ajmg.a.38465 [PubMed: 28884975]

Heller JH, Spiridigliozzi GA, Crissman BG, Sullivan-Saarela JA, Li JS, & Kishnani PS (2006). Clinical trials in children with Down syndrome: issues from a cognitive research perspective. American Journal of Medical Genetics Part C: Seminars in Medical Genetics, 142C(3), 187–195. 10.1002/ajmg.c.30103

Heller JH, Spiridigliozzi GA, Crissman BG, Sullivan JA, Eells RL, Li JS, Doraiswamy PM, Krishnan KR, & Kishnani PS (2006). Safety and efficacy of rivastigmine in adolescents with Down syndrome: a preliminary 20-week, open-label study. Journal of Child & Adolescent Psychopharmacology, 16(6), 755–765. [PubMed: 17201619]

Hessl D, Sansone SM, Berry-Kravis E, Riley K, Widaman KF, Abbeduto L, Schneider A, Coleman J, Oaklander D, & Rhodes KC (2016). The NIH Toolbox Cognitive Battery for intellectual disabilities: three preliminary studies and future directions. Journal of Neurodevelopmental Disorders, 8(1), 35. [PubMed: 27602170]

Jarrold C, & Baddeley AD (1997). Short-term memory for verbal and visuospatial information in Down's syndrome. Cognitive Neuropsychiatry, 2(2), 101–122. [PubMed: 25420199]

Jarrold C, Nadel L, & Vicari S. (2008). Memory and neuropsychology in Down syndrome. Down Syndrome Research and Practice, 12, 68–73.

Jarrold C, Thorn AS, & Stephens E. (2009). The relationships among verbal short-term memory, phonological awareness, and new word learning: Evidence from typical development and Down syndrome. Journal Of Experimental Child Psychology, 102(2), 196–218. [PubMed: 18707692]

Kishnani PS, Heller JH, Spiridigliozzi GA, Lott I, Escobar L, Richardson S, Zhang R, & McRae T. (2010). Donepezil for treatment of cognitive dysfunction in children with Down syndrome aged 10–17. American Journal of Medical Genetics Part A, 152(12), 3028–3035.

Klein BP, & Mervis CB (1999). Contrasting patterns of cognitive abilities of 9-and 10-year-olds with Williams syndrome or Down syndrome. Developmental Neuropsychology, 16(2), 177–196.

Kogan CS, Boutet I, Cornish K, Graham GE, Berry-Kravis E, Drouin A, & Milgram NW (2009). A comparative neuropsychological test battery differentiates cognitive signatures of Fragile X and Down syndrome. Journal of Intellectual Disability Research, 53(2), 125–142. 10.1111/j.1365-2788.2008.01135.x [PubMed: 19054268]

Koo TK, & Li MY (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15, 155–163. 10.1016/j.jcm.2016.02.012 [PubMed: 27330520]

Korkman M, Kirk U, & Kemp S. (2007). NEPSY Second Edition (NEPSY-II). Harcourt Assessment.

Lanfranchi S, Cornoldi C, & Vianello R. (2004). Verbal and visuospatial working memory deficits in children with Down syndrome. American Journal on Mental Retardation, 109(6), 456–466. [PubMed: 15471512]

Laws G, MacDonald J, & Buckley S. (1996). The effects of a short training in the use of a rehearsal strategy on memory for words and pictures in children with Down syndrome. Down Syndrome Research and Practice, 4(2), 70–78.

Lee NR, 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. International Review of Research in Developmental Disabilities, 51, 1–40.

Moyer A. (2021). Too Good to Be True: Reflections on a Down Syndrome Clinical Trial. Johns Hopkins Medicine. Retrieved 8/5 from https://biomedicalodyssey.blogs.hopkinsmedicine.org/2021/03/too-good-to-be-true-reflections-on-a-down-syndrome-clinical-trial/

Purser HR, & Jarrold C. (2005). Impaired verbal short-term memory in Down syndrome reflects a capacity limitation rather than atypically rapid forgetting. Journal of Experimental Child Psychology, 91, 1–23. [PubMed: 15814093]

Roche. (2016). Statement on CLEMATIS trial. Retrieved 8/5 from http://www.edsa.eu/roche-clematis-trial-discontinued/

Roid G. (2003). Stanford-Binet Intelligence Scale: Fifth Edition. Riverside.

Sansone SM, 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), 1–15. [PubMed: 24433325]

Schworer EK, Esbensen AJ, Fidler DJ, Beebe DW, Carle A, & Wiley S. (2021). Evaluating working memory outcome measures for children with Down syndrome. Journal of Intellectual Disability Research. 10.1111/jir.12833

Schworer EK, Hoffman EK, & Esbensen AJ (2021). Psychometric evaluation of social cognition and behavior measures in children and adolescents with Down syndrome. Brain Sciences, 11(7). 10.3390/brainsci11070836

Seung HK, & Chapman R. (2000). Digit span in individuals with Down syndrome and in typically developing children: temporal aspects. J Speech Lang Hear Res, 43(3), 609–620. 10.1044/jslhr.4303.609 [PubMed: 10877432]

Silverman W. (2007). Down syndrome: cognitive phenotype. Mental Retardation and Developmental Disabilities Research Reviews, 13(3), 228–236. [PubMed: 17910084]

Spanò G, & Edgin JO (2016). Everyday memory in individuals with Down syndrome: Validation of the Observer Memory Questionnaire – Parent Form. Child Neuropsychology, 23(5), 523–535. 10.1080/09297049.2016.1150446 [PubMed: 26981787]

Sparrow SS, Cicchetti DV, & Saulnier CA (2016). Vineland Adaptive Behavior Scales, Third Edition. Pearson.

Su C-Y, Lin Y-H, Wu Y-Y, & Chen C-C (2008). The role of cognition and adaptive behavior in employment of people with mental retardation. Research in Developmental Disabilities, 29, 83–95. 10.1016/j.ridd.2006.12.001 [PubMed: 17210243]

Tomaszewski B, Fidler D, Talapatra D, & Riley K. (2018). Adaptive behaviour, executive function and employment in adults with down syndrome. Journal of Intellectual Disability Research, 62, 41–52. 10.1111/jir.12450 [PubMed: 29214700]

Twomey C, O'Connell H, Lillis M, Tarpey SL, & O'Reilly G. (2018). Utility of an abbreviated version of the stanford-binet intelligence scales in estimating 'full scale'IQ for young children with autism spectrum disorder. Autism Research, 11(3), 503–508. [PubMed: 29282895]

Vicari S, Marotta L, & Carlesimo GA (2004). Verbal short-term memory in Down's syndrome: an articulatory loop deficit?. Journal of Intellectual Disability Research, 48(2), 80–92. [PubMed: 14723651]

Wilkinson KM, & McIlvane WJ (2013). Perceptual factors influence visual search for meaningful symbols in individuals with intellectual disabilities and Down syndrome or autism spectrum disorders. American Journal on Intellectual and Developmental Disabilities, 118(5), 353–364. 10.1352/1944-7558-118.5.353 [PubMed: 24245729]

What this paper adds?

The current study informs the selection of STM outcome measures for future clinical trials in DS. Of the five evaluated measures, one direct measure and one parent questionnaire met feasibility criteria. The feasibility of other measures was not adequate to be recommended for use in future research. Children with higher cognitive abilities were more likely to complete measures with low feasibility. Direct STM measures had problematic floor effects, moderate test-retest reliability, some practice effects, and minimal convergent validity. The parent reported memory questionnaire showed good psychometric properties, with a normal distribution of scores, moderate test-retest reliability, and negligible differences between parent responses at Time 1 and 2, but no convergent validity with direct STM measures. This study also described commonalities in participant responses and identified primacy and recency effects on list memory tasks. Performance was correlated with IQ for some, but not all measures. Overall, this study confirms that the parent questionnaire is appropriate for use in clinical trials and advises that the evaluated direct assessments should be used with caution or with restricted subgroups of individuals with DS in treatment studies for youth with DS. We also make recommendations for items to monitor in a larger trial, such as the overlap of animal stimuli in various tests and visual components of tasks altering recency effects in word list memory tasks.

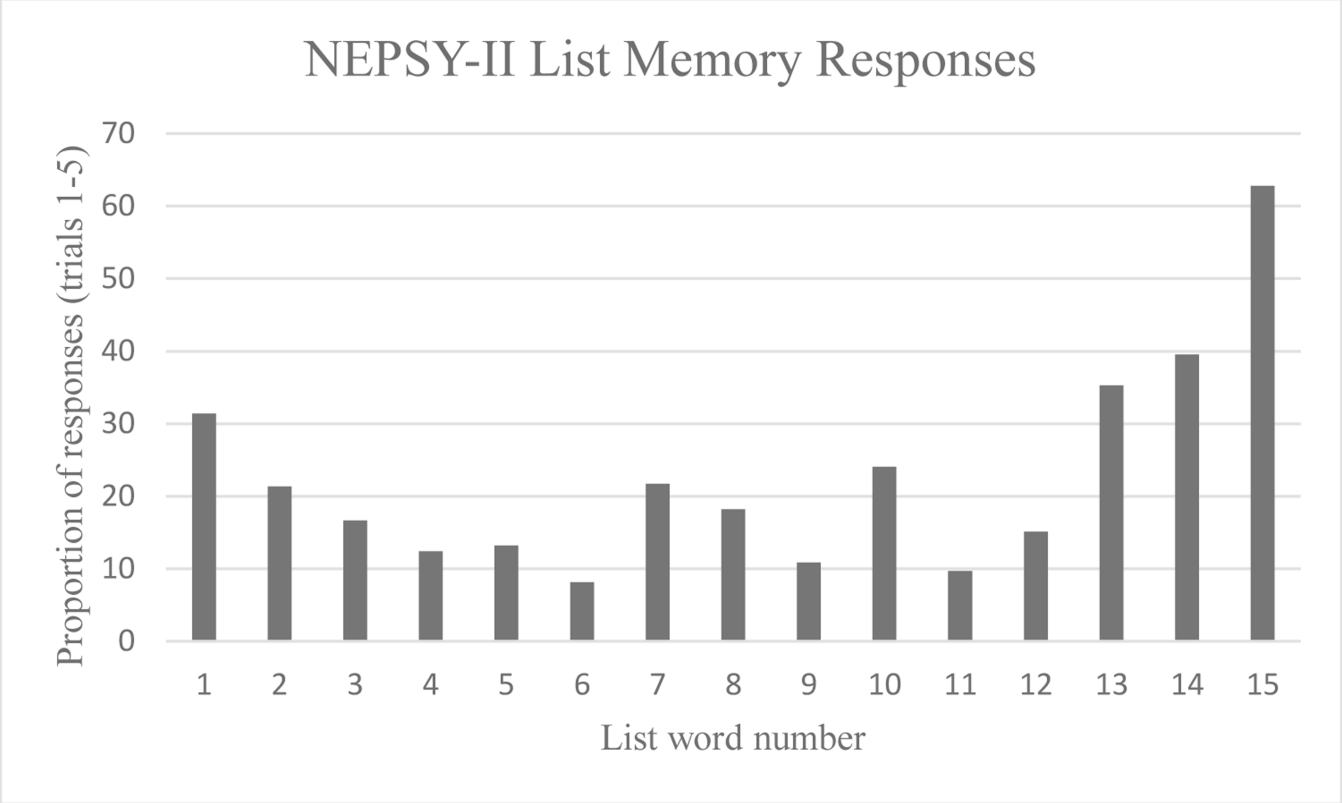

Figure 1.
NEPSY-II List Memory response proportions at Time 1

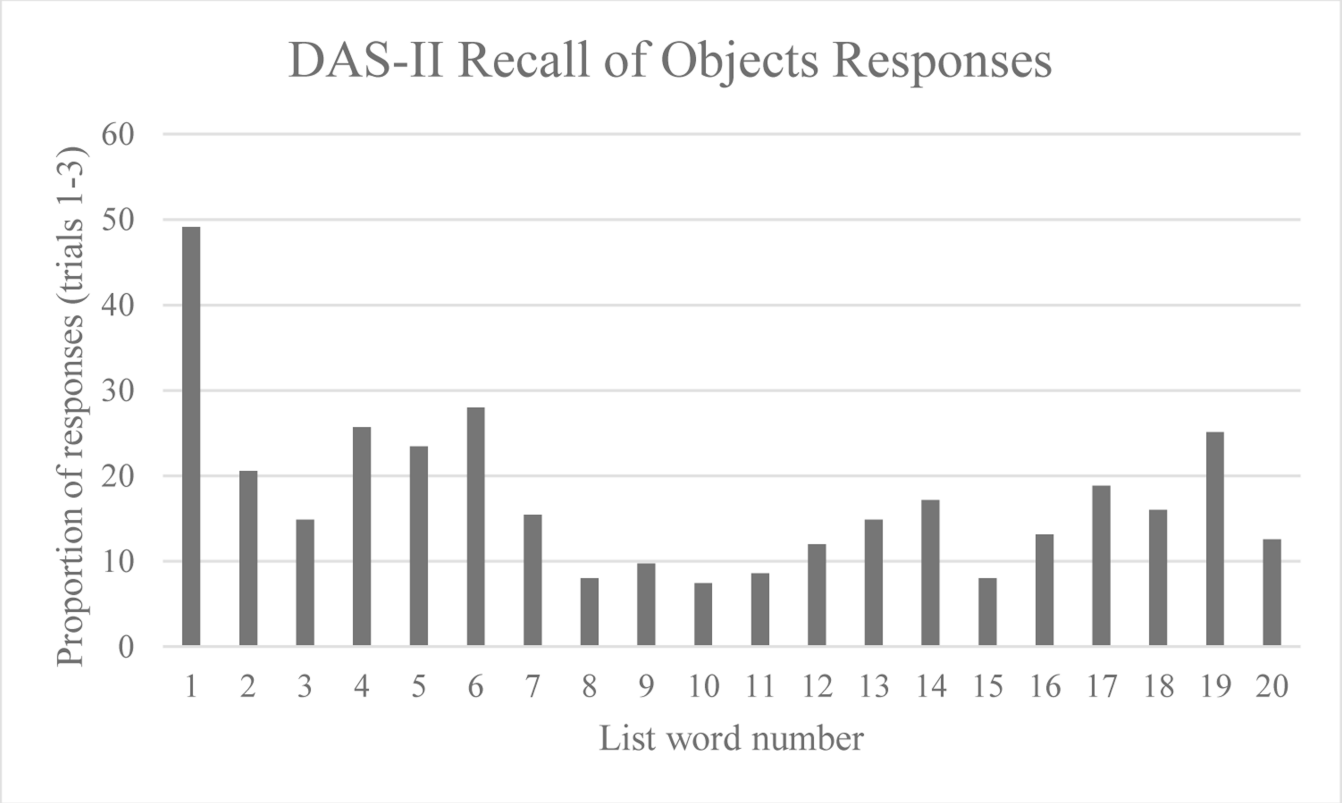

Figure 2.
DAS-II Recall of Objects response proportions at Time 1

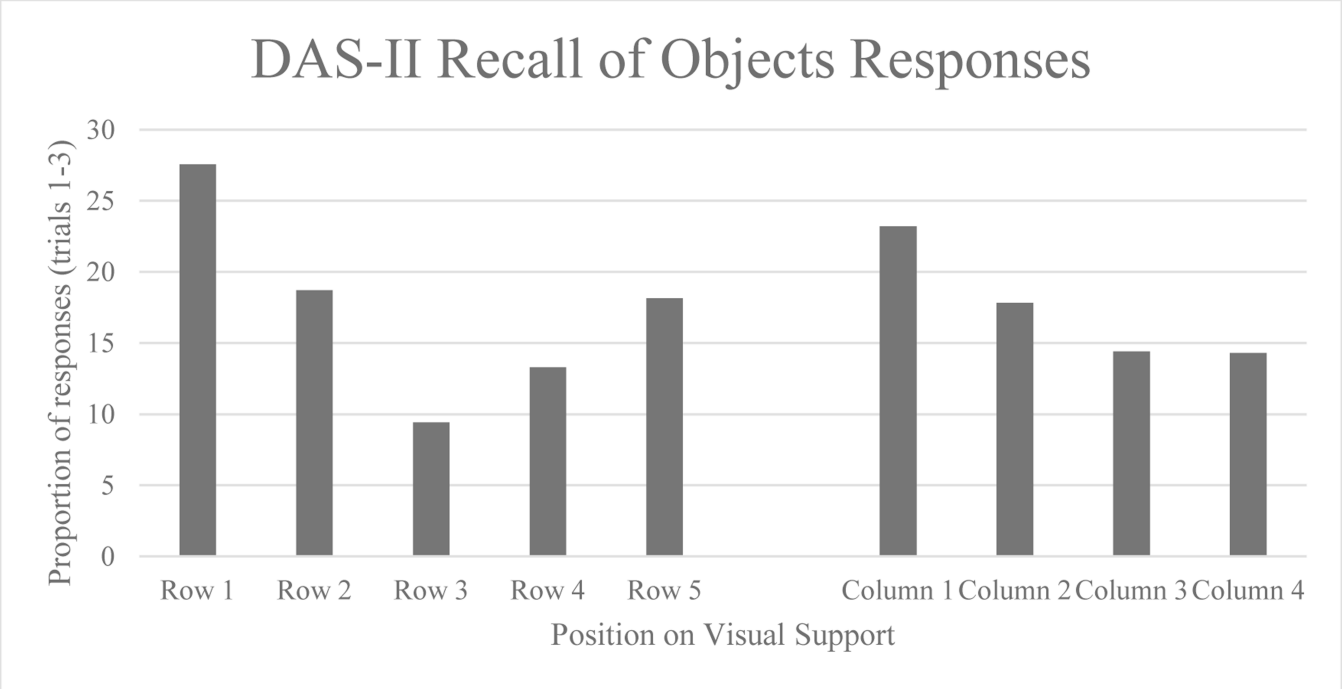

Figure 3.
DAS-II Recall of Objects response proportions at Time 1 collapsed into rows and columns of visual support.

Table 1.

Short-term memory task performance, feasibility, and floor effects at Time 1

	Median (range)	Skew	Kurtosis	Feasibility[a] n (%)	Participants at floor n (%)[b]
NEPSY-II List Memory Total Correct	16 (0–43)	0.65	0.22	49 (69.0%)	23/71 (32.4%)
NEPSY-II List Memory Scaled Score	1 (1–5)	1.71	1.63		58/71 (81.7%)
DAS-II Recall of Objects Ability Score	75 (10–143)	−0.02	−0.02	60 (81.1%)	16/74 (21.6%)
DAS-II Recall of Objects T-score	10 (10–31)	1.49	1.23		50/74 (67.6%)
CMS Dot Locations Total Raw Score[c]	53.65 (29.17–96.88)	0.55	0.90	52 (70.3%)	22/74 (29.7%)
CMS Dot Locations Learning Raw Score[c]	54.17 (29.17–95.83)	0.56	0.51		22/74 (29.7%)
CMS Dot Locations Short Delay Raw Score[c]	50.00 (12.50–100)	0.16	1.02		22/74 (29.7%)
CMS Dot Locations Total Scaled Score	4 (1–12)	1.02	1.65		36/74 (48.6%)
CMS Dot Locations Learning Scaled Score	3.5 (1–11)	1.12	1.13		38/74 (51.4%)
CMS Dot Locations Short Delay Scaled Score	5 (1–12)	1.05	1.37		23/74 (31.1%)
CANTAB PAL First Attempt Memory	5 (0–12)	.23	−1.22	22 (31.9%)[d]	49/69 (71.0%)[d]
CANTAB PAL Mean Errors to Success	1 (0–8)	1.86	4.98		48/69 (69.6%)[d]
CANTAB PAL Total Errors Adjusted	47 (12–69)	−0.50	−1.05		48/69 (69.6%)[d]
OMQ-PF	88 (60–115)	−0.11	−0.14	70 (94.6%)	0/74 (0%)

[a]Feasibility was defined as the percentage of participants who provided responses (correct or incorrect) at Time 1 and Time 2;

[b]Participants at floor included non-completers and participants with the lowest score on the measure;

[c]CMS Dot Locations raw scores reported as percentage correct because of the varying grid sizes that correspond with chronological age;

[d]Technology error (n=5) was removed from feasibility calculations for the CANTAB PAL, as it did not reflect individuals' ability to complete the task; NEPSY-II = Developmental Neuropsychological Assessment, second edition; DAS-II = Differential Abilities Scale, second edition; CMS = Children's Memory Scale; CANTAB PAL = Cambridge Neuropsychological Test Automated Battery Paired Associate Learning; OMQ-PF = Observer Memory Questionnaire-Parent Form

Table 2.

Practice effects, test-retest reliability, and associations with broader developmental domains

	Time 1, Mean (SD)	Time 2, Mean (SD)	Mean Difference	Cohen's *d*	ICC	Age	VABS-3 ABC	SB-5 ABIQ[a]	SB-5 Fluid Reasoning[a]	SB-5 Knowledge[a]
NEPSY-II List Memory Total Correct	17.86 (9.37)	22.53 (11.54)	4.67***	0.45	.69	.37**	.38**	.41**	.28	.51***
NEPSY-II List Memory Scaled Score[b]	1.63 (1.18)	2.37 (1.78)	0.73***	0.49	.61	.07	.35*	.16	.03	.29*
DAS-II Recall of Objects Ability Score	74.05 (28.99)	81.92 (31.44)	7.87*	0.26	.66	.31*	.34**	.49***	.40**	.53***
DAS-II Recall of Objects T-score[b]	13.83 (5.88)	15.78 (7.48)	1.95**	0.29	.67	−.01	.44**	.24	.16	.35**
CMS Dot Locations Total Raw Score[c]	53.59 (13.39)	57.61 (13.04)	4.02*	0.30	.58	.38**	.16	.36**	.32*	.36*
CMS Dot Locations Learning Raw Score [c]	54.03 (13.82)	57.08 (12.47)	3.05	0.23	.14	.39**	.05	.29*	.26	.29*
CMS Dot Locations Short Delay Raw Score [c]	52.88 (15.73)	54.41 (19.04)	1.53	0.08	.53	.21	.43**	.45**	.38**	.45**
CMS Dot Total Scaled Score[b]	3.63 (2.39)	4.10 (2.52)	0.47	0.19	.51	−.04	.07	.15	.11	.17
CMS Dot Learning Scaled Score[b]	3.38 (2.40)	4.08 (2.37)	0.70	0.29	.41	.10	.06	.15	.12	.18
CMS Dot Short Delay Scaled Score[b]	5.12 (2.24)	5.44 (2.66)	0.32	0.13	.59	−.27	.32*	.25	.24	.17
CANTAB PAL First Attempt Memory	5.91 (4.02)	6.41 (4.54)	0.50	0.12	.68	.36	−.50*	.06	.07	.03
CANTAB PAL Mean Errors to Success[b]	1.76 (1.90)	2.29 (1.35)	0.53	0.33	.39	.19	−.43	.01	.14	−.14
CANTAB PAL Total Errors Adjusted	44.41 (18.29)	39.32 (20.42)	5.09	0.26	.68	−.37	.41	−.08	−.06	−.10
OMQ-PF	87.75 (11.98)	87.06 (10.84)	0.69	0.06	.70	.12	.20	.18	.15	.19

*$p<.05$

**$p<.01$

*** $p<.001$;

[a] Stanford-Binet, fifth edition deviation/z scores;

[b] Spearman correlation used for associations with age, VABS-3, and SB-5;

[c] CMS Dot Locations raw scores reported as percentage correct because of the varying grid sizes that correspond with chronological age; VABS-3 ABC = Vineland Adaptive Behavior Scale, third edition Adaptive Behavior Composite; NEPSY-II = Developmental Neuropsychological Assessment, second edition; DAS-II = Differential Abilities Scale, second edition; CMS = Children's Memory Scale; CANTAB PAL = Cambridge Neuropsychological Test Automated Battery Paired Associate Learning; OMQ-PF = Observer Memory Questionnaire-Parent Form

Res Dev Disabil. Author manuscript; available in PMC 2023 January 01.

Table 3.

Convergent validity correlations at Time 1

	1	2	3	4	5	6	7	8
1. NEPSY-II List Memory Total Correct								
2. DAS-II Recall of Objects Ability Score	.68**							
3. CMS Dot Locations Total Raw Score	.36*	.30*						
4. CMS Dot Locations Learning Raw Score	.35*	.28*	.97**					
5. CMS Dot Locations Short Delay Raw Score	.23	.26	.78**	.61**				
6. CANTAB PAL First Attempt Memory	−.02	.31	.54*	.53*	.49*			
7. CANTAB PAL Mean Errors to Success	.27	.20	.16	.14	.12	.53**		
8. CANTAB PAL Total Errors Adjusted	−.17	−.39	−.61**	−.59**	−.56*	−.97**	−.57**	
9. OMQ-PF	.07	−.10	.02	.01	.10	.13	−.01	−.17

* $p < .05$;

** $p < .01$;

NEPSY-II = Developmental Neuropsychological Assessment, second edition; DAS-II = Differential Abilities Scale, second edition; CMS = Children's Memory Scale; CANTAB PAL = Cambridge Neuropsychological Test Automated Battery Paired Associate Learning; OMQ-PF = Observer Memory Questionnaire-Parent Form

Table 4.

Sensitivity and specificity for measures below feasibility criterion

	NEPSY-II List Memory				CMS Dot Locations				CANTAB PAL			
	Sensitivity		*Specificity*		*Sensitivity*		*Specificity*		*Sensitivity*		*Specificity*	
	Age 8	**Age 10**	**Age 8**	**Age 10**	**Age 8**	**Age 10**	**Age 8**	**Age 10**	**Age 8**	**Age 10**	**Age 8**	**Age 10**
No ABIQ Deviation Restriction	94%	88%	9%	18%	94%	92%	23%	41%	86%	86%	10%	19%
ABIQ Deviation ≥ 20	94%	88%	35%	43%	92%	90%	43%	59%	86%	86%	20%	29%
ABIQ Deviation ≥ 30	73%	71%	80%	81%	68%	68%	76%	82%	67%	67%	50%	53%
ABIQ Deviation ≥ 40	38%	38%	95%	95%	36%	36%	95%	95%	33%	33%	76%	76%
ABIQ Deviation ≥ 50	15%	15%	100%	100%	14%	14%	100%	100%	19%	19%	94%	94%

NEPSY-II = Developmental Neuropsychological Assessment, second edition; CMS = Children's Memory Scale; CANTAB PAL = Cambridge Neuropsychological Test Automated Battery Paired Associate Learning

Table 5.

Psychometric evaluation summary for short-term memory measures

	Minimal floor effects	Feasibility	Test-retest	Negligible practice effects	Convergent validity
NEPSY-II List Memory Total Correct	–	–	–	–	With Recall of Object
NEPSY-II List Memory Scaled Score[a]	–	–	–	–	
DAS-II Recall of Objects Ability Score	–	+	–	–	With List Memory
DAS-II Recall of Objects T-score[a]	–	+	–	–	
CMS Dot Locations Total Raw Score	–	–	–	–	With CANTAB PAL
CMS Dot Locations Learning Raw Score	–	–	–	+	With CANTAB PAL
CMS Dot Locations Short Delay Raw Score	–	–	–	+	With CANTAB PAL
CMS Dot Locations Total Scaled Score[a]	–	–	–	–	
CMS Dot Locations Learning Scaled Score[a]	–	–	–	+	
CMS Dot Locations Short Delay Scaled Score[a]	–	–	–	–	
CANTAB PAL First Attempt Memory	–	–	–	+	With CMS Dot
CANTAB PAL Mean Errors to Success	–	–	–	–	–
CANTAB PAL Total Errors Adjusted	–	–	–	–	With CMS Dot
OMQ-PF	+	+	–	+[b]	–

+ indicates study criterion met: < 20% floor effects, ≥ 80% feasibility, ≥ .75 test-restest ICC, small and non-significant practice effects

– indicates study criterion not met: ≥ 20% floor effects, < 80% feasibility, < .75 test-retest ICC, medium/large and significant practice effects

[a] Scaled scores not assessed for convergent validity;

[b] Agreement between T1 and T2