Research in Developmental Disabilities 160 (2025) 104965

Contents lists available at ScienceDirect

Research in Developmental Disabilities

journal homepage: www.elsevier.com/locate/redevdis

The Dice Trails Test: A modified Trail Making Test for children and adults with Down Syndrome

Katja Sandkühler [a,*], Elisabeth Wlasich [a], Hannah Stadler [a], Armelle Müller [a], Lena Heiß [a], Olivia Wagemann [a,b], Georg Nübling [a,b], Thomas Schenk [c], Adrian Danek [a], Johannes Levin [a,b,d], Sandra V. Loosli [a,e,**]

[a] Department of Neurology, University Hospital, Ludwig-Maximilians-University (LMU) Munich, Munich, Germany
[b] German Center for Neurodegenerative Diseases (DZNE) Munich, Munich, Germany
[c] Lehrstuhl für Klinische Neuropsychologie, Ludwig-Maximilians-Universität München, Munich, Germany
[d] Munich Cluster for Systems Neurology (SyNergy), Munich, Germany
[e] Department of Neurology, University Hospital Zurich and University of Zurich, Zurich, Switzerland

ARTICLE INFO

Keywords:
Down syndrome
Trisomy 21
Intellectual disability
Cognitive flexibility
Measurement
Neuropsychology
Cognition

ABSTRACT

Background: Psychometrically sound instruments to assess cognitive flexibility in people with Down Syndrome (DS) are lacking. The Trail Making Test (TMT) is well-established but requires reading letters and numerals, limiting its applicability for people with DS.
Aims: To evaluate the psychometric properties and developmental sensitivity of a newly developed TMT adaptation without letters and numerals – the Dice Trails Test (DTT).
Methods: The DTT was administered to 39 children (8–14 years, 46 % female) and 57 adults (18–57 years, 47 % female) with DS. We evaluated feasibility (proportion of participants completing the task), distributional properties, construct validity, developmental sensitivity, and split-half reliability. Nineteen individuals were reassessed for test-retest reliability. Individuals with DS were compared to typically developing (TD) groups matched on chronological and mental age.
Results: The DTT showed adequate feasibility ($\geq$ 80 %) for individuals with DS and mild intellectual disability (ID), no relevant floor effects, acceptable construct validity, developmental sensitivity, good split-half reliability, and preliminary evidence for good test-retest reliability in DS. DTT performance differed between DS and TD individuals matched on chronological age, but not when matched on mental age.
Conclusions: Although limited in applicability for individuals with DS and moderate ID, the DTT shows potential as a direct measure of cognitive flexibility in DS across a broad age range.

1. Introduction

Down Syndrome (DS, trisomy 21) is the leading cause of intellectual disability (ID; De Graaf et al., 2017). Relative to the global

* Correspondence to: Department of Neurology, University Hospital, Ludwig-Maximilians-University (LMU) Munich, Marchioninistraße 15, 81377 Munich, Germany.
** Correspondence to: Department of Neurology, University Hospital Zurich, Frauenklinikstr. 26, 8091 Zürich, Switzerland.
E-mail addresses: katja.sandkuehler@med.uni-muenchen.de (K. Sandkühler), sandra.loosli@usz.ch (S.V. Loosli).

https://doi.org/10.1016/j.ridd.2025.104965
Received 24 October 2024; Received in revised form 28 February 2025; Accepted 28 February 2025
Available online 17 March 2025

intellectual impairment, DS is associated with a distinct cognitive phenotype with syndrome-specific strengths and weaknesses. In addition to significant language deficits (Grieco et al., 2015), executive functions (EF) have been identified as a notable area of weakness for people with DS. This is evident from comparisons to typically developing (TD) individuals matched on mental age (MA; i.e., matching based on the cognitive level instead of chronological age), showing a reduced performance for the DS group on EF tasks (for a meta-analysis see Tungate & Conners, 2021) or rating scales (Csumitta et al., 2022).

EF encompass higher-level cognitive processes crucial for goal-directed behaviour (Diamond, 2013). Cognitive flexibility (CF; "shifting") is a core EF, involving the ability to switch between rules, tasks, or mental sets (Miyake et al., 2000). CF plays an important role for everyday functioning in individuals with DS, demonstrated through its links to childhood social cognition (Amadó et al., 2016), adaptive behaviour (Sabat et al., 2020), and employment status (Tomaszewski et al., 2018). Moreover, CF deficits and, more generally, executive dysfunction have been recognised as an early symptom of Alzheimer's dementia in elderly individuals with DS. Due to the location of the amyloid precursor protein gene on chromosome 21, people with DS are at very high risk for Alzheimer's Disease (Lautarescu et al., 2017).

The critical role of CF across developmental stages highlights its potential as a target for interventions and clinical trials. A precondition for the evaluation of such interventions is the availability of sound measures to assess CF deficits and changes therein (Esbensen et al., 2017).

1.1. Cognitive flexibility measures in Down syndrome

Most studies assessing CF in DS rely on behavioural rating scales (Csumitta et al., 2022; Daunhauer et al., 2014), rather than performance-based ("direct") measures. A recent meta-analysis on EF in people with DS conducted by Tungate and Conners (2021) identified only 13 studies using performance-based CF tasks, likely due to a lack of feasible and psychometrically acceptable tests for DS individuals (Schworer et al., 2023).

Attempts have been made to use relatively simple or adapted versions of existing tasks to make them more suitable for this population. For example, a modified Category Switching task, which involves verbal switching between semantic categories (e.g., animals and food), has been successfully used in people with DS (Carney et al., 2013). However, DS-specific language difficulties (Grieco et al., 2015) suggests that non-verbal tests might be more appropriate (Edgin et al., 2010). Card sorting tasks like the Dimensional Change Card Sort Test (DCCS, Zelazo, 2006) have been extensively used in DS research (De Sola et al., 2015; Will et al., 2017). Comparable yet simpler alternatives include reverse categorization tasks, where items are initially sorted based on one rule, followed by adjustment to a reversed rule in the post-switch phase (Van Deusen et al., 2023).

The performance in sorting tasks is typically quantified as a total score within a predefined range, like the number of correctly sorted items or responses post-rule switching (Van Deusen et al., 2023; Will et al., 2017). However, fixed scoring ranges (e.g. 0–10 points) inherently reduce response options, thereby limiting the test's ability to capture a broad performance spectrum. Consequently, the test may lack sensitivity to differentiate lower performance levels, leading to clustering near the minimum (floor effects), which is particularly problematic in studies involving individuals with ID (e.g., De Sola et al., 2015; Schworer et al., 2023). Continuous outcome measures, as used in Trail Making Tests, may reduce this risk by providing a wider range of potential scores.

1.2. Trail Making Tests

The most frequently used paper-pencil version of Trail Making Tests (TMT; Reitan & Wolfson, 1993) consists of two parts, each with 25 items: TMT-A involves connecting numbered circles, while TMT-B requires alternating between numbers and letters. Completion time in part A serves as an index of processing speed, whereas part B as a marker of EF, especially CF. Yet, derived scores such as the difference between part A and B (TMT(B-A)) are commonly used to control for factors affecting performance in both parts, including motor speed and visuo-perceptual demands (Bowie & Harvey, 2006).

A limitation is that the TMT requires the ability to read the alphabet. Language is known to be one of the most impaired cognitive domains in DS, including difficulties with verbal literacy (Van Bysterveldt et al., 2006). This is particularly evident in younger individuals, whose literacy skills have not yet developed, and older individuals, where literacy abilities may have declined over time and are generally weaker due to the lower levels of education in earlier generations (Bochner et al., 2001). Even if they know the alphabet, reading the letters and remember the order may require significant cognitive resources. This added effort could interfere with the TMT's intended assessment of CF rather than reading ability.

Alternatives like the Children's Color Trails Test (Williams et al., 1995), in which participants connect numbers while alternating between colours, were designed to develop easier and culture-fair versions without letters. Nevertheless, effortless reading of Arabic numerals up to fifteen is necessary, which may be challenging for people with DS, given their reduced counting range and interindividual differences in numerical skills (Nye et al., 2001). A study by Van Biesen et al. (2023) using the Color Trails Test in individuals with ID reported low completion rates, suggesting the test may not be well-suited for this population.

More broadly, since individuals with DS have both global cognitive impairments and specific EF difficulties, developing an alternative with reduced demands — incorporating simplified stimuli, lower complexity, and shorter task duration — would be useful.

1.3. The current study

This study introduces the Dice Trails Test (DTT), a novel TMT version, specifically aimed at assessing CF in individuals with DS. Our goal was to create a task with time-based scoring similarly to the TMT, but without letters and numerals. We aimed to evaluate the

DTT's psychometric properties in children and adults with DS, addressing the lack of psychometrically evaluated tasks in DS research (Esbensen et al., 2017). This evaluation included feasibility (proportion of participants completing the task), reliability (split-half, test-retest) and validity.

We hypothesized that the DTT should correlate higher with CF-related measures (convergent validity) compared to CF-unrelated measures (divergent validity) within a DS sample. In a complementary attempt to assess DTT's validity, we also evaluated its developmental sensitivity by analysing differences in age and ID levels within DS. This can be considered a form of criterion validity, with developmental domains (age, ID level) serving as external criteria. Along those lines, we compared people with DS to age-matched TD groups. To be considered a sensitive tool, the DTT should differentiate DS from TD individuals. We hypothesized a substantial difference between DS and TD individuals matched on chronological age (CA). Considering literature indicating a DS-specific CF weakness, we also expected a less pronounced but similar difference between TD and DS individuals when those individuals were matched on MA (i.e., with similar cognitive level). Finally, we correlated DTT performance with the original TMT in TD individuals to further examine DTT's convergent validity.

We adhere to the COSMIN reporting guidelines for studies on measurement properties of patient-reported outcome measures (Gagnier et al., 2021), extending the application of those guidelines to a performance-based measure.

2. Methods

2.1. Participants

Participants were recruited from two larger observational studies involving people with DS and TD controls. We retrospectively selected a subset consisting of all DTT-assessed participants. This sample comprises 96 individuals with DS and 100 TD individuals, consistent with sample sizes in similar studies reporting significant effects (e.g., Pinks et al., 2023). A study flowchart detailing the participants included in each analysis is available in Supplementary Material S1.

2.1.1. DS sample

Eligibility criteria for the DS group were: (a) diagnosis of DS; (b) age between 8 and 14 years (children) and 18–60 years (adults); (c) German as a primary language. Exclusion criteria were: (a) severe uncorrected sensory impairments; (b) acquired severe structural

Table 1
Demographics for the DS samples.

	Children		Adults		
	Full Sample	**Feasible Sample**	**Full Sample**	**Feasible Sample**	**Test-Retest**
N	39	27	57	45	19
Sex (male)[a], *N (%)*	21 (53.8 %)	15 (55.6 %)	30 (52.6 %)	23 (51.1 %)	6 (31.6 %)
ID level, *N (%)*					
Mild	9 (23.1 %)	9 (33.3 %)	40 (70.2 %)	38 (84.4 %)	8 (42.1 %)
Moderate	30 (76.9 %)	18 (66.7 %)	17 (29.8 %)	7 (15.6 %)	11 (57.9 %)
DS type, *N (%)*					
Trisomy 21	36 (92.3 %)	24 (88.9 %)	53 (93.0 %)	42 (93.3 %)	19 (100.0 %)
Mosaic	1	1	1	1	
Translocation	1	1	0	0	
Unknown	1	1	3	2	
Parental Education[b], *N (%)*					
High	29 (74.4 %)	22 (81.1 %)			
Middle	8 (20.5 %)	4 (14.8 %)			
Low	1	0			
Unknown	1	1			
Employment Status, *N (%)*					
Paid community-based job			7 (12.3 %)	6 (13.3 %)	
Facility-based job			41 (71.9 %)	33 (73.3 %)	
Not employed			8 (14.0 %)	6 (13.3 %)	
Others			1	0	
Diagnosis, *N (%)*					
No CD			45 (79.0 %)	37 (82.2 %)	18 (94.7 %)
Dementia			5	1	0
MCI			2	2	0
Secondary CD			4	4	1
Unclassifiable CD			1	1	0
Age years, *M (SD)*	11.21 (1.98)	11.67 (1.73)	33.47 (10.73)	31.89 (9.34)	19.47 (11.57)
Age years, range	8–14	9–14	18–57	18–53	9–43

Note: DS = Down Syndrome; ID = Intellectual Disability; MCI = Mild Cognitive Impairment; CD = Cognitive Decline; Full Sample = Originally included sample; Feasible Sample = Participants for whom the Dice Trails Test was feasible;

[a] assigned at birth

[b] high: university entrance qualification or higher, middle: secondary school diploma or vocational training, low: lower secondary education or below.

brain damages (e.g., stroke, traumatic brain injuries); (c) severe or profound ID. Table 1 presents the characteristics of the DS samples. Adults had a higher proportion of individuals with mild ID (70.2 %) compared to moderate ID, a difference that was significant against the children's sample (mild ID = 23.1 %), χ^2 (1, N = 96) = 18.71, $p < .001$.

2.1.1.1. ID level. The level of ID was classified as mild, moderate, severe or profound according to DSM-V criteria determined by a neuropsychologist based on caregivers' reports of the individuals' adaptive functioning. Current classification systems (DSM-5 and ICD-11) recommend this approach rather than using IQ cut-off scores, as adaptive functioning more accurately reflects the level of support needed in daily life.

2.1.1.2. Dementia diagnosis. For DS adults, a consensus diagnosis by two neurologists of "dementia", "mild cognitive impairment" (MCI), "secondary cognitive decline" (i.e., decline due to other reasons such as psychiatric disorders), or "unclassifiable cognitive decline" was determined based on ICD-10 criteria (Nübling et al., 2022). We included adults regardless of their diagnosis to accurately reflect the older DS population, given the high prevalence of dementia (Lautarescu et al., 2017), and to investigate the applicability of our newly developed test across diagnostic groups.

2.1.1.3. Socio-economic status. We assessed the highest parental education level for children with DS, categorising it as high (university entrance qualification or higher), middle (secondary school diploma or vocational training), or low (lower secondary education or below). For DS adults, we classified their employment status as paid community-based job, paid facility-based job (e.g., sheltered workshops), or not employed. Most attended either special schools or were integrated into mainstream schools with support.

2.1.1.4. Retest subsample. To assess test-retest reliability, we re-administered the DTT to a DS subsample (n = 19, see Table 1) within a 1–24 days interval (M = 4.42 days). We administered the DTT a second time to all individuals with DS who attended two test sessions.

2.1.2. TD sample

We recruited 100 TD participants (47 male) spanning the same age bands as for DS, along with two additional groups (4–6 years, 61–80 years). The mean age of the TD sample was 28.26 years (SD = 23.10). Exclusion criteria were: (a) IQ below 85 and above 120; (b) German not being primary language; (b) reported history of neurological, developmental, or psychiatric disorders (e.g., dyslexia, autism spectrum disorder, depression, dementia). We used the Montreal Cognitive Assessment (MoCA, cut-off ≤ 26; Nasreddine et al., 2005) to screen for dementia in participants aged 50 years and above. Socio-economic level was classified based on the educational level of the participants or of the parents, following the same classification used for the DS sample. The respective distribution was high (65 %), middle (32 %), and low (3 %).

For comparisons between DS and TD, we subdivided the samples as follows:

2.1.2.1. Chronological age-matched (TD-CA). Based on chronological age, we matched 57 TD to 57 DS participants who completed the DTT and were classified as having "no cognitive decline", to mirror our TD sample. The CA-matched groups showed no significant age differences (M_{DS} = 24.58, M_{TD} = 22.53, t(112) = 0.83, p = 0.41). See Table 2 for demographics.

2.1.2.2. Mental age-matched (TD-MA). Based on the Raven's Progressive Matrices 2 raw score (Raven's 2; Raven et al., 2018), we matched 18 TD children aged 4–6 years to 18 DS individuals from our participant pool who completed the DTT and had no signs of cognitive decline. There were no significant differences between the MA-matched groups on Raven's 2 (M_{DS} = 20.28, M_{TD} = 20.61, t (34) = 0.33, p = 0.75). See Table 2 for demographics.

2.2. Procedure

DS adults were recruited from the outpatient clinic at the Department of Neurology at LMU University Hospital in Munich between 2021 and 2024. Other groups, including children with DS and TD individuals, were recruited through targeted advertising and schools between 2021 and 2023. Each participant or their legal proxies provided written consent. The study was conducted in accordance with the Declaration of Helsinki in its latest revision. The study protocol has been approved by the local ethics committee of the LMU medical faculty (Project 17–107 and Project 18–920). The DTT was conducted on-site as part of a larger test battery, with each

Table 2
Demographics for the age-matched groups.

	CA-matched		MA-matched	
	DS	**TD**	**DS**	**TD**
N	57	57	18	18
Sex (male)[a], *N (%)*	28 (49.1 %)	28 (49.1 %)	10 (58.8 %)	11 (64.7 %)
Age years, *M (SD)*	24.58 (14.22)	22.53 (12.00)	25.89 (13.17)	5.83 (1.20)
Age years, range	9–51	9–51	9–51	4–8

Note: DS = Down Syndrome; TD = Typically Developing; CA = Chronological age; MA = Mental age.
[a] assigned at birth

participant individually assessed in one or two sessions. Individuals with DS were assessed by neuropsychologists with long-standing clinical experience with this population. TD individuals were assessed by trained psychology students.

2.3. Cognitive measures

2.3.1. The Dice Trails Test (DTT)

The DTT is a paper-pencil test on A3 format. It is a modified TMT version, incorporating elements of the Children's Color Trails Test. See Table 3 for the DTT's components and their rationale. In short, research indicates that individuals with DS experience difficulties in reading both letters and numerals, despite their ability to count verbally. To mitigate this, we employed dice as stimuli, eliminating symbolic representation while enhancing engagement through their association with gameplay. Additionally, the number of stimuli was reduced to align with the typically lower counting range in DS and to minimize attentional demands.

In DTT-A, participants connect dice sequentially based on the number of dots, with odd numbers in blue and even numbers in yellow. Participants receive verbal instructions and a demonstration using a template of four dice before engaging in a practice trial with the same template. The main test consists of two segments, DTT-A1 and DTT-A2, each with six dice. In DTT-B, each dice, except the first one, exists twice, one in blue and one in yellow. Participants must connect the dice in ascending order, alternating between colours (i.e., one dot blue – two dots yellow – three dots blue… etc.), while disregarding dice with the wrong colour. Verbal instructions and a demonstration with a template of dice with one to four dots are provided (Fig. 1). The same template serves as a practice trial, followed by a second unfamiliar practice template. Both segments of the main test (DTT-B1 and DTT-B2) feature dice ranging from one to six dots.

Scoring is based on completion time in seconds. In case of errors, the clock runs on, and the participant is guided back to the preceding dice. Completion time for DTT-A and DTT-B is determined by averaging their respective segments (DTT-A1 and DTT-A2; DTT-B1 and DTT-B2). DTT's CF index is calculated as the difference in completion time between DTT-A and DTT-B, labelled DTT (B-A). Note that lower DTT(B-A) values signify a higher degree of CF.

2.3.2. Trail Making Test (TMT)

We employed the TMT from the German extended test battery of the Consortium to Establish a Registry for Alzheimer's Disease (CERAD-Plus; Schmid et al., 2014). Only alphabet-proficient participants could be tested, excluding preschoolers and DS individuals. The TMT shows good test-retest reliability ($r = 0.79$ to $r = 0.89$; Dikmen et al., 1999). The order of DTT and TMT was counterbalanced.

2.3.3. Category switching

Verbal CF was assessed with a modified Category Switching task from the Regensburger Wortflüssigkeits-Test, demonstrating high test-retest reliability ($r = 0.72$ - $r = 0.89$, Aschenbrenner et al., 2000). The test requires participants to generate words while switching between categories (here animals and food). The outcome is the number of correct responses within one minute.

2.3.4. Raven's 2

The Raven's 2 was assessed to evaluate nonverbal abstract reasoning, representing a higher EF (Diamond, 2013). The test demonstrates high reliability ($r = 0.73$–0.93, Raven et al., 2018). Set A-C raw scores were used for group comparisons between DS and TD. To assess the individuals' IQ, following the manual guidelines, set A-C was administered to controls aged 4;0–8;11 and to people with DS, while set B-E was used for controls aged ≥ 9 years.

2.3.5. Boston naming test, short form (BNT)

The BNT from the CERAD-Plus (Schmid et al., 2014) assesses naming ability through black-and-white illustrations of 15 objects.

Table 3
Rationale for DTT components for individuals with DS.

DTT Components	Rationale
Nonverbal response	Difficulties in expressive language (Grieco et al., 2015).
No letters as stimuli	Great variability in literacy skills (Ratz, 2013). Reducing cultural bias
Dice as stimuli, no Arabic numerals	Symbolic numerical representations (i.e., numerals) develop later in life than non-symbolic ones (Li et al., 2018). Given the delayed development in DS (Grieco et al., 2015), non-symbolic stimuli can increase applicability in DS children. Spatially structuring dots into patterns improves numerosity estimation (Anobile et al., 2020); DS individuals exhibit relative strengths in visuospatial skills (Grieco et al., 2015). Familiarity with dice patterns through gameplay; enhancing motivation
Reduced number of numeral stimuli	Lower counting range in DS compared to MA-matched controls (Up to six objects, Nye et al., 2001). Reducing attentional demands
Colours blue and yellow, A3 format	Larger stimuli with high contrast colours to account for reduced contrast sensitivity (Rocco et al., 1997).
Processing speed as outcome measure	Higher sensitivity to detect small effects by providing a wide range of potential scores compared to measures with a fixed score range (e.g. 0–6 points) Continuous outcome to prevent high floor and ceiling effects

Note: DTT = Dice Trails Test; DS = Down Syndrome.

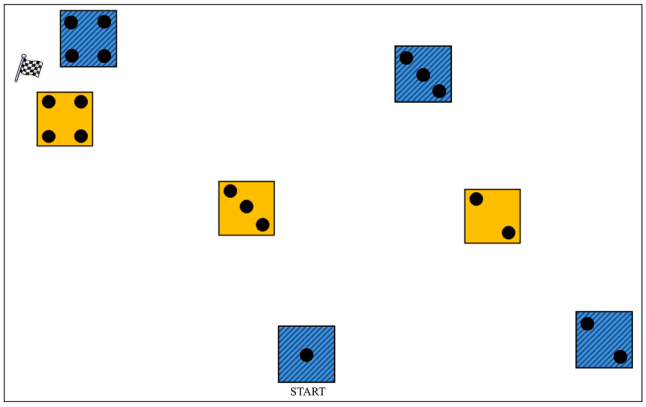

Fig. 1. Practice trial for dice trails test (DTT) Part B. Stripes on blue dice were added here to enhance visual distinguishability.

The number of objects correctly named is calculated. Cronbach's alpha of BNT short forms ranges between 0.37 and 0.67 (Fastenau et al., 1998).

2.3.6. *Drawing & copying (DC)*

This subscale of the Cambridge Cognitive Examination for older Adults with Down's Syndrome (CAMCOG-DS, Nübling et al., 2020) assesses constructional praxis and visuospatial abilities. It involves copying figures and drawing a clock, with a maximum score of 8. Retest-reliability for the items ranges from $\tau = 0.53$ to $\tau = 0.69$ in DS (Loosli et al., 2024).

2.4. *Statistical analysis*

Statistical analyses were conducted in R (version 4.2.0; R Core Team, 2022). Significance was set at $p < 0.05$. Missing data were handled via pairwise deletion, removing cases only from analyses where data were missing while retaining them for others, maximizing available data use. We defined extreme values as values beyond 3 SD from the mean. However, we chose not to exclude participants due to this criterion, given the broad variability in cognitive abilities within DS (Ratz, 2013). Considering the higher proportion of DS adults with mild ID compared to children (see Section 2.1.1.), we conducted main analyses additionally with ID level adjustment. Due to small sample sizes ($n < 10$) when stratifying by age group (children vs. adults) and ID level (mild vs. moderate), these findings should be treated as preliminary.

2.4.1. *Feasibility*

Feasibility was defined as the percentage of participants completing DTT-A and DTT-B in each group (DS and TD) separately. Good feasibility is assumed when 80 % or more DS individuals manage to complete the test (see Schworer et al., 2023). Statistical analyses were calculated with task completion as a binary variable (completed/not-completed).

2.4.2. *Distributional properties*

DTT completion time served as the main outcome measure. Therefore, classical definitions of "floor" and "ceiling" effects (clustering near minimum or maximum scores) were not applicable due to the absence of a fixed scoring range. Instead, we evaluated the DTT's skewness: Negative skew implied a floor effect, positive skew a ceiling effect. Values within $\pm$ 2 were considered acceptable.

2.4.3. *Reliability*

We evaluated split-half reliability with Spearman-Brown adjusted Pearson correlations. Test-retest reliability was calculated using intraclass correlations (ICC). ICCs between 0.5 and 0.75 indicate moderate test-retest reliability, and values $\geq$ 0.75 good test-retest reliability (Koo & Li, 2016).

2.4.4. *Validity*

In DS, DTT's convergent validity was assessed via age-adjusted Pearson correlations with another CF measure (Category Switching) and a higher-level EF measure (abstract reasoning; Raven's 2), with good validity defined as $r \geq 0.50$ (Schworer et al., 2023). Divergent validity was evaluated by assessing the correlations of DTT with less CF-related measures (BNT and DC). In TD, DTT's convergent validity was examined by its correlation with the original TMT.

Within-group differences in DS were analysed with two-sample t-tests and linear regression. To ensure that observed age effects in DTT performance could be solely attributed to non-pathological, i.e. purely age-related, changes in the DS sample, participants diagnosed with dementia, MCI, and secondary/unclassifiable forms of cognitive decline were excluded from the regression analyses. The data of those excluded participants were not further analysed as there was not enough data to analyse them separately. Between-group differences between DS and TD were examined using a mixed-design analysis of variance (mixed-design ANOVA).

Some variables were not normally distributed (Shapiro-Wilk test). Since parametric tests can be sensitive to non-normality and since in non-parametric tests valuable information is lost when continuous values are converted into ranks, we used bootstrapping methods (5.000 resamples) alongside parametric tests to confirm main results. Parameters and confidence intervals obtained from bootstrapping are provided in the Supplementary Material S2. Unless indicated otherwise, bootstrapping confirms the parametric analyses.

3. Results

3.1. Feasibility and distributional properties

Table 4 shows DTT descriptive statistics and feasibility scores. In TD, both DTT-A and DTT-B exceeded the ≥ 80 % cut-off for good feasibility, with 99 % and 94 %, respectively. In DS, while DTT-A showed high feasibility for children (92 %) and adults (88 %), DTT-B fell below that cut-off, with 69 % of children and 79 % of adults completing it. Among adults, only one out of five individuals with dementia completed DTT-B. Without this subgroup, feasibility among DS adults increased to 85 %.

Considering the higher proportion of mild ID in the DS adult sample, we further investigated the relationship between feasibility and ID level. Fisher's exact test revealed a higher completion rate for DTT-B among DS adults with mild ID (95 %) versus moderate ID (42 %, $p < .001$). Similarly, DS children with mild ID achieved 100 % completion, while those with moderate ID showed only 60 % completion rate ($p = .036$).

A ceiling effect for DTT was observed within the TD sample, characterised by high (> 2) positive kurtosis and skewness. In DS, DTT exhibited minimal distributional concerns, with elevated values observed only in children for DTT-A (Table 4). Low skewness indicates no relevant floor effect for DTT in the DS sample.

3.2. Reliability

Split-half reliability between DTT-A1 and DTT-A2 was high for DS (0.85) and TD (0.93, Spearman-Brown coefficient). Similarly, for DTT-B1 and DTT-B2, split half reliability was 0.90 in DS and 0.79 in TD, indicating good internal consistency.

Age comparison between the whole DS sample and the test-retest subsample ($M = 19.47$) yielded no significant difference, $t(18) = -1.15$, $p = .26$. The ICC(3,2) values were 0.76 for DTT-A, 0.90 for DTT-B and 0.88 for DTT-(B-A), providing preliminary evidence of strong test-retest reliability.

3.3. Construct validity

Table 5 shows age-adjusted correlations between DTT and cognitive tasks in the DS samples. In adults, DTT measures correlated moderately with CF-related tasks, namely Category Switching and Raven's 2 (all $p < 0.05$), but not with CF-unrelated tasks (BNT and DC; all $r < .29$, n.s). Only DTT-B and Category Switching met the predefined cut-off for good convergent validity ($r \geq 0.50$). For children, a significant moderate relationship exceeding the cut-off ($r = 0.51$) was observed only between Category Switching and the main outcome DTT(B-A). The Supplementary Material S3 provides correlations adjusted for ID level. This adjustment did not yield relevant changes in the DTT and Category Switching relationship but a slight decrease in the correlation between DTT and Raven's 2.

Among alphabet-proficient TD participants ($n = 77$), we exhibited a significant moderate relationship between DTT(B-A) and TMT (B-A) after age adjustment ($r = 0.34$, $p < 0.001$). However, the ceiling effect observed for DTT in TD participants likely led to compressed variability, thereby probably lowering the correlation. Focusing on older participants within this group between 50 and 80 ($M = 64.09$, $SD = 8.93$, 36.4 % male), the positive skew in DTT measures disappeared (skewness < 2), indicating a reduced ceiling effect. An exploratory analysis of this older group ($n = 22$) showed a strong correlation between DTT(B-A) and TMT(B-A), $r = 0.63$, $p < 0.001$.

Table 4
Descriptive statistics for DTT.

	Feasibility *n (%)*	Min	Max	*M (SD)*	Skew	Kurtosis
Overall TD						
DTT-A	99 (99.0 %)	1.30	28.08	4.68 (3.70)	3.64	17.13
DTT-B	94 (94.0 %)	3.34	48.47	10.68 (8.41)	2.60	6.75
DTT(B-A)		0.64	32.51	6.00 (5.47)	2.48	6.70
Children DS						
DTT-A	36 (92.3 %)	4.86	47.96	12.84 (8.05)	3.07	10.65
DTT-B	27 (69.2 %)	11.44	69.64	31.08 (13.90)	0.80	0.07
DTT(B-A)		1.95	37.75	18.24 (10.29)	0.27	−1.26
Adults DS						
DTT-A	50 (87.7 %)	3.59	56.00	18.23 (12.98)	1.38	1.20
DTT-B	45 (79.0 %)	8.73	112.49	35.46 (23.60)	1.40	1.35
DTT(B-A)		0.65	56.49	17.23 (12.55)	1.27	1.27

Note: DTT = Dice Trails Test, Outcome: completion time in seconds; DS = Down Syndrome; TD = Typically Developing; Skew = Skewness; Kurtosis = Excess Kurtosis.

Table 5
Construct validity of the DTT in DS, partial correlations controlling for age.

	Convergent Validity		Divergent Validity	
	Category Switching	Raven's 2	BNT	DC
Children				
	($n = 25$)	($n = 27$)	($n = 27$)	($n = 27$)
DTT-A	0.09	−0.01	−0.25	−0.15
DTT-B	−0.31	−0.09	−0.29	−0.19
DTT(B-A)	−0.51**	−0.10	−0.16	−0.12
Adults				
	($n = 43$)	($n = 36$)	($n = 44$)	($n = 44$)
DTT-A	−0.41**	−0.44**	−0.18	−0.28
DTT-B	−0.51**	−0.44**	−0.18	−0.20
DTT(B-A)	−0.49**	−0.36*	−0.14	−0.08

Note: DTT = Dice Trails Test, Outcome: completion time in seconds; DS = Down Syndrome; BNT = Boston Naming Test; Raven's 2 = Raven's Progressive Matrices 2, Raw score set A-C; DC = Drawing & Copying from the Cambridge Cognitive Examination for older Adults with Down's Syndrome (CAMCOG-DS).
* $p < 0.05$
** $p < 0.01$.

3.4. Criterion validity

3.4.1. Between-group differences

A 2 (group) x 2 (DTT part) mixed-design ANOVA comparing the CA-matched TD and DS group on completion time for DTT-A and DTT-B revealed a significant main effect for group ($F(1,112) = 99.98, p < .001, \eta_p^2 = .47$) and DTT part ($F(1,112) = 206.19, p < .001, \eta_p^2 = .65$). Completion times were longer for DTT-B than for DTT-A, and TD individuals completed both DTT parts faster. Importantly, a significant interaction effect emerged ($F(1,112) = 77.45, p < .001, \eta_p^2 = .41$), showing that the increase in completion time in DTT-B compared to DTT-A was more pronounced in DS compared to the CA-matched controls (Fig. 2A), hinting at stronger CF skills in TD CA-matched individuals compared to DS.

For MA-matched groups, a similar pattern could be observed, but the relevant ANOVA yielded only a significant main effect for the factor DTT part ($F(1,34) = 60.43, p < .001, \eta_p^2 = .64$, see Fig. 2B). No significant main effect for group and no interaction effect emerged (all $p > 0.05$), suggesting no difference in CF skills between DS and MA-matched controls.

3.4.2. Within-group differences in DS

No significant differences occurred between ID levels in DTT-A and DTT-B (all $p > 0.07$). Nevertheless, individuals with DS and mild ID ($M = 13.85$) outperformed those with moderate ID ($M = 20.96$) on DTT(B-A), $t = -2.79, p < 0.01$. No significant sex differences were found (all $p > 0.20$).

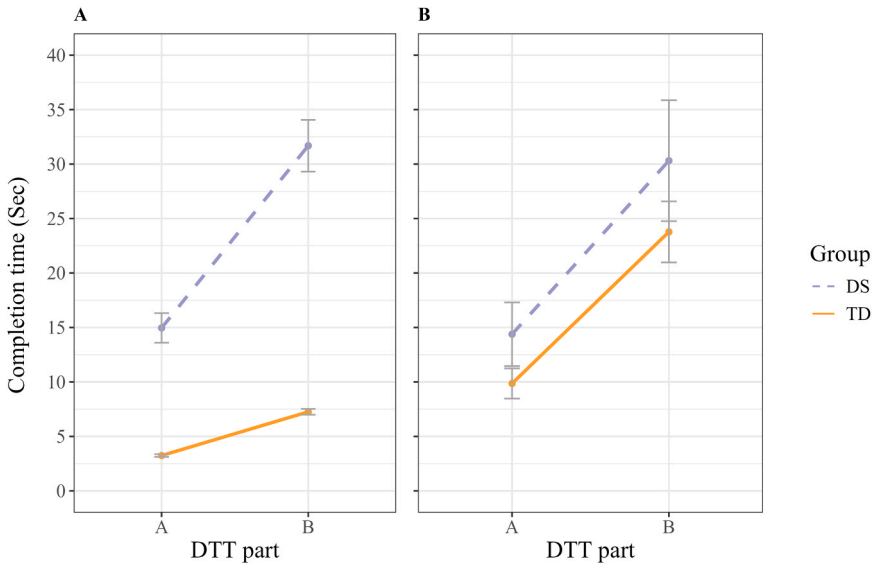

Fig. 2. Mean Completion Time on the Dice Trails Test (DTT) Split by Group and Part. (A) Groups Matched on Chronological Age (CA, $n = 114$) and (B) Mental Age (MA, $n = 36$). Error Bars Represent Standard Errors. DS = Down Syndrome; TD = Typically Developing; Sec = Seconds.

Age was found to be a significant predictor for DTT(B-A) in DS children ($\beta = -3.18$, $p < 0.0{,}1$ $R^2 = 0.29$) and adults ($\beta = 0.48$, $p = 0.01$, $R^2 = 0.17$). Notably, age was associated with better task performance in children, indicated by a decrease in DTT(B-A) with increasing age. Conversely, the reverse trend was evident in DS adults (Fig. 3).

Given the observed difference between DS individuals with mild and moderate ID in DTT(B-A), we conducted the same analysis including ID level as a potential moderator. For adults, age remained a significant predictor ($\beta = 0.48$, $p = 0.01$), with no significant ID level ($p = 0.82$) or interaction effects ($p = 0.67$). For children, the effect of age disappeared ($p = 0.40$). Instead, a significant main effect of ID level ($\beta = 60.04$, $p = 0.02$) and interaction effect ($\beta = -4.55$, $p = 0.04$) emerged, hinting at a moderating role of ID level. DS children with moderate ID seem to exhibit a pronounced improvement in DTT performance with increasing age compared to those with mild ID. This model explained more variance than age alone ($\Delta R^2 = 0.18$, $F(2,22) = 4.01$, $p = 0.03$). However, due to small sample sizes when stratifying by age and ID level, this result should be considered preliminary.

4. Discussion

Research indicates that CF impacts various life outcomes in DS individuals across developmental stages (Amadó et al., 2016; Lautarescu et al., 2017), highlighting its potential as a target for interventions and clinical trials. Yet, properly evaluated neuropsychological tasks for this population are lacking. When evaluated, they frequently demonstrated floor effects, with many individuals scoring at the low end, rendering them insensitive to differentiate between lower performance levels (Esbensen et al., 2017). Hence, we developed a modified TMT specifically aimed at assessing CF in DS – the DTT. To the best of our knowledge, despite numerous TMT versions (e.g., Chan & Morgan, 2018; Williams et al., 1995), this is the first adaptation for people with DS. Rather than simply making the TMT easier, we tailored it to meet the unique needs of individuals with DS. By eliminating the use of numerals and letters, the DTT is designed to be more accessible for individuals with limited literacy skills and counting abilities. Despite low feasibility for some DS subgroups, the DTT largely met predefined psychometric criteria, demonstrated developmental sensitivity and revealed clear differences in performance between DS and TD individuals.

4.1. Feasibility and distributional properties

The DTT showed completion rates ranging between 69 % in children and 85 % in adults without a dementia diagnosis, placing it within or above the range of other common CF measures in DS (3–84 %; Schworer et al., 2023). The initial age group differences appeared to result from a higher proportion of people with DS and mild ID among adults compared to a higher proportion of moderate ID in children. Indeed, over 95 % of DS individuals with mild ID across age groups completed DTT-B, exceeding the predefined 80 % threshold, while those with moderate ID fell significantly short, clearly indicating DTT's better applicability for people with DS and mild ID. To our knowledge, no prior studies have examined EF task feasibility differences by ID level, making direct comparisons challenging. However, Schworer et al. (2023) reported lower completion rates among DS individuals with lower IQ, indicating a similar pattern.

While acknowledging this limitation, it is important to highlight that no relevant floor effects were found for DS participants who completed the DTT. This suggests it is sufficiently sensitive to CF differences in people with DS and confirms our expectation that continuous outcome measures reduce floor effects, though ceiling effects were observed in healthy TD individuals. Our result is

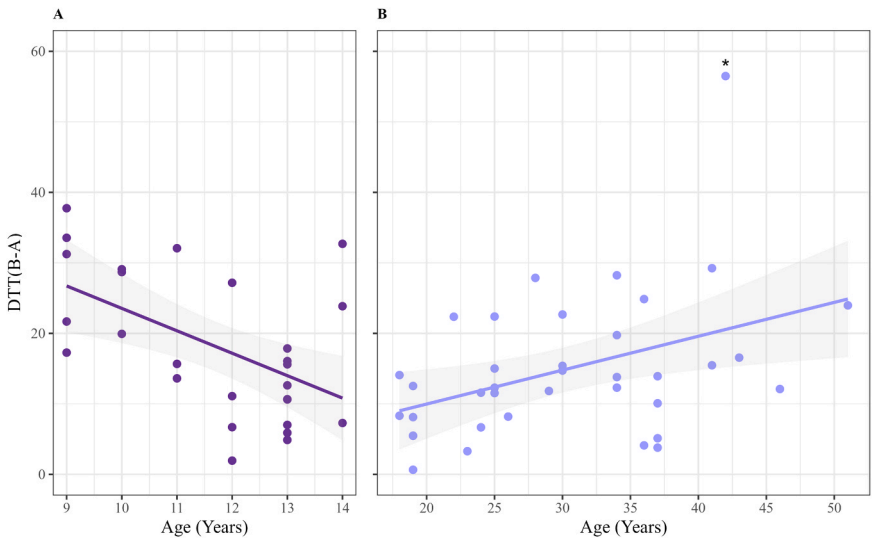

Fig. 3. Linear relationship between age and performance on the dice trails test (DTT, difference in completion time in seconds between Part A and B), for A) Children ($n = 27$) and B) Adults ($n = 37$) with down syndrome (DS) and no diagnosis of cognitive decline. Linear regression results remained significant without the extreme value (*).

noteworthy given that Tungate and Conners (2021) found in their meta-analysis that over 70 % of studies investigating CF show floor-related skewness in DS. Moreover, we applied a rather stringent definition of "feasibility", requiring participants to complete DTT-A1, DTT-A2, DTT-B1 and DTT-B2. By comparison, Van Deusen et al. (2023) determined the feasibility of their reverse categorization task based on whether participants gave one or more correct responses during the pre-switch phase (equivalent to DTT-A), without taking the post-switch phase into account, where over half of the participants did not reach this criterion. In summary, although feasibility was significantly lower in people with DS and moderate ID, high feasibility for people with mild ID applied across a broad age range, from young children to adults over 50. Thus, the DTT could be a suitable choice for future studies investigating effects across a wide age range, such as developmental studies.

However, improving the completion rate for individuals with moderate ID would be beneficial, as they represent a substantial portion of the DS population. While we did not assess reasons for non-completion, high feasibility in DTT-A compared to DTT-B implies that, rather than participant engagement or the task of connecting stimuli itself, the instruction complexity may be the primary issue. This could explain higher completion rates in mild ID, given the link between cognitive level and language development in DS (Grieco et al., 2015). Simplifying instructions with a more step-by-step approach during practice (e.g., starting with colour switching, then adding numbers) might help. However, due to the complexity of EF (Diamond et al., 2013), CF tasks will likely always pose feasibility challenges for a proportion of people with DS.

4.2. Reliability

The DTT demonstrated good internal consistency. Additionally, this study provided promising evidence for good test-retest reliability, though preliminary due to a small retest sample and a short interval averaging less than one week. Shorter intervals can overestimate the reliability (Polit, 2014), so our result should be interpreted with caution. Nonetheless, few studies have examined retest reliability in this population, often with modest results (e.g., Pinks et al., 2023; Schworer et al., 2023).

4.3. Construct validity

The correlation patterns provide promising evidence for the DTT's construct validity, with the strongest association between DTT and Category Switching, and no significant correlation with CF-unrelated tasks in DS. Additionally, we found a strong relationship between TMT and DTT in older TD individuals who did not exhibit ceiling effects. Considering the general decline in EF skills with increasing age (Zelazo et al., 2004), this finding supports the DTT's suitability for people with lower cognitive capacities.

Due to the complexity of EF tasks, they typically engage not only the targeted EF but also additional cognitive processes such as attention (*task impurity problem*; Miyake et al., 2000). Consequently, correlations between EF tasks tend to be weaker due to the interference from these additional processes, as seen for other TMT versions when correlating with CF tasks (Chan & Morgan, 2018). Considering this, our results, with correlations around 0.5 between DTT and both Category Switching and TMT, are encouraging, indicating they tap into the same construct, namely CF. Unexpectedly, our results showed no relationship between DTT and Raven's 2 among DS children. This difference may be due to other cognitive processes, like attention, having a greater impact on children solving such task.

DTT-B contains twice as many stimuli as DTT-A, increasing visual scanning demands. Previous research demonstrates that visual complexity can impact TMT results (Gaudino et al., 1995). Adding distractor stimuli to the DTT-A version could mitigate this risk. Nevertheless, our study's correlation between DTT and EF tasks clearly indicates EF involvement.

Overall, the DTT seems to be a valuable alternative to assess CF in DS, given the relatively low convergent validity of conventional tasks in this population (Schworer et al., 2023).

4.4. Criterion validity

4.4.1. Between-group differences

As expected DTT-B's completion time exceeded that of DTT-A due to its higher demands on EF. This increase was pronounced in DS compared to CA-matched individuals, reflecting lower CF in DS compared to TD peers of similar age. Surprisingly, compared to an MA-matched group (i.e., with comparable cognitive level), the difference disappeared, challenging the prevailing view of a DS-specific CF weakness. In line with this finding, Tungate and Conners (2021) found that performance differences between DS and MA-matched controls were smaller for non-verbal compared to verbal tasks. Specifically, there was no difference in non-verbal shifting tasks. Additionally, many studies that identified a CF deficit employed rating scales (Csumitta et al., 2022), which can yield different results than performance-based measures (Toplak et al., 2013). The heterogeneous findings highlight the need for additional research, alongside performance-based tasks specifically designed for individuals with DS.

Nonetheless, a similar performance pattern for DS and MA-matched controls, in contrast to the differences between DS and CA-matched controls, offers evidence that the DTT outcome varies with the cognitive level — a crucial characteristic when developing a neuropsychological task.

4.4.2. Within-group differences in DS

The DTT showed developmental sensitivity in DS, in terms of both ID level and chronological age. People with mild ID outperformed those with moderate ID on DTT's flexibility score. Moreover, we found differences in DTT performance within DS depending on age, a crucial consideration in task development (Edgin et al., 2010): As age increased, children exhibited a decrease in

DTT(B-A), reflecting increased CF, whereas adults displayed an opposite trend. Interestingly, our results hint at a moderating role of ID level in DS children. Despite lower DTT performance among DS children with moderate ID, they showed greater improvement in CF with age, potentially suggesting a developmental delay in EF compared to those with mild ID.

We focused on DTT's developmental sensitivity for healthy age-related changes due to small sample sizes within diagnostic groups, representing a crucial initial step. Evaluating the DTT's ability to distinguish healthy age-related from disease-related changes, for example to diagnose MCI, would be of interest for future clinical trials involving anti-amyloid targeting medication (Strydom et al., 2018).

4.5. Limitations and future research

Although we assessed a substantial sample of 96 DS individuals, conclusions were limited by small group sizes when stratifying by age, ID level and diagnostic group. This was particularly evident in the children subsample, which restricts the generalizability of our findings. Further research is warranted to validate our findings with larger groups, especially the differences between ID levels.

Additionally, our feasibility analysis focused exclusively on DS individuals with mild and moderate ID, omitting those with severe ID, despite their significant representation within the population. Given that the DTT only achieved acceptable feasibility for individuals with mild ID but not for those with moderate ID, it suggests that the task may not be suitable for individuals with severe ID. Future research could benefit from exploring adaptations of the DTT or developing alternative tasks that are better suited for this subgroup. Another potential approach could involve using only DTT-A, which may be more accessible for individuals with greater support needs. It is essential to recognize that individuals with severe and profound ID are frequently overlooked in research, and this gap needs to be addressed in future studies.

To further validate our findings, future studies should support DTT's construct validity by including a nonverbal CF task, like a card sorting task, to correlate it with a test in the same modality, as this can impact the results (Tungate & Conners, 2021). Moreover, while the original TMT and its existing adaptations may not be well suited for individuals with DS, future research could still benefit from comparing their feasibility with that of the DTT.

The DTT demonstrates potential for broader application beyond individuals with DS. Its design, free from letters and numerals, positions it as a culture-free TMT version. The DTT could be a promising alternative for pre-school TD children, for individuals with severe cognitive impairments (e.g., dementia in the later stages), and those with lower educational levels, particularly those struggling with the traditional TMT.

5. Conclusion

Currently, no CF measure has been specifically designed and rigorously evaluated for use across both children and adults with DS. The DTT fills this critical gap by demonstrating adequate validity, sensitivity for between-group and within-group differences, and promising preliminary test-retest reliability. Notably, unlike many existing cognitive measures, the DTT does not exhibit floor effects for individuals with DS — a major limitation of traditional assessments. While feasibility analysis suggests that the DTT is most suitable for individuals with DS and mild ID rather than moderate ID, it has proven effective across a broad age range from 9 to 53 years. This makes the DTT a promising tool for developmental studies, where different tools are typically used across different age groups. Furthermore, the demonstrated developmental sensitivity of the DTT provides strong evidence of its potential as a reliable outcome measure for future clinical trials.

Funding

This work was supported by the Else Kröner-Fresenius Stiftung (Grant number 2020_EKEA.09), the Fondation Jérôme LeJeune, the VERUM Foundation, the Deutsche Forschungsgemeinschaft DFG, German Research Foundation, under Germany's Excellence Strategy within the framework of the Munich Cluster for Systems Neurology (Grant number EXC 2145 SyNergy – ID 390857198), and the MOMENTE program for excellent young scientists of the medical faculty of the LMU.

CRediT authorship contribution statement

Müller Armelle: Writing – review & editing, Investigation, Data curation. **Heiß Lena:** Writing – review & editing, Investigation. **Stadler Hannah:** Writing – review & editing, Investigation. **Sandkühler Katja:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Wlasich Elisabeth:** Writing – review & editing, Investigation. **Levin Johannes:** Writing – review & editing, Funding acquisition, Conceptualization. **Loosli Sandra V.:** Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Schenk Thomas:** Writing – review & editing. **Danek Adrian:** Writing – review & editing. **Wagemann Olivia:** Writing – review & editing, Investigation, Data curation. **Nübling Georg:** Writing – review & editing, Data curation.

Declaration of Competing Interest

Johannes Levin reports speaker fees from Bayer Vital, Biogen, EISAI, TEVA, Zambon, Esteve, Merck and Roche, consulting fees from Axon Neuroscience, EISAI and Biogen, author fees from Thieme medical publishers and W. Kohlhammer GmbH medical publishers and

is inventor in a patent "Oral Phenylbutyrate for Treatment of Human 4-Repeat Tauopathies" (PCT/EP2024/053388) filed by LMU Munich. Additionally, he reports compensation for serving as chief medical officer for MODAG GmbH, is beneficiary of the phantom share program of MODAG GmbH and is inventor in a patent "Pharmaceutical Composition and Methods of Use" (EP 22 159 408.8) filed by MODAG GmbH, all activities outside the submitted work.

Acknowledgements

We thank Nina Smrzka for study coordination, the participants and caregivers for participation.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ridd.2025.104965.

Data Availability

The authors do not have permission to share data.

References

Amadó, A., Serrat, E., & Vallès-Majoral, E. (2016). The role of executive functions in social cognition among children with down syndrome: Relationship patterns. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.01363

Anobile, G., Castaldi, E., Moscoso, P. a M., Burr, D. C., & Arrighi, R. (2020). Groupitizing": a strategy for numerosity estimation. *Scientific Reports, 10*(1). https://doi.org/10.1038/s41598-020-68111-1

Aschenbrenner, S., Tucha, O., & Lange, K. W. (2000). *Regensburger wortflüssigkeits-test:Rwt*. Hogrefe: Verlag für Psychologie.

Bochner, S., Outhred, L., & Pieterse, M. (2001). A study of functional literacy skills in young adults with down syndrome. *International Journal of Disability, Development and Education, 48*(1), 67–90. https://doi.org/10.1080/10349120120036314

Bowie, C. R., & Harvey, P. D. (2006). Administration and interpretation of the trail making test. *Nature Protocols, 1*(5), 2277–2281. https://doi.org/10.1038/nprot.2006.390

Carney, D. P., Brown, J. H., & Henry, L. A. (2013). Executive function in Williams and Down syndromes. *Research in Developmental Disabilities, 34*(1), 46–55. https://doi.org/10.1016/j.ridd.2012.07.013

Chan, A. Y. C., & Morgan, S. (2018). Assessing children's cognitive flexibility with the Shape Trail Test. *PLoS ONE, 13*(5), Article e0198254. https://doi.org/10.1371/journal.pone.0198254

Csumitta, K. D., Stephan, C. M., LaQuaglia, R. I., Miller, E., & Lee, N. R. (2022). Updated profiles of everyday executive function in youth with Down syndrome using the BRIEF-2. *Journal of Intellectual Disability Research, 66*(1–2), 68–80. https://doi.org/10.1111/jir.12879

Daunhauer, L. A., Fidler, D. J., Hahn, L., Will, E., Lee, N. R., & Hepburn, S. (2014). Profiles of everyday executive functioning in young children with down syndrome. *American Journal on Intellectual and Developmental Disabilities, 119*(4), 303–318. https://doi.org/10.1352/1944-7558-119.4.303

De Graaf, G., Buckley, F., Dever, J., & Skotko, B. G. (2017). Estimation of live birth and population prevalence of down syndrome in nine U.S. states. *American Journal Of Medical Genetics Part A, 173*(10), 2710–2719. https://doi.org/10.1002/ajmg.a.38402

De Sola, S., De La Torre, R., Sánchez-Benavides, G., Benejam, B., Cuenca-Royo, A., Del Hoyo, L., Rodríguez, J., Catuara-Solarz, S., Sanchez-Gutierrez, J., Dueñas-Espin, I., Hernandez, G., Peña-Casanova, J., Langohr, K., Videla, S., Blehaut, H., Farre, M., Dierssen, M., Principe, A., Civit, E., & Legout, V. (2015). A new cognitive evaluation battery for down syndrome and its relevance for clinical trials. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.00708

Diamond, A. (2013). Executive functions. *Annual Review of Psychology, 64*(1), 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Dikmen, S. S., Heaton, R. K., Grant, I., & Temkin, N. R. (1999). Test–retest reliability and practice effects of Expanded Halstead–Reitan Neuropsychological Test Battery. *Journal of the International Neuropsychological Society, 5*(4), 346–356. https://doi.org/10.1017/s1355617799544056

Edgin, J. O., Mason, G. M., Allman, M. J., Capone, G. T., Deleon, I., Maslen, C., Reeves, R. H., Sherman, S. L., & 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

Esbensen, A. J., Hooper, S. R., Fidler, D., Hartley, S. L., Edgin, J., D'Ardhuy, X. L., Capone, G., Conners, F. A., Mervis, C. B., Abbeduto, L., Rafii, M., Krinsky-McHale, S. J., Urv, T., & Group, O. M. W. (2017). Outcome measures for clinical trials in Down syndrome. *American Journal on Intellectual and Developmental Disabilities, 122*(3), 247–281. https://doi.org/10.1352/1944-7558-122.3.247

Fastenau, P. S., Denburg, N. L., & Mauer, B. A. (1998). Parallel short forms for the boston naming test: psychometric properties and norms for Older adults. *Journal of Clinical and Experimental Neuropsychology, 20*(6), 828–834. https://doi.org/10.1076/jcen.20.6.828.1105

Gagnier, J. J., Lai, J., Mokkink, L. B., & Terwee, C. B. (2021). COSMIN reporting guideline for studies on measurement properties of patient-reported outcome measures. *Quality of Life Research, 30*(8), 2197–2218. https://doi.org/10.1007/s11136-021-02822-4

Gaudino, E. A., Geisler, M. W., & Squires, N. K. (1995). Construct validity in the trail making test: What makes part B harder? *Journal of Clinical and Experimental Neuropsychology, 17*(4), 529–535. https://doi.org/10.1080/01688639508405143

Grieco, J. A., Pulsifer, M. B., Seligsohn, K., Skotko, B. G., & Schwartz, A. (2015). Down syndrome: Cognitive and behavioral functioning across the lifespan. *American Journal of Medical Genetics Part C Seminars in Medical Genetics, 169*(2), 135–149. https://doi.org/10.1002/ajmg.c.31439

Koo, T. K., & Li, M. Y. (2016). A Guideline of selecting and Reporting Intraclass correlation coefficients for Reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lautarescu, B. A., Holland, A. J., & Zaman, S. H. (2017). The early presentation of dementia in people with down syndrome: A systematic review of longitudinal studies. *Neuropsychology Review, 27*(1), 31–45. https://doi.org/10.1007/s11065-017-9341-9

Li, Y., Zhang, M., Chen, Y., Deng, Z., Zhu, X., & Yan, S. (2018). Children's non-symbolic and symbolic numerical representations and their associations with mathematical ability. *Frontiers in Psychology, 9*. https://doi.org/10.3389/fpsyg.2018.01035

Loosli, S. V., Neumann, L. C., Wlasich, E., Prix, C., Koll, L., Weidinger, E., Vöglein, J., Wagemann, O., Danek, A., Nübling, G., & Levin, J. (2024). Measurement properties of the German version of the Cambridge examination for mental disorders of older people with Down syndrome and others with intellectual disabilities (CAMDEX-DS). *Journal of Intellectual Developmental Disability*, 1–13. https://doi.org/10.3109/13668250.2024.2317794

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "Frontal lobe" tasks: a latent variable analysis. *Cognitive Psychology, 41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The montreal cognitive assessment, MOCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society, 53*(4), 695–699. https://doi.org/10.1111/j.1532-5415.2005.53221.x

Nübling, G., Loosli, S. V., Wlasich, E., Prix, C., Schönecker, S., Freudelsperger, L., Smrzka, N., Strydom, A. M., Zaman, S. H., Benejam, B., Missios, J., Meister, R., Danek, A., & Levin, J. (2020). Eine deutsche fassung der cambridge examination for mental disorders of older people with down's syndrome and others with intellectual disabilities. *Zeitschrift Für Gerontologie Und Geriatrie, 53*(6), 546–551. https://doi.org/10.1007/s00391-019-01591-7

Nübling, G., Wagemann, O., Loosli, S. V., Wlasich, E., Danek, A., & Levin, J. (2022). Diagnostik der alzheimer-demenz bei menschen mit einem down-syndrom nach ICD, DSM-V und A/T/N-Klassifikation. *Fortschritte Der Neurologie · Psychiatrie, 90*(09), 416–420. https://doi.org/10.1055/a-1771-6273

Nye, J., Fluck, M., & Buckley, S. (2001). Counting and cardinal understanding in children with Down syndrome and typically developing children. *Down Syndrome Research and Practice, 7*(2), 68–78. https://doi.org/10.3104/reports.116

Pinks, M. E., Van Deusen, K., Prince, M. A., Esbensen, A. J., Thurman, A. J., Patel, L. R., Abbeduto, L., Walsh, M. M., Daunhauer, L. A., Feigles, R. T., Nguyen, V., & Fidler, D. J. (2023). Psychometric evaluation of a working memory assessment measure in young children with Down syndrome. *Research in Developmental Disabilities, 139*, Article 104564. https://doi.org/10.1016/j.ridd.2023.104564

Polit, D. F. (2014). Getting serious about test–retest reliability: a critique of retest research and some recommendations. *Quality of Life Research, 23*(6), 1713–1720. https://doi.org/10.1007/s11136-014-0632-9

R Core Team. (2022). R (4.2.0): A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ⟨https://www.R-project.org/⟩.

Ratz, C. (2013). Do students with Down syndrome have a specific learning profile for reading? *Research in Developmental Disabilities, 34*(12), 4504–4514. https://doi.org/10.1016/j.ridd.2013.09.031

Raven, J., Rust, J., Chan, F., & Zhou, X. (2018). *Raven's 2 progressive matrices, clinical edition (Raven 2)*. Frankfurt, Germany: Pearson.

Reitan, R., & Wolfson, D. (1993). *The Halstead-Reitan neuropsychology battery: Theory and clinical interpretation*. Tucson, AZ: Neuropsychology Press.

Rocco, F. J., Cronin-Golomb, A., & Lai, F. (1997). Alzheimer-like visual deficits in down syndrome. *Alzheimer Disease Associated Disorders, 11*(2), 88–98.

Sabat, C., Arango, P., Tassé, M. J., & Tenorio, M. (2020). Different abilities needed at home and school: The relation between executive function and adaptive behaviour in adolescents with Down syndrome. *Scientific Reports, 10*(1). https://doi.org/10.1038/s41598-020-58409-5

Schmid, N. S., Ehrensperger, M. M., Berres, M., Beck, I. R., & Monsch, A. U. (2014). The Extension of the German CERAD neuropsychological assessment battery with tests assessing subcortical, executive and frontal functions improves accuracy in dementia diagnosis. *Dementia and Geriatric Cognitive Disorders Extra, 4*(2), 322–334. https://doi.org/10.1159/000357774

Schworer, E. K., Soltani, A., Altaye, M., Fidler, D. J., & Esbensen, A. J. (2023). Cognitive flexibility assessment in youth with Down syndrome: Reliability, practice effects, and validity. *Research in Developmental Disabilities, 133*, Article 104416. https://doi.org/10.1016/j.ridd.2022.104416

Strydom, A., Coppus, A., Blesa, R., Danek, A., Fortea, J., Hardy, J., Levin, J., Nuebling, G., Rebillat, A., Ritchie, C., Van Duijn, C., Zaman, S., & Zetterberg, H. (2018). Alzheimer's disease in Down syndrome: an overlooked population for prevention trials. *Alzheimer States Dementia Translational Research Clinical Interventions, 4*(1), 703–713. https://doi.org/10.1016/j.trci.2018.10.006

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*(1), 41–52. https://doi.org/10.1111/jir.12450

Toplak, M. E., West, R. F., & Stanovich, K. E. (2013). Practitioner review: Do performance-based measures and ratings of executive function assess the same construct? *Journal of Child Psychology and Psychiatry, 54*(2), 131–143. https://doi.org/10.1111/jcpp.12001

Tungate, A. S., & Conners, F. A. (2021). Executive function in down syndrome: A meta-analysis. *Research in Developmental Disabilities, 108*, Article 103802. https://doi.org/10.1016/j.ridd.2020.103802

Van Biesen, D., Van Damme, T., Pineda, R., & Burns, J. (2023). The impact of intellectual disability and sport expertise on cognitive and executive functions. *Journal of Intellectual Disabilities, 27*(1), 104–120. https://doi.org/10.1177/17446295211036331

Van Bysterveldt, A. K., Gillon, G. T., & Moran, C. (2006). Enhancing phonological awareness and letter knowledge in preschool children with down syndrome. *International Journal Of Disability Development And Education, 53*(3), 301–329. https://doi.org/10.1080/10349120600847706

Van Deusen, K., Prince, M. A., Thurman, A. J., Esbensen, A. J., Patel, L. R., Abbeduto, L., Walsh, M. M., Daunhauer, L. A., Feigles, R. T., & Fidler, D. J. (2023). Evaluating an adapted reverse categorisation task to assess cognitive flexibility in young children with Down syndrome. *Journal of Intellectual Disability Research, 67*(8), 734–745. https://doi.org/10.1111/jir.13040

Will, E., Fidler, D. J., Daunhauer, L., & Gerlach-McDonald, B. (2017). Executive function and academic achievement in primary – grade students with Down syndrome. *Journal of Intellectual Disability Research, 61*(2), 181–195. https://doi.org/10.1111/jir.12313

Williams, J., Rickert, V., Hogan, J., Zolten, A. J., Satz, P., D'Elia, L. F., … Light, R. (1995). Children's color trails. *Archives of Clinical Neuropsychology, 10*(3), 211–223. https://doi.org/10.1016/0887-6177(94)00041-n

Zelazo, P. D. (2006). The dimensional change card sort (DCCS): a method of assessing executive function in children. *Nature Protocols, 1*(1), 297–301. https://doi.org/10.1038/nprot.2006.46

Zelazo, P. D., Craik, F. I., & Booth, L. (2004). Executive function across the life span. *Acta Psychologica, 115*(2–3), 167–183. https://doi.org/10.1016/j.actpsy.2003.12.005