Received: 30 August 2016 | Revised: 24 July 2017 | Accepted: 29 July 2017
DOI: 10.1002/ajmg.a.38416

RESEARCH REVIEW

WILEY AMERICAN JOURNAL OF medical genetics PART A

Challenges in measuring the effects of pharmacological interventions on cognitive and adaptive functioning in individuals with Down syndrome: A systematic review

Lori A. Keeling[1] | Gail A. Spiridigliozzi[2] | Sarah J. Hart[1] | Jane A. Baker[1] | Harrison N. Jones[3] | Priya S. Kishnani[1]

[1] Department of Pediatrics, Duke University Medical Center, Durham, North Carolina

[2] Department of Psychiatry and Behavioral Sciences, Duke University Medical Center, Durham, North Carolina

[3] Department of Surgery, Duke University Medical Center, Durham, North Carolina

Correspondence
Gail A. Spiridigliozzi, PhD, Department of Psychiatry and Behavioral Sciences, Duke University Medical Center, Box 3364, Durham, NC 27710.
Email: gail.spiridigliozzi@duke.edu

We systematically reviewed the measures used in pharmaceutical trials in children/adults with Down syndrome without dementia. Our purpose was to identify developmentally appropriate outcome measures capable of detecting changes in cognitive and adaptive functioning in this population. Eleven studies were included and used diverse outcome measures across the domains of language, memory, attention, behavior, and executive/adaptive functioning. Our results highlight the challenges in selecting measures capable of capturing improvements in pharmaceutical trials in individuals with DS. We offer suggestions to enhance future research, including: conducting studies with larger samples of participants with a range of developmental abilities; modifying existing/developing novel outcome measures; incorporating advances from related areas and DS observational studies; and considering alternative analytic techniques to characterize treatment effects.

KEYWORDS
cognitive assessment, developmentally appropriate outcome measures, Down syndrome, treatment effects

1 | MEASURING TREATMENT EFFECTS IN DOWN SYNDROME

Individuals with Down syndrome (DS) demonstrate overall deficits in cognitive and adaptive functioning, including language, learning, and memory impairments. In general, visuospatial processing and implicit long-term memory are relative cognitive strengths, while working memory, episodic long-term memory, expressive language, and executive function are relative cognitive weaknesses (Liogier d'Ardhuy et al., 2015). Early medical interventions in individuals with DS prioritized the treatment of relatively more acute conditions, such as gastro-intestinal disorders (e.g., duodenal stenosis/atresia and Hirschsprung disease) (Freeman et al., 2009), heart-related conditions (Freeman et al., 2008), thyroid disorders (Graber, Chacko, Regelmann, Costin, & Rapaport, 2012), and acute leukemia (Seewald, Taub, Maloney, & McCabe, 2012). However, there has been increased interest in the use of pharmacological agents targeting cholinergic or GABAergic systems to enhance cognitive function and activities of daily living (Liogier d'Ardhuy et al., 2015).

The majority of pharmaceutical clinical trials in DS to date have targeted cholinergic function, as data suggest cognitive deficits in DS may be linked to inherent cholinergic dysfunction with known abnormalities in peripheral and central cholinergic functions (Beccaria et al., 1998). Central cholinergic functions are essential for cognitive performance (e.g., memory, attention) and balanced mood (Casanova, Walker, Whitehouse, & Price, 1985). DS is also associated with reduced cholinergic neurons (Casanova et al., 1985) and altered cortical connectivity (Becker, Mito, Takashima, & Onodera, 1990;

The authors wish to dedicate this manuscript to the memory of James H. Heller, whose expertise, passion, and commitment to the field of Down syndrome research continues to inspire our work.

Berger-Sweeney, 2003). These findings suggest that interventions targeting cholinergic function may enhance neural connectivity and improve cognitive functioning (Becker et al., 1990; Berger-Sweeney, 2003). Accordingly, Kishnani, Sullivan, et al. (1999) initially investigated the safety and efficacy of a cholinesterase inhibitor (ChEI; i.e., donepezil hydrochloride) for the treatment of cognitive functioning in adults with DS without dementia/Alzheimer's disease (AD). Although this agent had previously received FDA approval for the treatment of AD, this study was the first in a series of studies investigating the use of a pharmacologic agent in individuals with DS with a scientific rationale to enhance cognition.

Since this investigation, a number of other clinical trials have explored the use of ChEIs, NMDA antagonists (Boada et al., 2012) and inverse GABA agonists (e.g., ClinicalTrials.gov identifiers: NCT02484703 and NCT02024789) with children, adolescents, and adults with DS without dementia (see Figure 1). Early open-label trials provided preliminary data in support of the cognitive benefits from ChEI intervention (Heller et al., 2003; Heller et al., 2004; Heller, Spiridigliozzi, Crissman, Sullivan et al., 2006; Kishnani, Sullivan et al., 1999; Spiridigliozzi et al., 2007), leading to double-blind, placebo-controlled efficacy trials (Boada et al., 2012; Johnson, Fahey, Chicoine, Chong, & Gitelman, 2003; Kishnani et al., 2009; Kishnani et al., 2010; Spiridigliozzi et al., 2016). Data from this body of research have yielded mixed results in terms of the treatment effects of pharmacological agents in individuals with DS. A primary challenge has been the selection of appropriate outcome measures to capture potential changes in clinical, cognitive, and adaptive functioning in individuals with DS. Most measures employ normative values based on data from typically developing children and adults, and were not developed to assess functioning in individuals with DS. In this paper, we systematically review the outcome measures used in clinical trials to date (by domain) and assess the suitability of these measures for future work.

2 | METHODS

2.1 | Eligibility criteria

Studies eligible for this review were completed pharmaceutical trials in participants with DS where the primary study outcomes included measures of cognitive/adaptive functioning. Studies were excluded if there was evidence subjects had dementia/AD, the participants' mean age was greater than 35 years old, and/or >25% of participants were above age 35, to avoid the potential impact of emerging symptoms of dementia or AD on the outcome. In addition, studies of a non-pharmaceutical treatment (e.g., high dose vitamins) or a non-Federal Drug Administration approved agent (e.g., piracetam) were excluded from this review.

2.2 | Search strategy

We searched Medline/PubMED, PsychINFO, BIOSIS Citation Index, Data Citation Index, and Web of Science, Core Collection to find all pharmaceutical intervention trials in adults and children with DS, without dementia, that targeted improving cognitive and adaptive functioning. Open-trials and randomized, placebo-controlled clinical trials that were published from 1980 through December 2015 were included. The search terms used were: "Down syndrome OR trisomy 21" AND "pharmaceutical, clinical, drug, treatment, open-trial, trial, study, OR intervention," AND "longitudinal, efficacy, effectiveness, OR effect," AND "cognitive, language, memory, OR adaptive functioning."

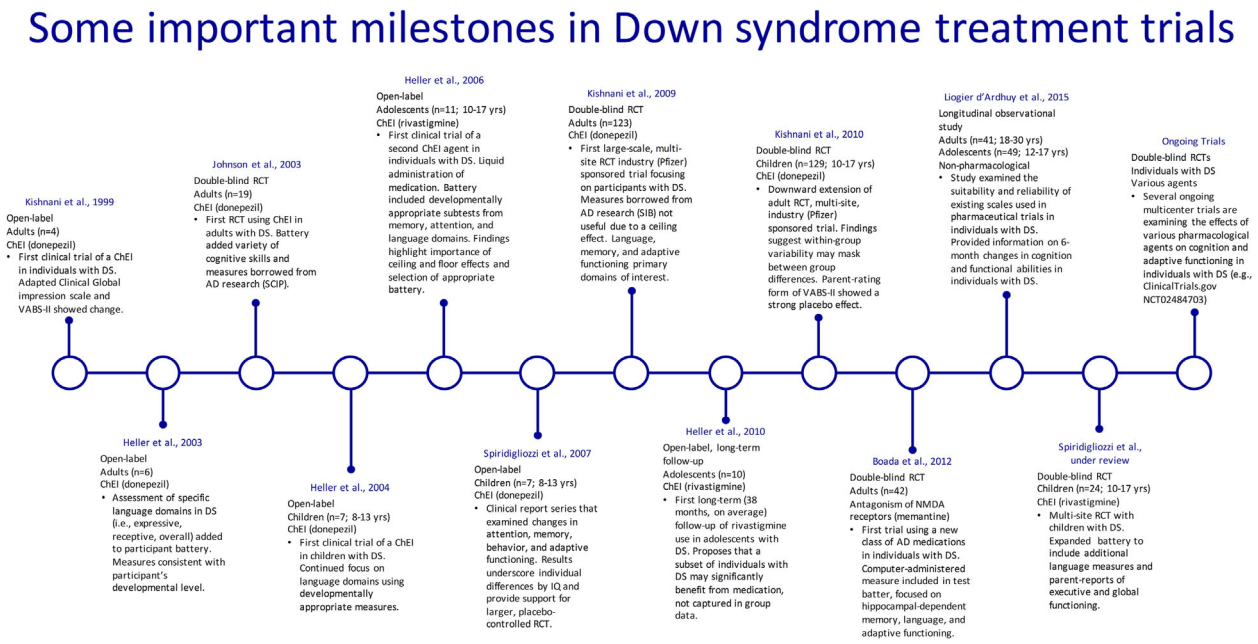

FIGURE 1 Some important milestones in Down syndrome treatment trials

2.3 | Study selection process

Judgement of study eligibility for inclusion was performed by a consensus review of the literature search results by all authors.

2.4 | Data collection process

For each study, we extracted descriptive data (e.g., age, gender), study characteristics (e.g., study type, intervention agent, treatment duration, assessment interval), and outcome measures (e.g., assessment battery, specific subtest(s), developmental level of measures). Outcomes were further defined by type (i.e., clinician assessment, parental-report, teacher-report, direct performance measure) and domain (e.g., memory, attention, adaptive functioning, attention, language). For study type, outcome measures were considered to be clinician assessments if the measure was completed by a study physician/clinician and included an evaluation of participants' functioning. The parental-report and teacher-report measures were assessments completed by parents or teachers, respectively, evaluating the participants' typical functioning. Direct performance measures were those assessments administered by trained study raters to the participants with DS.

3 | RESULTS

Eleven studies met the inclusion/exclusion criteria and used a variety of measures, including parental-reports of adaptive and behavioral functioning, clinician global assessments, and standardized direct/behavioral measures of cognitive functioning. Across these 11 studies, the outcome measures most commonly targeted the domains of language, memory/attention, and adaptive functioning (see Table 1).

3.1 | Overview of useful measures

Based on our historical and comprehensive review, we identified subtests/measures in the domains of language, memory/attention, and adaptive functioning that appeared to be useful in detecting a potential change in response to an intervention. This was based on whether or not the study participants with DS could actually complete the measure, the absence of floor or ceiling effects, and whether or not there was any difference between the placebo and treatment groups on the measure.

3.2 | Language domain

Subtests from several editions of the Clinical Evaluation of Language Fundamentals (CELF) assessment have been more sensitive to change in trials than other standardized language measures reported, such as the Test of Problem Solving (TOPS) (Zachman, Jorgensen, Huisingh, & Barrett, 1984), Differential Abilities Scale-II (DAS-II) (Elliott, 2007), Peabody Picture Vocabulary Test-Third Edition (PPVT-III) (Dunn & Dunn, 1997), and the Test of Reception of Grammar, Version 2 (TROG-II) (Bishop, 2003) with some notable caveats. The Clinical Evaluation of Language Fundamentals-Preschool (CELP-P) Expressive Language Score and Total Language Score (Wiig, Secord, & Semel, 1992) detected treatment effects among adolescents in an open trial (Heller, Spiridigliozzi, Crissman, Sullivan, et al., 2006), with some evidence of ceiling effects, but was not sensitive to performance changes in a longer-term follow-up study (Heller et al., 2010), or when treatment group improvements were compared to placebo or comparison group changes (Kishnani et al., 2010; Spiridigliozzi et al., 2016). The Clinical Evaluation of Language Fundamentals-Third Edition (CELF-3) (Semel, Wiig, & Secord, 1995) Receptive and Expressive Language Scores and Total Score demonstrated utility in detecting treatment effects in an open-label trial among children (Heller et al., 2004), but not in a large randomized controlled trial (RTC) with adults (Kishnani et al., 2009), which found a substantial number of participants at the floor. The CELF-Revised (CELF-R) (Semel, Wiig, & Secord, 1986) was not useful in detecting changes in an open-label trial in adults (Heller et al., 2003). One noteworthy non-standardized measure, the Test of Verbal Expression and Reasoning (TOVER) (Heller, Spiridigliozzi, & Kishnani, 2000), was used in four trials to assess expressive language (Heller, Spiridigliozzi, Crissman, Sullivan, et al., 2006; Heller et al., 2010; Kishnani et al., 2010; Spiridigliozzi et al., 2016). Although it only detected treatment effects in one open-label trial among adolescents (Heller, Spiridigliozzi, Crissman, Sullivan, et al., 2006), it is mentioned here as a potentially useful non-standardized measure for future development and research in DS and other neurodevelopmental disorders.

3.3 | Memory/attention domains

Six of the eleven trials included at least one assessment of memory/attention, with mixed results found across studies. Subtests from the NEPSY: A Developmental Neuropsychological Assessment (NEPSY) (Korkman, Kirk, & Kemp, 1998) were relatively more sensitive to change than the other standardized measures of attention/memory reported (i.e., Leiter International Performance Scale-Revised (Leiter-R) (Roid & Miller, 1997); Rivermead Behavioral Memory Test for Children (RBMT-C) (Wilson, Ivani-Chalian, & Aldrich, 1991); Severe Impairment Battery (SIB) (Saxton, McGonigle, Swihart, & Boller, 1993); Cambridge Neuropsychological Test Automated Battery (CANTAB) (Luciana & Nelson, 2002); California Verbal Learning Test-Second Edition (CVLT-II) (Delis, Kramer, Kaplan, & Ober, 2000). Specifically, the Narrative Memory and Memory for Names subtests from the NEPSY detected treatment effects in two open trials with adolescents (Heller, Spiridigliozzi, Crissman, Sullivan, et al., 2006; Spiridigliozzi et al., 2007). The NEPSY Visual Attention subtest also showed improvement in one of these trials (Spiridigliozzi et al., 2007). Of note, the NEPSY memory subtests were not sensitive to improvements in the long-term follow-up study with adolescents (Heller et al., 2010). Finally, subtests from the Leiter-R were useful in detecting treatment effects in attention, but not memory; and were used to identify a developmentally appropriate attention measure for adolescents with DS (Heller, Spiridigliozzi, Crissman, Sullivan, et al., 2006). Specifically, the Leiter-R Attention Sustained B subtest (normed for typically-developing

TABLE 1 Pharmaceutical trials in individuals with Down syndrome using outcome measures of cognitive/adaptive functioning

Study	Study design	Drug	N	Gender (*M* = Male, *F* = Female)	Age range in years	Language	Memory	Attention	Behavior/ executive functioning	Adaptive functioning	Clinical global
Kishnani, Sullivan, et al. (1999)	Prospective, open-label	Donepezil	4	*M* = 3, *F* = 1	24-64				Caregiver diary^{P}	VABS^{P-I}	PGI^{C}
Heller et al. (2003)	Prospective, open-label	Donepezil	6	*M* = 5, *F* = 1	20-41	TOPS^{B} CELF-R^{B}					
Johnson et al. (2003)	RCT, double-blind, placebo-controlled	Donepezil	19	*M* = 11, *F* = 8	17-50				SCIP (AD)$^{P-Q}$ SIB-R(SP)$^{P-Q}$	SCIP (AD)$^{P-Q}$	
Heller et al. (2004)	Prospective, open-label	Donepezil	7	*M* = 2, *F* = 5	8-13	TOPS^{B} CELF-3^{B}					
Heller, Spiridigliozzi, Crissman, Sullivan, et al. (2006)	Prospective, open-label	Rivastigmine	11	*M* = 8, *F* = 3	10-17	CELF-P^{B} TOVER (DS)^{B}	Leiter-R^{B} NEPSY^{B}	Leiter-R^{B} NEPSY^{B}		VABS-II^{P-I}	CIBIS/ CIBIC^{C}
Spiridigliozzi et al. (2007)	Prospective, open-label	Donepezil	7	*M* = 2, *F* = 5	8-13		NEPSY^{B}	NEPSY^{B}	Conners-R^{P-Q}	VABS-II^{P-I}	
Kishnani et al. (2009)	RCT, double-blind, placebo-controlled	Donepezil	123	Drug: *M* = 38, *F* = 24 Placebo: *M* = 39, *F* = 22	18-35	CELF-3^{B}	RBMT-C^{B}	SIB(AD)$^{P-Q}$		VABS-II^{P-I}	
Heller et al. (2010)	Long-term follow-up of open-label	Rivastigmine	10	*M* = 8, *F* = 2	12-22	CELF-P^{B} TOVER (DS)^{B}	Leiter-R^{B} NEPSY^{B}	Leiter-R^{B} NEPSY^{B}		VABS-II^{P-I}	
Kishnani et al. (2010)	RCT, double-blind, placebo-controlled, industry funded	Donepezil	129	Drug: *M* = 36, *F* = 26 Placebo: *M* = 30, *F* = 35	10-17	TOVER (DS)^{B}				VABS-II^{P-I}	
Boada et al. (2012)	RCT, double-blind, placebo-controlled	Memantine	38	Drug: *M* = 7, F = 12 Placebo *M* = 7, *F* = 12	18-32	DAS-II (SP)^{B} PPVT-III^{B} TROG-II^{B}	CANTAB^{B} CVLT-II^{B} RBMT-C^{B} DAS-II (SP)^{B}		SIB-R (SP)$^{P-Q}$	DAS-II (SP)^{B} SIB-R (SP)$^{P-Q}$	
Spiridigliozzi et al. (2016)	RCT, double-blind, placebo-controlled	Rivastigmine	22	Drug: *M* = 4, *F* = 8 Placebo: *M* = 4, *F* = 6	10-17	CELF-P^{B}-2 TOVER (DS)^{B}	NEPSY^{B}	NEPSY^{B}	BRIEF-P^{P-Q} REAL (DS)$^{P-Q}$	VABS-II^{P-I}	REAL (DS)$^{P-Q}$

Measures are BRIEF-P, Behavior Rating Inventory of Executive Function-Preschool Version; CANTAB, Cambridge Neuropsychological Test Automated Battery; CELF-R, Clinical Evaluation of Language Fundamentals—Revised; CELF-P, Clinical Evaluation of Language Fundamentals—Preschool; CELF-P-2, Clinical Evaluation of Language Fundamentals-Preschool-2; CELF–3, Clinical Evaluation of Language Fundamentals–Third Edition; CVLT-II, California Verbal Learning Test-II; CIBIS/CIBIC, Clinician's Interview-Based Impression of Severity/Change; Conners-R, Conners' Parent Rating Scale-Revised; DAS-II, Differential Abilities Scale-II; Leiter-R, Leiter International Performance Scale-Revised; NEPSY, NEPSY: A Developmental Neuropsychological Assessment; PGI, Physician Global Impression; PPVT-III, Peabody Picture Vocabulary Test-Third Edition RBANS, Repeatable Battery for the Assessment of Neuropsychological Status; RBMT-C, Rivermead Behavioral Memory Test for Children; REAL, Rating of Everyday Activities and Life Skills; SCIP, Severe Cognitive Impairment Profile; SIB, Severe Impairment Battery; SIB-R, Scales of Independent Behavior-Revised; TROG-II, Test of Reception of Grammar, Version 2; TOVER, Test of Verbal Expression and Reasoning; TOPS, Test of Problem Solving; VABS, Vineland Adaptive Behavior Scales; VABS-II, Vineland Adaptive Behavior Scales-Second Edition, Interview Edition (Survey Form).
Coding for measures: ^{B}, behavioral/direct performance measure; ^{P}, parent report measure; ^{C}, clinician assessment; ^{I}, interview; ^{Q}, questionnaire/rating form/checklist; (AD), Assessment used to test treatment effects in Alzheimer's Disease; (DS), Measure was developed specifically for use with individuals with DS; (SP), Special populations included in norms, for example, individuals with developmental disabilities.

children ages 4–5) had no ceiling effects, and was sensitive enough to detect improvements in attention among adolescents with DS aged 10–17 (Heller, Spiridigliozzi, Crissman, Sullivan, et al., 2006). In contrast, the Leiter-R Attention Sustained A subtest (normed for typically-developing children ages 2–3) evidenced a ceiling effect at baseline among this sample (Heller, Spiridigliozzi, Crissman, Sullivan, et al., 2006).

3.4 | Adaptive functioning domain

The parent interview versions of the Vineland Adaptive Behavior Scales (VABS) (Sparrow, Balla, & Cicchetti, 1984) and the Vineland Adaptive Behavior Scales, Second Edition (VABS-II) (Sparrow, Cicchetti, & Balla, 2005) were the most useful measures of adaptive behavior across studies (i.e., Adaptive Behavior Composite score and Communication, Daily Living Skills, and Socialization domain standard scores). These measures demonstrated utility in detecting treatment effects among adults and adolescents participating in RCTs, with no evidence of ceiling or floor effects (Kishnani et al., 2009; Spiridigliozzi et al., 2016). The fact that the VABS and VABS-II scales were designed for individuals from birth through age 90 is particularly helpful. However, it is notable that performance on the VABS-II may show significant variability and/or placebo effects, as our recent RCT in children and adolescents with DS noted relative improvement on VABS-II scores in the group receiving placebo, but not the treatment group (Spiridigliozzi et al., 2016).

4 | DISCUSSION

For nearly two decades (e.g., [Kishnani et al., 1999]), researchers have sought to evaluate the treatment effects of drug therapies designed to improve a variety of cognitive domains in individuals with DS. The current review highlights the challenges faced in selecting outcome measures that are valid, reliable, and sensitive to change in individuals with DS. Despite these challenges, research teams have attempted to overcome existing limitations and barriers by: (i) collaborating across diverse disciplines, institutions, and industries to foster novel ways of approaching DS treatment; (ii) engaging parents, caregivers, and participants in the research process, and modifying research questions and approach based on stakeholder feedback; and (iii) utilizing best-available assessment batteries/subtests, and when those were determined insufficient, designing novel measures specific to the needs of individuals with DS.

Additionally, researchers have sought to improve existing batteries by examining the suitability and reliability of outcome measures in individuals with DS designed to capture the neurocognitive phenotype of DS (de Sola et al., 2015; Edgin et al., 2010; Liogier d'Ardhuy et al., 2015) and natural changes over time (Edgin et al., 2010; Liogier d'Ardhuy et al., 2015). Edgin et al. (2010) evaluated a multi-informant neurocognitive assessment battery (Arizona Cognitive Test Battery) with a sample of individuals with DS (n = 74, ages 7–38) and mental-age matched controls (n = 50, ages 7–38). This battery included tasks (such as nonverbal subtests from the CANTAB) targeting prefrontal, hippocampal, and cerebellar functioning relevant to individuals with DS, as identified in neuroimaging studies. The authors also retested a subset of individuals with DS (n = 10) to examine changes over a 1.55 year interval and to estimate test-retest reliability. In addition, de Sola and colleagues (de Sola et al., 2015) developed a cognitive and functional assessment battery (TESDAD Battery) for individuals with DS (n = 86, ages 16–34), for use in their RTC of a green tea dietary supplement for this population. Liogier d'Ardhuy et al. (2015) examined several outcome measures (e.g., subtests from the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) (Randolph, Tierney, Mohr, & Chase, 1998) used in DS pharmacological studies by following a cohort of individuals with DS longitudinally (n = 89, ages 12–30), to investigate developmental changes over time, test-retest reliability, and practice effects.

Even with these contributions, numerous gaps and unanswered questions remain. It is unclear to what extent study participants with DS exhibit fluctuating level of task engagement and performance on study measures, as described by Wishart (Wishart, 1995). The inclusion of a visit prior to the baseline visit for participants to become familiar with study personnel and the setting may be helpful in reducing any anxiety and promoting optimal performance. Another factor to consider is that even the large RTCs completed to date may not be sufficiently powered to detect small (but significant) changes in individuals with DS. It is also possible that the treatments themselves are not having a beneficial effect on the older children, adolescents and adults who have been included in the RTCs. It may be necessary to include younger children with DS in these trials to see a benefit from treatments targeting improved cognition.

4.1 | Evolution of assessment batteries

Figure 1 illustrates the experience of the Duke Down Syndrome Research Team and contributions from other groups from an historical perspective in developing study assessment batteries. Measures used to assess treatment effects were first informed by theory and hypothesized mechanisms of action, then expanded based on qualitative observations made by parents, clinicians, and team members comprising the Duke Down syndrome Research Team. These observations and direct experience administering the measures with individuals with DS suggested that the initial parsimonious battery assembled by our group did not fully capture the range of perceived/observed changes in language and cognitive functioning. After standardized language and cognitive measures were added, our research team continued to expand the assessment batteries, which typically included a combination of participant performance measures, observer-reported measures (i.e., parent and/or teacher-report), and clinician assessments of functioning. The goal was to capture the range of potential treatment effects. With the addition of more standardized neurocognitive measures, it became clear that existing measures were inadequate at capturing the impact of treatment in participants with DS, and specific subtests needed to be adjusted/included to more appropriately match the actual skill levels of participants. For example,

when our team was identifying an appropriate measure of attention, we selected Attention Sustained A from the Leiter-R battery, which was designed for typically developing children aged 2–3 years (Heller et al., 2010). However, when we used this measure with our adolescent participants with DS, we observed significant ceiling effects (Heller, Spiridigliozzi, Crissman, Sullivan, et al., 2006). Furthermore, we noted substantial floor effects using the CELF-3 in our RCT among adults with DS (Kishnani et al., 2009), but not in our open trial with children (Heller et al., 2004). These examples and the evolution of assessment batteries in DS research highlight the importance of selecting developmentally appropriate measures, and the difficulty in achieving this aim when using standardized tests designed for typically-developing individuals.

4.2 | Evolution of the language domain

In the first open-label treatment trial in DS, caregivers reported improvements in parent-child communication and adaptive behavior via study diaries and interviews (e.g., child more talkative about day at school with parents, or child more receptive to working on homework after school) (Kishnani, Sullivan, et al., 1999). Parents also noted it was difficult to describe/quantify the changes in communication they had observed in their homes via the existing measures selected by the research team. Parents' narrative reports led to an expansion of our test battery, with measures added (such as the TOPS and CELF) to specifically probe receptive and expressive language domains (Heller et al., 2003; Heller et al., 2004). In our open trial among adults with DS, the TOPS appeared useful in detecting improvements in language (Heller et al., 2003), however, it did not appear sensitive enough to detect treatment effects in our open trial among children (Heller et al., 2004), and we subsequently discontinued use of this measure in our studies. These early studies helped solidify a focus on the language domain in DS treatment trials, but as no gold standard measure existed for use with this population, subsequent studies continued to include alternate ways to investigate improvements in language, using direct performance and parental-report measures. We continued to explore the language domain using the CELF, and (later) the TOVER, (Heller et al., 2000), while Boada et al. (2012) explored the language domain using the DAS-II, the PPVT-III and the TROG-II. The most commonly used language measure was the CELF.

Overall, findings across our six DS treatment trials suggest utility in the CELF's expressive subtests, with mixed results for the receptive subtests and for the different versions of the test. Across these trials, we included three versions of the CELF (CELF-R, CELF-3; CELF-P—each normed for different age ranges), and examined a variety of individual subtest scores (Expressive Language subtests such as Word Structure, Formulating Sentences, Recalling Sentences; and Receptive Language subtests, such as Sentence Structure, Concepts and Directions, and Word Classes), domain scores (Expressive Language, Receptive Language), and Total Language Scores (Heller et al., 2003; Heller et al., 2004; Heller et al., 2010; Heller, Spiridigliozzi, Crissman, Sullivan, et al., 2006; Kishnani et al., 2009; Spiridigliozzi et al., 2007; Spiridigliozzi et al., 2016). The CELF-R, which was normed for typically developing individuals aged 5–17 years old, was not an appropriate fit for our adult participants with DS (Heller et al., 2003). Specifically, we identified differences in CELF-R subtest scores by language level and IQ, such that participants with higher language levels and IQs scored much higher on the CELF-R subscales at baseline. A floor effect was also apparent for all participants in this study on the Sentence Assembly subtest. Similarly, a significant number of young adults with DS scored at the floor on the CELF-3 subtests in a large RCT of donepezil (Kishnani et al., 2009).

In contrast, the CELF-P was the most sensitive version of the test, across our six trials. Mainly, the CELF-P detected treatment effects among our adolescents (aged 10–17) in the open-label trial (Heller, Spiridigliozzi, Crissman, Sullivan, et al., 2006), and revealed significant within-group effects in our adolescent RCT on: (i) the Word Classes subtest, Expressive Language Scores and Total Scores, among the treatment arm; and (ii) a Recalling Sentences subtest score among the placebo arm (Spiridigliozzi et al., 2016). It is notable that between-group differences on the CELF-P in our RCT with adolescents may have not been detectable due to limitations in statistical power and large within-group variability. Taken together, results from those trials including standardized measures of language highlight the importance of probing expressive and receptive language domains in DS interventions. However, these findings also underscore the fact that language improvements may not be fully captured in participants with DS, using standardized measures alone. Future studies should aim to enroll larger numbers of participants, with more specific inclusion criteria regarding their baseline expressive and receptive language abilities designed to minimize the effects of within-group variability (and potentially mask treatment effects).

In an effort to overcome the limitations of using existing standardized language measures in DS research, Heller and colleagues (Heller et al., 2000) developed an expressive language measure (TOVER), which was designed to evaluate specific language improvements in individuals with DS. While the TOVER was useful in detecting treatment effects in our 22-week open-trial with children (Heller, Spiridigliozzi, Crissman, Sullivan, et al., 2006), it was not able to capture group differences in three subsequent trials: (i) the long-term follow-up with our child/adolescent sample (Heller et al., 2010); (ii) our large RCT with children (Kishnani et al., 2010); or (iii) our subsequent RCT with children (Spiridigliozzi et al., 2016). However, when we examined individual performance changes in the long-term follow-up study (Heller et al., 2010), we found that scores on the TOVER and CELF-P were not consistent across individuals with DS. Instead, a subset of individuals seemed to benefit from long-term treatment, as measured by the TOVER and CELF-P. Future work should examine the utility of CELF-P and TOVER in detecting language treatment effects among larger, more diverse samples that could identify meaningful subgroups who may benefit differentially from certain treatments.

4.3 | Strategies for future research

As trials of pharmacologic agents continue to target treatment of cognitive functioning in individuals with DS (e.g., RCT examining

effects of memantine, ClinicalTrials.gov identifier: NCT02304302), the following suggestions for future studies are offered with respect to: (i) recruiting diverse participants (specifically those who are non-verbal or minimally verbal) and large samples; (ii) modifying assessment measures/batteries; (iii) incorporating lessons learned from similar fields and observational studies; and (iv) utilizing analytic techniques that best capture treatment effects and accommodate potential study limitations.

The studies in our review were diverse in terms of age and sex, and included participants from childhood through young adulthood, with a balance of male and female subjects. However, the inclusion criterion across studies required participants with DS to be verbal, which was cited as a prerequisite for completing the neurocognitive measures. Future research should aim to include individuals with DS who are non-verbal or have minimal verbal skills and explore measures that can accommodate different communication styles and levels. One example is the Leiter-3 (Roid, 2013), which includes an Attention/Memory Battery in addition to its Cognitive Battery. This non-verbal measure uses gestures, pictures, and manipulatives to assess subjects. Given that a substantial number of individuals with DS are non-verbal, and that DS is a heterogeneous condition, it is important for future work to include large, diverse samples of participants in terms of their language abilities. At the same time, it will also be important to systematically assess the presence of behaviors and comorbid psychiatric conditions (such as attention-deficit/hyperactivity disorder and an autism spectrum disorder) within each sample, as these have the potential to confound the study results. Changes in behavior over time need to be controlled for if improved cognition is the primary outcome measure.

Larger samples may also help investigators identify subgroups who differentially benefit from drug treatments, help in the detection of treatment effects, and prevent the potentially significant placebo effects often found in DS treatment trials. While placebo effects are a hazard in any study (and for which controls need to be in place), in our review, placebo effects were noted in three out of the five RCTs, on one or more of the primary/secondary outcome measures (Kishnani et al., 2009; Kishnani et al., 2010; Spiridigliozzi et al., 2016). Caregivers and families who join our trials often express considerable hope, especially given the substantial unmet need for cognitive treatments in DS. The placebo effects observed in our review are consistent with a recent review of placebo responses in genetically determined intellectual disabilities (Curie et al., 2015). Namely, the authors found significant placebo effects across twenty-two RCTs, and noted a significant effect of age in that there were higher placebo responses in the treatment of younger subjects, and a significant effect of IQ with greater placebo effects among individuals with higher IQ (Curie et al., 2015).

Future research should also consider modifying existing assessment measures to more fully engage participants with DS, to ensure understanding of the tasks being used, and to more accurately characterize the potential cognitive improvements associated with interventions. First, future studies should consider modifying existing measures to accommodate the particular needs of individuals with DS, such as: (i) include more specific instructions at the beginning of each task, with simplified language that is clearly understandable to the participants; (ii) incorporate additional practice/teaching items for each task; (iii) allow examiners to use more re-direction cues when subjects' attention wanes and/or standardize the amount of these cues; (iv) extend review/learning periods for memory-specific stimuli (e.g., extend encoding time from 3 to 5 s); and (v) incorporate lessons learned from investigators working with other conditions in the developmental disabilities field, such as fragile X syndrome, autism spectrum disorder, and Williams syndrome, to understand what measures have been sensitive to treatment effects in these populations.

Second, future research should aim to create new measures or use innovative technologies to assess treatment effects. For example, a novel audio recording device such as the LENA Pro now exists that can record interactions between caregivers and their children in real-time, within ecologically valid settings, such as participants' homes without the direct presence of researchers (Thiemann-Bourque, Warren, Brady, Gilkerson, & Richards, 2014). This type of technology may have the potential to capture subtle changes in day-to-day communication via a non-obtrusive device, which could decrease participant burden, increase utility of the findings, and streamline data collection and analysis by using automatic coding to quantify a variety of outcomes. Standardized cognitive batteries available via the iPad or other tablets should also be explored as potential outcome measures in trials. Most of the participants in our studies are well versed in iPad use, which may increase their attention and engagement with standardized tasks, and potentially increase the accuracy/utility of the results. Finally, innovative technologies, such as cognitive "games" (e.g., CogMed, which can be completed at home/online) could be used as an assessment tool to estimate individual performance trajectories over the course of treatment (Bennett, Holmes, & Buckley, 2013). With multiple data points, these examples have the potential to be more sensitive measures of change and treatment effects.

In addition, assessment data from observational studies in individuals with DS should be collected to understand developmental trajectories in the absence of treatment, to shed light on reliability and validity issues specific to DS (e.g., to determine test-retest reliability thresholds and practice effects for measures that are not normed for individuals with DS), to inform average scores by age group, and to help identify relevant covariates (e.g., IQ). For example, Liogier d'Ardhuy and colleagues (Liogier d'Ardhuy et al., 2015) examined a variety of cognitive scales in their 6-month observational study among adults and adolescents with DS. They found significant floor effects on the Leiter-R, particularly among adults, and suggested future studies should use a different version of the test (Leiter-3). Earlier, we found evidence of ceiling effects on the Leiter-R in our open trial with adolescent (Heller, Spiridigliozzi, Crissman, Sullivan et al., 2006). For the Clinical Evaluation of Language Fundamentals-Preschool-2 (CELF-P-2) (Semel, Wiig, & Secord, 2004), Liogier D'ardhuy and colleagues (Liogier d'Ardhuy et al., 2015) found no evidence of ceiling effects on the Word Classes subtest among adolescents, but found the data were skewed toward the ceiling for adults. Across our three trials that used the CELF-P/CELF-P-2, the only one which showed ceiling effects was the

Total Language score, not Word Classes (Heller, Spiridigliozzi, Crissman, Sullivan, et al., 2006), and although the other two trials did not show ceiling effects on any of the CELF-P/CELF-P-2 subtests, they also did not demonstrate sensitivity to detecting treatment effects (Heller, Spiridigliozzi, Crissman, Sullivan, et al., 2006; Spiridigliozzi et al., 2016). Once again, these findings highlight the importance of determining developmentally appropriate measures across the lifespan in individuals with DS, and the importance of considering findings from non-treatment trials in DS.

Future studies should explore the use of analytic techniques that may better capture treatment effects and accommodate potential study limitations. In the long-term follow-up trial with adolescents (Heller et al., 2010), the authors suggested potential differences may exist in treatment response by subgroups. Future work should incorporate analytic techniques that are able to characterize relevant subgroups, while minimizing the effect of within-group differences. For example, investigators should consider use of "responder analysis" approaches, or mixed methods approaches that estimate the effects of time, treatment arm, and covariates. Other analytic techniques that account for the difficulty in identifying sensitive measures should be explored, such as using a composite measure as a "primary outcome," to take into account changes in cognitive functioning across several domains (e.g., language, adaptive functioning, and memory). In addition, analytic techniques that account for non-continuous data should be employed, as many of the studies reviewed examined the outcome measures as continuous variables, even though the range of scores were limited. Furthermore, studies should consider use of measures that are truly continuous, as these may be more sensitive to detecting treatment effects.

Our group is also obtaining MRI imaging data on children and adolescents with DS. This could potentially serve as a biomarker in future treatment studies. Other potential biomarkers that should be explored include genomic, proteomic, electrophysiological and alternative imaging approaches.

In sum, our findings across eleven drug treatment trials in individuals with DS highlight the importance of identifying developmentally appropriate assessments, across a variety of cognitive domains, which are sensitive to treatment effects. We have identified several useful measures across the language, memory/attention, and adaptive functioning domains, but much work remains in the comprehensive assessment of cognitive improvements if we are to collectively advance outcome research in DS treatment trials. Future research should focus on recruiting diverse participants and large samples, modifying assessment measures, incorporating lessons learned from similar fields and DS observational studies, and utilizing analytic techniques that best capture treatment effects and account for study limitations.

ACKNOWLEDGMENTS

We thank the Anna Michelle Merrills Foundation for Down Syndrome Research for their continued support of the Duke Down Syndrome Research Team. Thank you also to Cindy Li for her assistance with the preparation of this manuscript.

ORCID

Lori A. Keeling http://orcid.org/0000-0003-0151-7047
Gail A. Spiridigliozzi http://orcid.org/0000-0002-4933-1607
Sarah J. Hart http://orcid.org/0000-0003-0974-3209

REFERENCES

Beccaria, L., Marziani, E., Manzoni, P., Arvat, E., Valetto, M. R., Gianotti, L., & Chiumello, G. (1998). Further evidence of cholinergic impairment of the neuroendocrine control of the GH secretion in Down's syndrome. *Dementia and Geriatric Cognitive Disorders*, *9*(2), 78–81.

Becker, L., Mito, T., Takashima, S., & Onodera, K. (1990). Growth and development of the brain in Down syndrome. *Progress in Clinical and Biological Research*, *373*, 133–152.

Bennett, S. J., Holmes, J., & Buckley, S. (2013). Computerized memory training leads to sustained improvement in visuospatial short-term memory skills in children with Down syndrome. *American Journal on Intellectual and Developmental Disabilities*, *118*(3), 179–192. https://doi.org/10.1352/1944-7558-118.3.179

Berger-Sweeney, J. (2003). The cholinergic basal forebrain system during development and its influence on cognitive processes: Important questions and potential answers. *Neuroscience and Biobehavioral Reviews*, *27*(4), 401–411.

Bishop, D. (2003). *Test for Reception of Grammar*-Version 2. San Antonio, TX: Harcourt Assessment.

Boada, R., Hutaff-Lee, C., Schrader, A., Weitzenkamp, D., Benke, T. A., Goldson, E. J., & Costa, A. C. S. (2012). Antagonism of NMDA receptors as a potential treatment for Down syndrome: A pilot randomized controlled trial. *Translational Psychiatry*, *2*, e141. https://doi.org/10.1038/tp.2012.66

Casanova, M. F., Walker, L. C., Whitehouse, P. J., & Price, D. L. (1985). Abnormalities of the nucleus basalis in Down's syndrome. *Annals of Neurology*, *18*(3), 310–313. https://doi.org/10.1002/ana.410180306

Curie, A., Yang, K., Kirsch, I., Gollub, R. L., des Portes, V., Kaptchuk, T. J., & Jensen, K. B. (2015). Placebo responses in genetically determined intellectual disability: A meta-analysis. *PLoS ONE*, *10*(7), e0133316. https://doi.org/10.1371/journal.pone.0133316

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

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (2000). *California verbal learning test* (2nd ed.). San Antonio, TX: Pearson, Retrieved from https://books.google.com/books?id=OfFlmwEACAAJ

Dunn, L. M., & Dunn, L. M. (1997). *PPVT-III: Peabody picture vocabulary test-Third edition*. Circle Pines, MN: American Guidance Service.

Edgin, J. O., Mason, G. M., Allman, M. J., Capone, G. T., Deleon, I., Maslen, C., . . . Nadel, L. (2010). Development and validation of the arizona cognitive test battery for Down syndrome. *Journal of Neurodevelopmental Disorders*, *2*(3), 149–164. https://doi.org/10.1007/s11689-010-9054-3

Elliott, C. D. (2007). *Differential ability scales* (2nd ed.). San Antonio, TX: Harcourt Assessment.

Freeman, S. B., Torfs, C. P., Romitti, P. A., Royle, M. H., Druschel, C., Hobbs, C. A., & Sherman, S. L. (2009). Congenital gastrointestinal defects in Down syndrome: A report from the Atlanta and National Down Syndrome Projects. *Clinical Genetics*, *75*(2), 180–184. https://doi.org/10.1111/j.1399-0004.2008.01110.x

Freeman Sallie, B., Bean, L. H., Allen, E. G., Tinker, S. W., Locke, A. E., Druschel, C., . . . Sherman, S. L. (2008). Ethnicity, sex, and the incidence of congenital heart defects: A report from the National Down Syndrome Project. *Genetics in Medicine: Official Journal of the American College of*

Medical Genetics, *10*(3), 173–180. https://doi.org/10.1097/GIM.0b013e3181634867

Graber, E., Chacko, E., Regelmann, M. O., Costin, G., & Rapaport, R. (2012). Down syndrome and thyroid function. *Endocrinology and Metabolism Clinics of North America*, *41*(4), 735–745. https://doi.org/10.1016/j.ecl.2012.08.008

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

Heller, J. H., Spiridigliozzi, G. A., Crissman, B. G., Sullivan, J. A., Eells, R. L., Li, J. S., … Kishnani, P. S. (2006). Safety and efficacy of rivastigmine in adolescents with Down syndrome: A preliminary 20-week, open-label study. *Journal of Child and Adolescent Psychopharmacology*, *16*(6), 755–765. https://doi.org/10.1089/cap.2006.16.755

Heller, J. H., Spiridigliozzi, G. A., Crissman, B. G., Sullivan-Saarela, J. A., Li, J. S., & Kishnani, P. S. (2006). Clinical trials in children with Down syndrome: Issues from a cognitive research perspective. *American Journal of Medical Genetics Part C, Seminars in Medical Genetics*, *142C*(3), 187–195. https://doi.org/10.1002/ajmg.c.30103

Heller, J. H., Spiridigliozzi, G. A., Doraiswamy, P. M., Sullivan, J. A., Crissman, B. G., & Kishnani, P. S. (2004). Donepezil effects on language in children with Down syndrome: Results of the first 22-week pilot clinical trial. *American Journal of Medical Genetics Part A*, *130A*(3), 325–326. https://doi.org/10.1002/ajmg.a.30184

Heller, J. H., Spiridigliozzi, G. A., Sullivan, J. A., Doraiswamy, P. M., Krishnan, R. R., & Kishnani, P. S. (2003). Donepezil for the treatment of language deficits in adults with Down syndrome: A preliminary 24-week open trial. *American Journal of Medical Genetics Part A*, *116A*(2), 111–116. https://doi.org/10.1002/ajmg.a.10074

Heller, J., Spiridigliozzi, G., & Kishnani, P. (2000). *TOVER: Test of verbal expression and reasoning*. Durham, NC: Duke University Medical Center.

Johnson, N., Fahey, C., Chicoine, B., Chong, G., & Gitelman, D. (2003). Effects of donepezil on cognitive functioning in Down syndrome. *American Journal of Mental Retardation: AJMR*, *108*(6), 367–372. https://doi.org/10.1352/0895-8017(2003)108<367:EODOCF>2.0.CO;2

Kishnani, P. S., Sullivan, J. A., Walter, B. K., Spiridigliozzi, G. A., Doraiswamy, P. M., & Krishnan, K. R. (1999). Cholinergic therapy for Down's syndrome. *Lancet (London, England)*, *353*(9158), 1064–1065. https://doi.org/10.1016/S0140-6736(98) 05285-4

Kishnani, P. S., Heller, J. H., Spiridigliozzi, G. A., Lott, I., Escobar, L., Richardson, S., … McRae, T. (2010). Donepezil for treatment of cognitive dysfunction in children with Down syndrome aged 10-17. *American Journal of Medical Genetics Part A*, *152A*(12), 3028–3035. https://doi.org/10.1002/ajmg.a.33730

Kishnani Priya, S., Sommer, B. R., Handen, B. L., Seltzer, B., Capone, G. T., Spiridigliozzi, G. A., … McRae, T. (2009). The efficacy, safety, and tolerability of donepezil for the treatment of young adults with Down syndrome. *American Journal of Medical Genetics Part A*, *149A*(8), 1641–1654. https://doi.org/10.1002/ajmg.a.32953

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

Liogier d'Ardhuy, X., Edgin, J. O., Bouis, C., de Sola, S., Goeldner, C., Kishnani, P., … Khwaja, O. (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. https://doi.org/10.3389/fnbeh.2015.00300

Luciana, M., & Nelson, C. A. (2002). Assessment of neuropsychological function through use of the Cambridge neuropsychological testing automated battery: Performance in 4- to 12-year-old children. *Developmental Neuropsychology*, *22*(3), 595–624. https://doi.org/10.1207/S15326942DN2203_3

Randolph, C., Tierney, M. C., Mohr, E., & Chase, T. N. (1998). The repeatable battery for the assessment of neuropsychological status (RBANS): Preliminary clinical validity. *Journal of Clinical and Experimental Neuropsychology*, *20*(3), 310–319.

Roid, G. H. (2013). *Leiter international performance scale* (3rd ed.). Wood Dale, IL: Stoelting, Co.

Roid, G. H., & Miller, L. (1997). *Leiter international performance scale—Revised*. Wood Dale, IL: Stoelting, Co.

Saxton, J., McGonigle, K. L., Swihart, A. A., & Boller, F. (1993). *The severe impairment battery*. London: Thames Valley Test Company.

Seewald, L., Taub, J. W., Maloney, K. W., & McCabe, E. R. B. (2012). Acute leukemias in children with Down syndrome. *Molecular Genetics and Metabolism*, *107*(1-2), 25–30. https://doi.org/10.1016/j.ymgme.2012.07.011

Semel, E., Wiig, E. H., & Secord, W. (1986). *Clinical evaluation of language fundamentals, revised*. San Antonio, TX: Psychological Corporation.

Semel, E., Wiig, E. H., & Secord, W. (1995). *Clinical evaluation of language fundamentals* (3rd ed.). San Antonio, TX: Psychological Corporation.

Semel, E., Wiig, E. H., & Secord, W. (2004). *Clinical evaluation of language fundamental—Preschool-2*. San Antonio, TX: Pearson.

Sparrow, S., Cicchetti, D., & Balla, D. (2005). *Vineland adaptive behavior scales* (2nd ed.). Minneapolis: Pearson Assessment.

Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (1984). *Vineland Adaptive Behavior Scales*. Circle Pines, MN: American Guidance Service. Retrieved from https://books.google.com/books?id=Qn1ZnQAACAAJ

Spiridigliozzi, G. A., Hart, S. J., Heller, J. H., Schneider, H. E., Baker, J. A., Weadon, C., … Kishnani, P. S. (2016). 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

Spiridigliozzi, G. A., Heller, J. H., Crissman, B. G., Sullivan-Saarela, J. A., Eells, R., Dawson, D., … Kishnani, P. S. (2007). Preliminary study of the safety and efficacy of donepezil hydrochloride in children with Down syndrome: A clinical report series. *American Journal of Medical Genetics Part A*, *143A*(13), 1408–1413. https://doi.org/10.1002/ajmg.a.31790

Thiemann-Bourque, K. S., Warren, S. F., Brady, N., Gilkerson, J., & Richards, J. A. (2014). Vocal interaction between children with Down syndrome and their parents. *American Journal of Speech-Language Pathology*, *23*(3), 474–485. https://doi.org/10.1044/2014_AJSLP-12-0010

Wiig, E. H., Secord, W., & Semel, E. (1992). *CELF-Preschool: Clinical Evaluation of Language Fundamentals-preschool: Examiner's Manual*. San Antonio, TX: Psychological Corporation. Retrieved from https://books.google.com/books?id=nD-jjwEACAAJ

Wilson, B., Ivani-Chalian, R., & Aldrich, F. (1991). *Rivermead behavioral test for children*. Suffolk, U.K.: Thames Valley Test Company.

Wishart, J. G. (1995). Cognitive abilities in children with Down syndrome: Developmental instability and motivational deficits. *Progress in Clinical and Biological Research*, *393*, 57–91.

Zachman, L., Jorgensen, C., Huisingh, R., & Barrett, M. (1984). *Test of problem solving*. Moline, IL: Linguisystems.

How to cite this article: Keeling LA, Spiridigliozzi GA, Hart SJ, Baker JA, Jones HN, Kishnani PS. Challenges in measuring the effects of pharmacological interventions on cognitive and adaptive functioning in individuals with Down syndrome: A systematic review. *Am J Med Genet Part A*. 2017;173A:3058–3066. https://doi.org/10.1002/ajmg.a.38416