Citation: Henry, L. A., Messer, D. M., Poloczek, S., Dennan, R., Mattiauda, E. & Danielsson, H. (2024). The Reliability of Hebb Repetition Learning and its Association with Language and Reading in Adolescents with Intellectual Disabilities. Cortex, 177, pp. 253-267. doi: 10.1016/j.cortex.2024.05.012

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Research Report

The reliability of Hebb repetition learning and its association with language and reading in adolescents with intellectual disabilities

Lucy A. Henry [a,*], David J. Messer [a,b], Sebastian Poloczek [c], Rachel Dennan [a], Elisa Mattiauda [a] and Henrik Danielsson [d,e]

[a] Department of Language and Communication Science, City, University of London, London, UK
[b] Centre for Education and Educational Technology, WELS, The Open University, Milton Keynes, UK
[c] Department of Psychology, Goethe-Universität Institut für Psychologie, Frankfurt am Main, Germany
[d] Department of Behavioural Sciences and Learning, IBL Linköping University, Linköping, Sweden
[e] Linnaeus Centre HEAD, Swedish Institute for Disability Research, Linköping University, Linköping, Sweden

ARTICLE INFO

Article history:
Received 1 December 2023
Reviewed 9 January 2024
Revised 11 March 2024
Accepted 20 May 2024
Action editor Valentina Borghesani
Published online 31 May 2024

Keywords:
Hebb repetition learning
Reliability
Intellectual disability
Language
Reading

ABSTRACT

Hebb repetition learning (HRL) refers to neurodevelopmental processes characterised by repeated stimulus exposure without feedback, which result in changes in behaviour and/or responses, e.g., long-term learning of serial order. Here, we investigate effects of HRL on serial order memory. The present research aimed to assess the reliability of new HRL measures and investigate their relationships with language and reading skills (vocabulary, grammar, word reading) in adolescents with intellectual disability (ID). A comparison group of children of similar mental age with typical development (TD) was also assessed. ID and TD groups were tested on HRL tasks, evaluating test-retest and split-half reliability. The relationship between HRL and language and reading was analysed after accounting for the influence of mental age and verbal short-term memory. The HRL tasks displayed moderate test-retest (and split-half) reliability, HRL tasks with different stimuli (verbal, visual) were related, and we identified issues with one method of HRL scoring. The planned regression analyses failed to show relationships between HRL and language/reading skills in both groups when mental age, a very strong predictor, was included. However, further exploratory regression analyses without mental age revealed HRL's predictive capabilities for vocabulary in the ID group and reading in the TD group, results which need further investigation and replication. HRL displays promise as a moderately reliable metric and exhibits varied and interpretable predictive capabilities for language and reading skills across groups.

---

* *Corresponding author*. Department of Language and Communication Science, City, University of London, London, ECV1 0HB, UK.
E-mail addresses: lucy.henry.1@city.ac.uk (L.A. Henry), david.messer@open.ac.uk (D.J. Messer), Sebastian.Poloczek@gmx.de (S. Poloczek), henrik.danielsson@liu.se (H. Danielsson).

https://doi.org/10.1016/j.cortex.2024.05.012

1. Introduction

Intellectual Disability (ID) is a common neurodevelopmental condition (prevalence 1–3%), beginning in childhood, that is characterised by significant intellectual and adaptive functioning differences within conceptual, social, and practical domains (McKenzie et al., 2016; Patel et al., 2020). Despite the prevalence of ID, there is a lack of research concerning learning and memory, an absence that holds back practical support as well understanding of the mechanisms that are responsible for the learning differences. Our focus is on Hebb repetition learning (HRL) (Hebb, 1961) in adolescents with ID. Previous studies indicate that HRL may be commensurate with mental age level in those with ID (Henry et al., 2022; Mosse & Jarrold, 2010). Therefore, it is important to understand if and how HRL relates to language and reading (decoding) abilities, as these skills are vital to access classroom learning.

Hebb's (1949) innovative ideas about the strengthening of synaptic connections were among the first to provide a neurophysiological mechanism for learning and memory, and his ideas relate to the process of cellular consolidation (Sridhar et al., 2023). The term Hebb repetition learning (HRL) is used to describe neurophysiological changes that can occur when sequences of items are repeated, and this repetition results in long-term serial order learning. Investigations are beginning to uncover the brain structures involved in this process (Attout, Ordonez Magro et al., 2020; Loo et al., 2021). It seems likely that HRL involves neurophysiological changes where serial order information moves from short-term memory to a more robust long-term memory (Norris, 2017). These processes have been regarded as implicit and gradual, although recent evidence suggests most adults are aware of the repetition, and HRL at the individual level is characterised by abrupt improvements (Musfeld et al., 2023).

The fact that HRL involves serial order information moving from short-term memory to a more robust long-term memory trace means that it may play a critical role in the formation of stable linguistic representations such as phonological word forms (Attout, Ordonez Magro, et al., 2020). Several authors have reported HRL to be related to a range of language and reading skills in adults and children with and without neurodevelopmental conditions (e.g., Bogaerts et al., 2016). Furthermore, it has been suggested that HRL could provide a laboratory analogue for the acquisition of vocabulary and phonological word forms (e.g., Bogaerts et al., 2018; Norris et al., 2018; Page & Norris, 2008, 2009).

Relationships between HRL and vocabulary have been proposed both in children with typical development (TD) and those with specific neurodevelopmental conditions. Page and Norris (2009) suggested the processes involved in HRL are involved in the sequence-learning component of phonological word-form learning. Difficulties with serial-order HRL could negatively affect the long-term acquisition of phonological sequences, which are, in turn, important for acquiring new vocabulary. Several studies have reported links between HRL and the learning of novel phonological word-forms/pairings, and between HRL and assessments of receptive vocabulary (Archibald & Joanisse, 2013; Mosse & Jarrold, 2008; Smalle et al., 2018). Related research has demonstrated associations between novel word learning and short-term memory tasks that are designed, like HRL, to maximise the requirement for serial order retention (Majerus & Boukebza, 2013). Relationships are also present between these types of short-term memory tasks and receptive/productive vocabulary assessments (Attout, Grégoire, et al., 2020; Leclercq & Majerus, 2010). Finally, there is complementary evidence for challenges with HRL in children with developmental language disorder (Hsu & Bishop, 2014), although this is less consistent (Majerus et al., 2009).

HRL has additionally been linked to automatized reading (particularly of lexical items), whereby the creation of stable orthographic word forms involves the consolidation of long-term representations of grapheme sequences via repeated presentations (Bogaerts et al., 2018). The ability to learn novel sequential information via repeated exposure is hypothesised to be an underpinning skill involved in the sequential learning required in fluent reading (Attout, Ordonez Magro et al., 2020). Supporting evidence for this position includes relationships between HRL and reading in typical children (Attout, Ordonez Magro et al., 2020; Bogaerts et al., 2016; Smalle et al., 2018; although West et al., 2018 found no such relationships), and weaknesses in HRL in those with dyslexia (Bogaerts et al., 2015; Szmalec, et al., 2011).

There may also be relationships between HRL and grammar. The Procedural Deficit Hypothesis (Ullman & Pierpont, 2005) argues that difficulties with the procedural memory system in children with developmental language disorder lead to difficulties with both implicit sequence learning and grammar. It is assumed these difficulties cannot easily be compensated for by the declarative memory system. Challenges with grammatical skills represent one of the hallmarks of children with developmental language disorder (Bishop, 1997; Moraleda-Sepúlveda & López-Resa, 2022; Rice & Wexler, 1996) and these children may also show challenges with HRL (Hsu & Bishop, 2014) and other tasks that assess procedural memory (e.g., Kuppuraj et al., 2016; Lum et al., 2012; Lum et al., 2014). However, we lack evidence on whether HRL is related to vocabulary, grammatical skills, and/or reading in other groups with neurodevelopmental differences, such as ID.

Children and adolescents with ID can have below expected levels of reading (Lemons et al., 2013; Nilsson et al., 2021a, 2021b; Ratz & Lenhard, 2013), vocabulary and syntax abilities (van der Schuit et al., 2011). Therefore, if HRL is broadly in line with expected developmental level in those with ID (Henry et al., 2022), and is additionally related to language and reading, this offers the opportunity to capitalise on these neurodevelopmental serial order learning skills. By maximising the usefulness and contribution of automatic long-term learning processes involving stimulus repetition, we may be able to improve vocabulary, grammar and word reading in children and adolescents with ID, crucial first steps for improving their school success and social integration (e.g., greater facility in using social media). Thus, one important aim of the current study was to examine relationships between HRL, language and reading (assessed using a decoding task).

To properly examine these relationships, it is important to also consider standard measures of verbal short-term memory (VSTM), given extensive evidence of its relationships with vocabulary (Avons et al., 1998; Gathercole et al., 1992, 1997; Michas & Henry, 1994), reading (Cunningham et al., 2021) and grammatical abilities (Vulchanova et al., 2014) in children. Yanaoka et al. (2019) noted that VSTM was related to developmental changes in HRL in young children, so including both measures can reveal whether HRL explains variance in language and reading scores over and above that accounted for by VSTM.

Previous HRL and related research has raised important concerns about reliability (Bogaerts et al., 2018; Siegelman et al., 2017; West et al., 2018), although few studies have directly reported reliability data. West et al. (2018) expressed caution about their moderate to low split-half (internal consistency) and test-retest (stability over time) reliabilities (.58/.50 and .29 respectively) for children completing HRL tasks. Bogaerts et al. (2018) reported low split-half (.20–.43) and poor test-retest (−.26 to .28) reliabilities in adults. However, Mosse and Jarrold (2008) noted that different methods of measuring HRL in children produced both stronger and weaker split-half reliabilities. Finding a suitable way to measure HRL and minimise measurement error is important when investigating relations between HRL and other variables (Bogaerts et al., 2018).

Therefore, test-retest reliability of HRL tasks was assessed; this also provided split-half reliability estimates for our measures. In addition, correlations between two methods of scoring were evaluated, and analyses were conducted on the similarity between two different forms of stimuli that were presented in HRL tasks (visual and verbal modalities – given arguments for the domain generality of HRL, e.g., Couture & Tremblay, 2006; Mosse & Jarrold, 2008).

Halves scores have been used as indicator for HRL in several developmental studies (e.g., Archibald & Joanisse, 2013; Bogaerts et al., 2018; Mosse & Jarrold, 2008; Smalle et al., 2016). They represent the difference in Hebb vs filler performance on the second half of a HRL trial sequence, contrasted with the Hebb versus filler performance difference on the first half of trials. Therefore, they are a relatively crude measure of HRL learning. The advantage of these scores is that they are easy to calculate. However, the disadvantages are that by averaging the performance on second half trials, further HRL gains within the second half are ignored; equally, by averaging the performance on first half trials, gains on trials after the first one are treated as baseline performance level. Not using all the available trial information, therefore, may hamper accuracy and reliability.

As already outlined, our analyses focussed on an assessment of learning through repeated exposure, which was derived from the profile of memory performance over 16 trials in each of two sessions assessing HRL. This allowed us to identify a progressive difference between the 8 Hebb and 8 filler trials. It is important to emphasise that the focus on an overall assessment of learning meant that traditional assessments of split-half reliability, such as consistency between odd and even trials were less relevant.

Given concerns about weak HRL effects in very young children, no participants with chronological and mental ages younger than four years were included (Yanaoka et al., 2019). HRL tasks were designed to maximize reliability with: (a) two separate HRL testing sessions to increase the number of trials without increasing fatigue and to evaluate test-retest reliability (West et al., 2018); (b) individually-titrated supraspan Hebb and filler list lengths based on each participant's memory span level to prevent floor and ceiling effects (Archibald & Joanisse, 2013; Hsu & Bishop, 2014; Smalle et al., 2016; West et al., 2018; Yanaoka et al., 2019); (c) sensitive scoring methods to reflect correct recall of both item and position information (Smalle et al., 2016) *and* comparisons of different scoring methods (Kalm & Norris, 2016); (d) non-overlapping items in Hebb and filler lists to maximise Hebb effects (Smalle et al., 2016; Yanaoka et al., 2019); and (e) using a serial order reconstruction recall method to reliably produce Hebb effects at the group level (Johnson et al., 2017).

Two pre-registered research questions were addressed. Firstly, do the measures of HRL show adequate test-retest reliability? Related to this, do they show adequate split-half reliability when using our overall measure of HRL that combined the information from two sessions and, therefore, two halves of our experiment? Also, were two methods of scoring HRL related, and was learning from different types of stimuli (verbal, visual) related? It was predicted that these reliability values would be adequate for both groups, but the basis for this prediction was not strong given previous findings. Secondly, to what extent is HRL related to language and reading abilities (vocabulary, grammar, word reading) in adolescents with ID and children with TD? We predicted, based on a limited previous literature, positive relationships between HRL and all three measures of language and reading in both study groups. Moreover, we tentatively expected HRL to remain a significant predictor even after accounting for mental age and VSTM, given previous reports that VSTM is dissociable from HRL (Bogaerts et al., 2015; Mosse & Jarrold, 2010).

2. Materials and methods

We report how we determined our sample size, all data exclusions, all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study.

2.1. *Participants*

This study included 52 adolescents with ID (28 females, mean age 13 years 4 months, SD 15m) and 47 children with TD (28 females, mean age 7 years 5 months, SD 16m). Green's (1991) suggested formula (see also Tabachnick & Fidell, 2014) for calculating sample size, taking into account the size of the anticipated effect, was used. This calculation was based on expected medium effect sizes in the regression models, which were run separately for each group, hence: $N \geq (8/f^2) + (m - 1)$ where f^2 refers to effect size ($f^2 = .15$ for a medium effect size) and m refers to the number of independent variables. This value equates to a sample size of at least 55 in each group. Groups were significantly different on mental age, IQ, and chronological age, see Table 1. However, mental age ranges of the two groups were overlapping and there were few group differences in language, reading, and short-term memory

Table 1 – Mean score, standard deviations (*SD*), ranges on key study variables for adolescents with ID and children with TD of comparable mental age.

Variables	Measure	Adolescents with ID ($n = 52$)			Children with TD ($n = 47$)			Group differences
		M	SD	Range	M	SD	Range	
Chronological age	Years:Months or Months	13:04	15	137–191	07:05	16	60–123	$t(97) = 22.37, p < .001$
Mental age	Years:Months or Months	06:11	12	55–111	07:05	13	60–111	$t(97) = -2.16, p = .033$
Stanford Binet-5 Abbreviated IQ	IQ	62.67	9.93	47–79	98.60	10.48	85–121	$t(97) = -17.51, p < .001$
British Picture Vocabulary Scale-3	Raw Score	109.58	21.55	66–149	104.64	17.82	67–142	$t(97) = 1.24, p = .220$
Test for Reception of Grammar-2	Blocks Passed	10.62	4.33	2–18	13.30	3.78	3–18	$t(97) = -3.27, p = .001$
Test of Word Reading Efficiency-2 (Sight Word Efficiency)	Raw Score	61.75	15.48	28–97	56.36	20.15	1–87	$t(97) = 1.50, p = .137$
Test of Word Reading Efficiency-2 (Phonemic Decoding Efficiency)	Raw Score	27.88	13.61	5–54	29.68	14.68	0–62	$t(97) = -.63, p = .529$
Word List Recall	Span	3.45	.52	2.00–4.50	3.44	.61	2.25–5.00	$t(97) = .08, p = .938$
Visual Sequential Memory	Span	3.52	1.09	2–6	3.53	1.08	2–6	$t(97) = -.06, p = .954$
Vineland: Communication[a]	Standardised Score	74.35	8.69	54–90				
Vineland: Daily Living Skills[a]	Standardised Score	83.43	11.13	48–105				
Vineland: Socialisation[a]	Standardised Score	83.20	10.47	60–109				
Vineland: Adaptive Behaviour Composite[a]	Standardised Score	78.24	7.76	59–94				

[a] Missing data for one ID participant ($n = 51$).

measures (Table 1). All participants had spoken English for at least two years at the time of testing, as confirmed by teachers.

Due to the *COVID*-19 pandemic, recruitment and testing ceased one month earlier than planned. The resulting sample sizes were affected, as some children (largely in the TD group) had not completed the study. We applied the following rules about inclusion to all participants who fulfilled the recruitment criteria (see below). All adolescents with ID with relevant data available were included (Hebb session 1, language and reading measures); plus all children with TD who had relevant data available and who also had mental ages within the same range as the ID group. These criteria were designed to be transparent, avoid losing data on completed participants, and maximise participant numbers. Exact mental age matches were not required because predictive relations between HRL and language/reading were tested within each group.

Adolescents with non-specific ID (no biological cause for ID was identified) were recruited. We applied DSM-5 (American Psychiatric Association, 2013) criteria for ID, using both cognitive and adaptive functioning (Patel et al., 2020), with less emphasis on exact cut-off scores (Burack et al., 2021). The study was preregistered on the Open Science Framework (OSF) https://osf.io/gkpwh/ (dated 17.01.19); a minor change to the inclusion criteria for the ID group was registered on 08.08.19 under Transparent Changes on the OSF (citation: osf.io/a5724); and we report data relevant to the second and third sets of pre-registered research questions (first set of pre-registered research questions reported in Henry et al., 2022).

In the ID group, 11–15-year-olds were recruited from 27 mainstream secondary schools in England (Greater London, Hertfordshire, Yorkshire, Cambridgeshire, Nottinghamshire). Teachers identified eligible young people if they had ID and no other diagnoses such as autism or Down syndrome. Participants were excluded if, after testing, they did not have mild to moderate ID. The ID group had: 1) a score of 40–79 on the Stanford-Binet Abbreviated Intelligence Scales (SB-5 ABIQ: Roid, 2003); and 2) a standardized score of 40–85 on one or more of the core Vineland Adaptive Behaviour Scales (Vineland-3: Sparrow et al., 2016) domains (Communication, Daily Living Skills, Socialisation) or on the overall Adaptive Behaviour Composite (ABC). Thirteen participants with ID had IQ scores in the borderline (70–79), 28 in the mild (55–69), and 11 in the moderate (40–54) ranges; all participants showed evidence of adaptive difficulties.

In the TD group, 5–10-year-olds were recruited from 7 mainstream primary schools in England (Greater London and Yorkshire). These participating schools were comparable to those of the ID group on socio-economic status; in both groups a higher proportion of students than the national average were from ethnic minority groups, spoke English as an additional language, and were eligible for pupil premium funding. Teachers identified eligible children who did not have special educational needs or diagnosed neurodevelopmental conditions. Children were included if their mental age, based on the verbal and non-verbal assessments from the SB-5 ABIQ, was in the same range as the ID group, and if they did not have ID, defined as a standardised score of 85 or above on the abbreviated version of the SB-5 ABIQ. Full details of the samples are provided in Table 1.

Ethical approval for the study was granted by the relevant University. Written informed consent from parents/guardians and written and verbal assent from participants was gained before testing.

2.2. *Design*

For the correlation analyses to estimate reliability and relationships between different assessments of HRL, a mixed factorial quasi-experimental design was conducted with three within-subjects factors to extract the HRL indicators. The within-subjects factors were HRL task list type (Hebb, filler), trial position (eight trials for each list type), and type of material (verbal, visual Hebb task). To examine the relations of HRL to language and reading a multiple regression design was employed with HRL, mental age and VSTM as predictors and vocabulary, grammar and single word reading as dependant variables.

2.3. Materials

2.3.1. IQ and mental age

The abbreviated version of the Stanford Binet, fifth edition (SB-5, Roid, 2003) provided an estimate of IQ (ABIQ), including two subtests: verbal knowledge and non-verbal reasoning skills. Manual-derived mental ages in months were used to describe samples and these scores were used as control variables in analyses. The SB-5 has high split-half (.85–.92) and test-retest (.84) reliability, as well as high validity, for the present age and ability range (Roid, 2003).

2.3.2. Adaptive functioning

Parents/caregivers of participants in the ID group completed the Vineland Adaptive Behavior Scales, third edition, Domain-Level Parent/Caregiver Form (VABS-3, Sparrow et al., 2016) via a telephone interview. In a few cases, parents completed the questionnaire themselves or teachers filled out the VABS-3 Domain-Level Teacher Form (Sparrow et al., 2016). One participant with missing Vineland data was included, given their ABIQ of 52, school-reported difficulties with cognition, learning, and communication, and receipt of in-class support. Standardised measures of adaptive functioning in three domains (communication, daily living skills, socialisation) were used to calculate an overall adaptive behaviour composite (ABC). The VABS-3 is a reliable and valid assessment with excellent test-retest reliabilities (.81–.92) (Sparrow et al., 2016).

2.3.3. Vocabulary

The British Picture Vocabulary Scale, third edition (BPVS-3: Dunn et al., 2009) assessed receptive vocabulary. Testing started on the set corresponding with chronological age for the TD group and on the ‘Age 7’ set for the ID group (the experimenter tested backwards to find the basal set if necessary). Total raw scores were used, and we performed within-group (ID/TD) z-standardisation of raw scores for use in analyses. We did not use norm-based standard scores because many participants had the lowest chronological-age-norm-based values (the BPVS-3 gives no values below 70). The BPVS-3 has a reliability of .91, and strong correlations to other related tests, providing evidence for validity (Dunn et al., 2009).

2.3.4. Grammar

The Test for Reception of Grammar, second edition (TROG-2: Bishop, 2003) assessed understanding of grammatical constructs. Raw scores reflected the total number of blocks passed and we performed within-group (ID/TD) z-standardisation of raw scores for use in the analyses. We did not use norm-based standard scores because many children and adolescents had the lowest chronological-age-norm-based values (the TROG-2 gives no values below 55). Internal consistency of the TROG-2 is high (.88); parallel form reliability is moderate (.67); and validity is well-established (Bishop, 2003).

2.3.5. Single word reading

The Test of Word Reading Efficiency, second edition (TOWRE-2: Torgesen et al., 2012) assessed reading accuracy and fluency via Sight Word Efficiency (SWE – number of real words identified) and Phonetic Decoding Efficiency (PDE – number of non-words decoded). Raw scores on each subtest were used. For analyses, a combined z-score reflecting the average z-scores of the two tasks was calculated (giving equal weight to each subtest given the high correlation between SWE and PDE, $r = .81$). Within group z-standardisation of each task was carried out first, then, after averaging these scores, a second z-standardisation was carried out to produce a z-standardized dependent variable for overall reading ability. Reliability of the TOWRE-2 is high (.93 and .94 for SWE and PDE).

2.3.6. Word List Recall

Word List Recall from the Working Memory Test Battery for Children (WMTB-C: Pickering & Gathercole, 2001) assessed VSTM. Participants repeated back, in correct serial order, lists of one-syllable words spoken by the experimenter. Word lists increased incrementally in length, beginning with lists of a single word. There were six trials for each span list length to ensure sensitivity to performance differences between children, and if four out of six trials were recalled correctly, the next block at a higher list length was offered. Testing ceased if the participant scored 3/6 or fewer trials correct in a block. Reliability on this task is reported between .83 and .38 for primary school age children (Pickering & Gathercole, 2001). A memory span score was calculated to reflect both span level (the list length at which the child could recall at least 4 out of 6 trials), plus correct trials at list lengths above span giving credit for partial success beyond span level and increasing the sensitivity of the measure. For each list correctly recalled at the list length above span, .25 was added to the span score (e.g., a full pass at list length 3 plus two correct trials out of six at list length 4 gives a span score of 3.5). This score was used to measure VSTM and to titrate difficulty level on the verbal HRL task.

2.3.7. Visual Sequential Memory

Visual Sequential Memory from the Test of Memory and Learning, second edition (TOMAL-2: Reynolds & Voress, 2007) assessed visual STM. Participants saw sets of horizontally displayed nonsense visual stimuli for five seconds. These were removed and immediately re-presented in a different order, with the instruction, “point to the drawings in the order you saw them on the page before”. Sequence lengths increased incrementally with two trials for each sequence length. Testing ceased if the participant failed to recall any items in the correct order for two consecutive trials. This task has high internal consistency (.78–.92) and good test-retest reliability (.71) for the age range in the present study (Reynolds & Voress, 2007). Memory span scores were calculated using performance on different list lengths within this task (although note that this is not the way scores are calculated in the test manual). The child's span score was derived from the longest list length at which perfect recall in serial order was achieved (e.g., a perfect recall in one out of two trials at a list length of 3, would give a span of 3). This span score was used to titrate difficulty level on the visual HRL task.

Legal copyright restrictions prevent public archiving of the SB-5, VABS-3, BPVS-3, TROG-2, TOWRE-2, WMTB-C and TOMAL-2, which can be obtained from the copyright holders in the cited references.

2.3.8. *Hebb repetition learning tasks*

Participants received two HRL tasks presented on an iPad: a verbal Hebb task using easily nameable and identifiable pictures of common objects with one-syllable names; and a visual Hebb task using difficult-to-name, unfamiliar 'nonsense drawing' stimuli. In the verbal HRL task, participants simultaneously heard the item's name while seeing its picture on the screen, presented at a rate of one item every 1.5 sec and with a .5 sec interval between items. The nonsense shapes in the visual HRL task were presented visually only, without any accompanying sound, at the same rate.

There were always 8 Hebb trials (the to-be-remembered Hebb sequence was repeated 8 times), alternating with 8 filler trials (randomly generated novel sequences on each trial), totalling 16 trials. The 16 trials always started with filler trials and alternated thereafter between filler and Hebb trials for each HRL task. Items were drawn from different item sets for Hebb and filler trials (item sets for Hebb and filler stimuli were non-overlapping, Smalle et al., 2016). The item sets included two sets of 8 nonsense pictures; and two sets of 10 easily nameable one-syllable nouns illustrated as black and white line drawings (List A = dog, car, kite, chair, bell, ring, sun, fish, sock, house; List B = book, cat, bus, cup, bed, pear, comb, ball, duck, shirt). For examples of the items and response arrays, please see Fig. 1a and b. All images are available at: https://osf.io/9g5w2/

Further task information is available here: https://app.gorilla.sc/openmaterials/782819.

In each HRL task, participants were presented with a sequence of items one at a time. After presentation, participants were shown an array of items from the relevant item set, 10 items for the verbal task and 8 items for the visual task, placed in the lower half of the screen in two rows, randomly ordered on each trial. The top of the response screen included horizontal lines corresponding to the length of the list being recalled, to cue how many responses were needed (Fig. 1a and b). Participants recalled items in correct serial order by sequentially touching images on the array screen. After an item had been selected from the response array, a black circle appeared in the relevant line at the top of the screen to signify the participant had selected an item. Items could be selected more than once, although target sequences never contained repetitions.

Both Hebb tasks were introduced to participants with short practice trials using different item sets (verbal practice item set = tree, dress, bowl, star, shoe, key, drum, flag, fork, cake; visual practice item set = eight further nonsense drawings; see https://osf.io/9g5w2/). There were four practice trials, two presenting shorter list lengths (two items for visual and three items for verbal Hebb versions) and two presenting the same list length as the actual task (which differed depending on the difficulty level assigned). Participants received a virtual coin for each trial completed in every task (the coin appeared to 'land' inside a money bag with a 'ping') and a gold trophy was presented on the screen at the end of each task session or practice session.

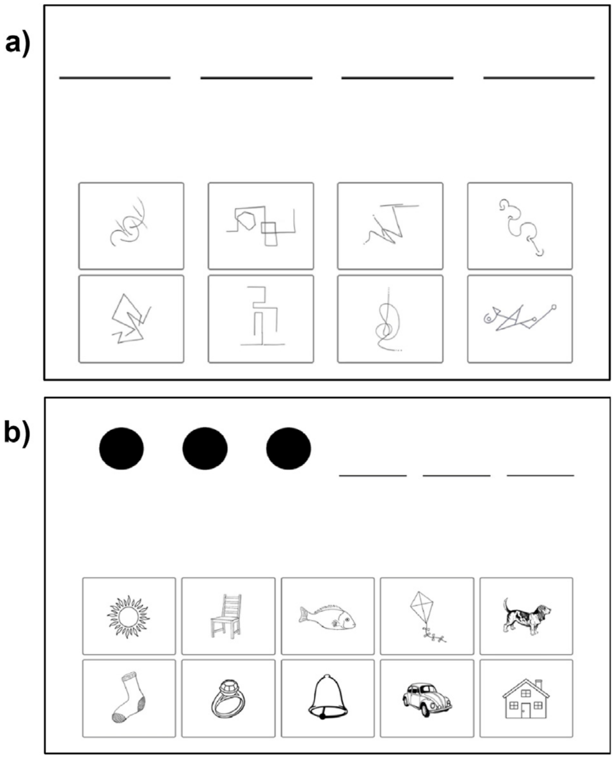

Fig. 1 – a. Response array from the visual Hebb repetition task. Here the participant has not started responding. b. Response array from the verbal Hebb repetition task. Here, the participant has made three responses as indicated by the black circles.

2.3.8.1. Hebb task allocation. Participants with VSTM spans of 3.5 or greater received longer verbal HRL lists (6 items; $N_{ID} = 26$, $N_{TD} = 25$); those with spans of 3.25 or less received shorter lists (5 items; $N_{ID} = 26$, $N_{TD} = 22$). Participants with visual STM spans of 3 or greater received longer visual HRL lists (4 items; $N_{ID} = 42$, $N_{TD} = 39$); those with spans of 2 or less received shorter lists (3 items; $N_{ID} = 10$, $N_{TD} = 8$). HRL lists were, therefore, at least one item above span (supraspan).

2.3.8.2. Nonsense drawing familiarisation. Before administering the visual Hebb task, participants were pre-familiarised with the nonsense drawing stimuli by playing a game of 'Snap' (see: Henry et al., 2022).

2.3.8.3. Counterbalancing. Two HRL sessions were administered up to five weeks apart to assess test-retest and split-half reliability (mean days between sessions ID = 15.2, range 7–35; TD = 11.8, range 1–16). Variability emerged due to *COVID*-19 truncating testing, particularly for the TD group. The order of presentation of Hebb tasks was counterbalanced across the two sessions, and items in each sequence were chosen via semi-randomized selection, but otherwise did not differ in format. Two parallel versions of each task were counterbalanced across participants: for session 2 the filler and Hebb stimuli set were reversed, and this was counterbalanced across participants. The HRL task the participant received first was randomised.

2.3.8.4. Hebb repetition learning task scoring. Two methods of scoring were used. For ordinal scoring, at every trial, participants scored one point for each item within a sequence correctly recalled (item score) plus a further point for each item in the correct serial order (order score). For each item, the ordinal information of 0 = 'no recall', 1 = 'item recall (not in position)' and 2 = 'item recall in position' was used for analysis. This ensured credit for correct item and position information, such that partial knowledge of the sequences was considered in the scores, important for obtaining reliable measures of inter-individual differences in HRL (Bogaerts et al., 2018).

An additional scoring method, Levenshtein edit-distance metrics, was calculated, "defined as the minimum number of edits needed to transform one string into another" (Kalm & Norris, 2016, p. 112). Edit distance was divided by list length and subtracted from 1 to derive a standardised metric, referred to as Levenshtein scores or scoring. These edit-distance metrics capture similarities between the target sequence and the recalled sequence, making minimal assumptions about what is being learned. The Levenshtein scoring method was exploratory as it was not pre-registered.

2.4. *Procedure*

Testing took place one-to-one in schools for both groups during lesson time, with session lengths adapted to children's needs and school schedules. For the ID group, the typical assessment time was 90 min, usually split across two sessions: session 1 included SB-5, STM measures, TOWRE-2, and first HRL tasks; session 2 included BPVS-3, TROG-2 and the second HRL tasks. Most of the TD group completed the activities in three sessions of approximately 30 min each: session 1 included SB-5, STM measures, and TOWRE-2; session 2 included first HRL tasks and TROG-2; session 3 included second HRL tasks and BPVS-3. Certificates were provided after the final session as a reward; children in the TD group also received stickers.

2.5. *Approach to analyses*

2.5.1. *Hebb repetition learning*

Scores were available for two types of sequences (Hebb lists, filler lists) from 8 trials in each case (trials 1 through 8). A positive interaction effect between list type (Hebb *vs* filler) and trial position (1 through 8) represents the degree of HRL, as it captures recall improvement over Hebb trials in comparison to no improvement or decline in performance on filler trials. Generalized Linear Mixed Models (GLMMs) were used in our analyses, which included both this fixed effect of list type × trial position, representing the average effect of HRL, and the random effect of list type × trial position, allowing for inter-individual differences in the degree of HRL. In mixed effects models the residuals of the random effects reflect how an individual differs from the group mean (fixed effects). Including a random intercept additionally captures individual differences in recall performance on the first Hebb and filler trials. Note that Bogaerts and colleagues (Bogaerts et al., 2016, 2018) introduced mixed logit models to analysing the development of HRL with the recall of items in the correct position as the dependent, binary variable (see also Yanaoka et al., 2019). All GLMMs were run with MLwiN 3.02 (Charlton et al., 2018) using MCMC estimation, with 100,000 iterations and thinning to 5000 estimates from R with the R2MLwiN package 0.8.7 (Zhang et al., 2016). Full data and R scripts are available at: https://osf.io/9g5w2/

2.5.2. *Reliability analyses*

In the preregistration, we planned to use HRL residuals from a model that combined both verbal and visual trials using 'ordinal' scoring. However, Levenshtein edit distance metrics are another suitable measure for learning the item sequence on repeated Hebb trials. Consequently, we extracted HRL residuals on parallel binomial models with normalized Levenshtein scores as the dependent variable. Test-retest reliability was calculated from the correlations between HRL residuals from session 1 and session 2. The same calculation provided an estimate of split-half reliability of our overall measure of HRL which was based on both sessions and both types of stimuli. To obtain overall reliability/split-half reliability, the test-retest correlations were adjusted according to the Spearman-Brown formula to take account of the reduced number of trials when calculating HRL residuals for either session 1 or 2 in comparison to all trials used when extracting the HRL predictor. In addition, we examined whether HRL with the two different types of stimuli were related, and whether the two methods of coding were related.

3. Results

Data on the reliability of HRL scores and the relationship between the two scoring methods are presented first, followed by analyses addressing the prediction of language and reading from HRL and other key variables. As no overall group differences in HRL were found in earlier analyses (Henry et al., 2022), we present the first set of reliability analyses carried out on combined ID and TD groups for ease of presentation and to obtain a larger sample size in the analyses.

3.1. *Reliability and convergent validity of the Hebb repetition learning scores*

Separate GLMM models on data from Sessions 1 and 2 were run to determine test-retest correlations and provide an estimate of split-half reliability for both ordinal and Levenshtein scoring. The visual and verbal data were combined in the initial analyses as specified at preregistration. For test-retest reliability of ordinal scoring, HRL residuals were significantly correlated $r_{\text{test-retest}} = .39^{**}$. This correlation also gives the split-half reliability for this method of coding, and an estimated split-half reliability of .56 with the Spearman-Brown adjustment. For Levenshtein test-retest reliability, HRL residuals were correlated $r_{\text{test-retest}} = .43^{**}$. This correlation also provides a measure of split-half reliability, which is increased to .60 with the Spearman-Brown formula. These reliability estimates for both scoring methods were consistent with each other but moderate.

Next, we considered relationships between the scoring methods in Sessions 1 and 2. The findings were as follows. (1) Trial level ordinal and Levenshtein scores were highly correlated, $r = .94$, indicating that these scoring methods both captured similar aspects of memory performance. (2) Fixed effects were comparable between the models based on ordinal and Levenshtein scoring (see supplement to Henry et al., 2022 and Table S2 in the supplement to this paper), indicating that in group-level analyses comparing the size of HRL between verbal and visual materials, or individuals with and without ID, both scoring methods provided similar results. (3) Correlations between HRL residuals of Levenshtein versus ordinal scoring for models with *both* materials, i.e., both the combined verbal and visual materials, were low and non-significant ($r = .00$ and $r = .09$; see Table 3). Further, HRL residuals for each participant from both materials with ordinal scoring were not well correlated with the residuals from either verbal material or visual material. Specifically, for ordinal scores, the part-whole correlations between residuals based on verbal materials and those on both materials were non-significant to low positive ($r = .18$ and $r = .28$**); for visual material the part-whole correlations were unexpectedly negative ($r = -.49$** and $r = -.52$**; see Table 2).

For a reliable and valid scoring and analysis approach, one would expect to find positive moderate to high part-whole correlations, as was the case for Levenshtein scoring ($r = .59$** to $r = .70$**). Additionally, models based on Levenshtein scores showed moderate test-retest correlations ($r = .36$** to $r = .43$**) and moderate cross-modal correlations (i.e., between the verbal and visual modalities, $r = .42$**; $r = .47$**) when HRL residuals were either based on verbal or on visual material (see Table 2).

Table 2 – Correlations within scoring method (Ordinal *vs* Levenshtein) across data slices differing in one dimension for GLMMs.

		Ordinal		Levenshtein	
		r	CI	*r*	CI
Test-Retest (Ses1-Ses2)	both materials	.39**	[.20, .55]	.43**	[.25, .59]
	verbal	.57**	[.42, .69]	.36**	[.17, .52]
	visual	.68**	[.55, .78]	.43**	[.25, .59]
Cross-modal (verbal-visual)	Session 1	.55**	[.40, .68]	.47**	[.30, .61]
	Session 2	.54**	[.38, .67]	.42**	[.23, .57]
Part-Whole (verbal-both)	Session 1	.18	[−.01, .37]	.59**	[.44, .70]
	Session 2	.28**	[.08, .45]	.70**	[.58, .79]
Part-Whole (visual-both)	Session 1	−.49**	[−.63, −.32]	.67**	[.55, .77]
	Session 2	−.52**	[−.65, −.35]	.70**	[.58, .79]

Table 3 – Correlations of Ordinal versus Levenshtein scoring within same data slice for HRL residuals from GLMMs.

	r	CI
Both materials – Session 1	.09	[−.11, .28]
Both materials – Session 2	.00	[−.20, .21]
Verbal – Session 1	.61**	[.47, .72]
Verbal – Session 2	.66**	[.53, .76]
Visual – Session 1	.79**	[.70, .85]
Visual – Session 2	.93**	[.89, .95]

These findings show that preregistration plans for regression analyses based on *combined ordinal scores* needed to be reconsidered, as ordinal scores from combined visual and verbal trials had lower correlations with other measures of HRL. In contrast, residuals based on *combined Levenshtein scores* from both types of materials were as reliable (.60) and had high part-whole correlations with both verbal and visual scores. Consequently, we based the HRL measure for predicting language and reading on the model closest to that in the preregistration document: combined HRL residuals of Levenshtein scores from both materials.

3.2. Hebb repetition learning as a predictor of language and reading

To determine whether HRL learning predicted language and reading, hierarchical regression models were set up for both groups, with each language and reading measure as a dependent variable. First, mental age was entered (Model A); second, HRL was added (Model B) to test whether HRL predicts language and reading skills after controlling for differences in mental age; third, VSTM was added (Model C) to examine whether a potential predictive effect of HRL remained once this variable was included.

Correlations between all variables are displayed in Table 4. Of note, in the ID group, there was a significant correlation between HRL and vocabulary and, in the TD group, there were significant correlations between HRL and both grammar and reading. In the ID group, HRL was not related to VSTM ($r = .11$), whereas in the TD group, these two variables were moderately related ($r = .40$). Almost all variables related to each other in the TD group; whereas fewer significant relationships emerged in the ID group.

Summary information about the regressions is presented in Tables 5–7. For the ID group, mental age, entered first, was a strong predictor of vocabulary and grammar, explaining 47% and 27% of the variance respectively; although it was a weak predictor of reading (6% of variance explained; Model 1a). For the TD group, mental age was a stronger predictor of all three language and reading measures (the amount of variance explained ranged between 37% and 62%).

When HRL and VSTM were entered second and third, respectively, they were not significant predictors in any of the models (for details see Table 5; models 1b and 1c). However, there were some marginal effects. In children with TD, the prediction of grammar showed weak evidence of improvement [by 5%; Model Comparison B *vs* A: $F(44,1) = 4.22$, $p = .046$] when HRL was added to mental age ($\beta = .24$, 95% CI [.00, .49], $p = .051$) and marginal further improvement [by another 4%; Model Comparison C *vs* B: $F(43,1) = 3.11$, $p = .08$], when VSTM was added ($\beta = .22$, 95% CI [−.03, .48], $p = .08$). For children with TD, reading skills were strongly linked to mental age (40% explained variance) with no additional variance explained by HRL or VSTM. For adolescents with ID neither mental age, HRL or verbal STM significantly predicted reading skills.

Table 4 – Correlations between mental age, memory, language and reading measures for the ID group (lower triangle) and the TD group (upper triangle).

	MA	VSTM	HRL	BPVS-3	TROG-2	TOWRE-2
Mental Age (MA)		.44** [.17, .64]	.40** [.12, .61]	.79*** [.65, .88]	.61*** [.39, .76]	.63*** [.42, .78]
Verbal Short-term Memory (VSTM)	.33* [.06, .55]		.40** [.12, .61]	.36* [.08, .59]	.48** [.23, .68]	.29*$^{\#}$ [.00, .53]
Hebb Repetition Learning (HRL)	.51*** [.28, .69]	.11 [−.17, .37]		.29^{+} [.00, .53]	.36* [.09, .59]	.42** [.15, .63]
Vocabulary (BPVS-3)	.68*** [.51, .81]	.31* [.04, .54]	.40** [.15, 61]		.64*** [.44, .79]	.44** [.18, .65]
Grammar (TROG-2)	.52*** [.28, .69]	.34* [.07, .56]	.26^{+} [−.01, .50]	.53*** [.30, .70]		.43** [.16, .64]
Reading (TOWRE-2)	.24^{+} [−.04, .48]	.29*$^{\#}$ [.02, .52]	.22 [−.06, .46]	.22 [−.06, .46]	.03 [−.25, .30]	

*$p < .05$. **$p < .01$; ***$p < .001$; ($^{+}p < .10$); $^{\#}$reduced significance with *Benjamini–Hochberg procedure* for controlling the false discovery rate.

Table 5 – Hierarchical regression models predicting vocabulary in the samples of students with ID and TD children.

	ID group ($n = 52$)					TD group ($n = 47$)				
	Model 1a	Model 1b	Model 1c	Model 2a	Model 2b	Model 1a	Model 1b	Model 1c	Model 2a	Model 2b
Mental Age	.68*** [.48, .89]	.65*** [.41, .89]	.61*** [.35,.87]			.79*** [.60, .97]	.80*** [.60, 1.00]	.79*** [.57, 1.00]		
Hebb Repetition Learning		.07 [−.17, .31]	.08 [−.16, .32]	.40** [.14, .66]	.37** [.12, .63]		−.03 [−.23, .17]	−.04 [−.25, .18]	.29+ [.00, .57]	.17 [−.13, .47]
Verbal Short-term Memory			.11 [−.12, .33]		.27* [.02, .52]			.03 [−.19, .25]		.29+ [−.01, .60]
R^2	.47	.47	.48	.16	.24	.62	.62	.62	.08	.15
F	44.1***	22.0***	14.9***	9.7**	7.6**	73.7***	36.2***	23.6***	4.0+	4.0*
ΔR^2		.00	.01		.07		.00	.00		.07
ΔF		.35	.92		4.66*		.09	.08		3.77+

*$p < .05$. **$p < .01$; ***$p < .001$; (+$p < .10$).

Table 6 – Hierarchical regression models predicting grammar in the samples of students with ID and TD children.

	ID group ($n = 52$)					TD group ($n = 47$)				
	Model 1a	Model 1b	Model 1c	Model 2a	Model 2b	Model 1a	Model 1b	Model 1c	Model 2a	Model 2b
Mental Age	.52*** [.27, .76]	.52*** [.23, .80]	.44** [.15, .75]			.61*** [.37, .85]	.55*** [.29, .81]	.47** [20, .74]		
Hebb Repetition Learning		.00 [−.29, .28]	.01 [−.27, .29]	.26+ [−.01, .54]	.23+ [−.04, .49]		.15 [−.11, .40]	.08 [−.18, 34]	.36* [.08, .64]	.20 [−.08, .49]
Verbal Short-term Memory			.19 [−.07, .49]		.31* [.05, .58]			.24+ [−.02, .51]		.40** [.12, .68]
R^2	.27	.27	.30	.07	.17	.37	.39	.43	.13	.27
F	18.2***	8.9***	6.8***	3.69+	4.87*	26.35***	13.9***	10.9***	6.87*	8.03**
ΔR^2		.00	.03		.10		.02	.04		.13
ΔF		.00	2.23		5.70*		1.36	2.06+		8.09**

*$p < .05$. **$p < .01$; ***$p < .001$; (+$p < .10$).

3.3. *Exploratory post-hoc regressions*

Controlling for mental age can only answer questions about whether HRL and VSTM predict language and reading beyond the contribution of mental age. Yet mental age might account for so much variance in language and reading performance that other important relationships are obscured, especially because mental age was based on both non-verbal and verbal abilities. Given interest in the relative roles of HRL and VSTM in predicting language and reading, exploratory regression models were run without first controlling for mental age: HRL was entered first (Model 2a) and VSTM second (Model 2b). These analyses are considered exploratory as they were not pre-registered. Please see Tables 4–6 for details.

3.3.1. *Vocabulary*

For adolescents with ID, both HRL (β = .37, 95% CI [.12, .63], $p = .005$) and VSTM (β = .27, 95% CI [.02, .52], $p = .04$; Model 2b)

Table 7 – Hierarchical regression models predicting word reading in the samples of students with ID and TD children.

	ID group ($n = 52$)					TD group ($n = 47$)				
	Model 1a	Model 1b	Model 1c	Model 2a	Model 2b	Model 1a	Model 1b	Model 1c	Model 2a	Model 2b
Mental Age	.24+ [−.04, .51]	.17 [−.15, .49]	.08 [−.25, .41]			.63*** [.40, .87]	.55*** [.31, .80]	.57*** [.30, .84]		
Hebb Repetition Learning		.13 [−.19, .45]	.15 [−.17, .47]	.22 [−.06, .49]	.19 [−.09, .46]		.20 [−.05, .45]	.21 [−.05, .47]	.42** [.15, .69]	.36* [.06, .66]
Verbal Short-term Memory			.25+ [−.04, .54]		.27+ [.00, 54]			−.05 [−.31, .22]		.14 [−.15, .44]
R^2	.06	.07	.12	.05	.12	.40	.44	.44	.18	.19
F	2.93+	1.78	2.22+	2.44	3.27*	30.23***	16.93***	11.1***	9.54**	5.24**
ΔR^2		.01	.05		.07		.03	.00		.02
ΔF		.67	2.96+		3.95+		2.52	.12		0,95

*$p < .05$. **$p < .01$; ***$p < .001$; (+$p < .10$).

were significant predictors of vocabulary. For children with TD, when HRL was the only predictor, it was marginally significantly linked to vocabulary (β = .29, 95% CI [.00, .57], $p = .05$). When VSTM was also included in the model, HRL was no longer significant (β = .17, 95% CI [−.13, .47], $p = .26$), and VSTM was not significant either (β = .29, 95% CI [−.01, .60], $p = .06$). Thus, in the ID group HRL and VSTM were significant predictors of vocabulary, but in the TD group these variables were not significant predictors of vocabulary.

3.3.2. *Grammar*

In adolescents with ID, HRL (β = .23, 95% CI [−.04, .49], $p = .09$) was not a significant predictor of grammar, whereas VSTM was (β = .31, 95% CI [.05, .58], $p = .02$). For children with TD, HRL was a significant predictor of grammar (β = .36, 95% CI [.08, .64], $p = .01$) when entered alone, but with both predictors entered, only VSTM was significant (β = .40, 95% CI [.12, .68], $p = .01$). These analyses suggest that in both groups VSTM was a better predictor of grammar than HRL.

3.3.3. *Reading*

For adolescents with ID, reading skills were not predicted by HRL (β = .19, 95% CI [−.09, .46], $p = .18$), but they were marginally significantly predicted by VSTM (β = .27, 95% CI [.00, .54], $p = .05$). For children with TD, HRL (β = .36, 95% CI [.06, .66], $p = .02$) was a significant predictor of reading, but VSTM (β = .14, 95% CI [−.15, .44], $p = .33$) was not. Consequently, for the ID group reading was not well predicted by HRL and VSTM, whereas, for the TD group, HRL was a significant predictor variable.

Given the large confidence intervals around the regression coefficients for HRL in both samples and for the various outcome measures, and the exploratory nature of these regressions, results must be interpreted with caution.

4. Discussion

This investigation provided new information about the reliability of measures of HRL, and the predictive relations of HRL to vocabulary, grammar, and reading (decoding) in adolescents with ID. The HRL task showed moderate test-retest and (related) split-half reliability. Learning within the verbal and visual HRL tasks appeared to be similar, but there were important issues around scoring HRL. Partly in line with predictions, correlations between HRL and measures of language and reading were positive in both study groups, however, they were more limited than expected: in the ID group, a significant correlation between HRL and vocabulary emerged; in the TD group, significant correlations were present between HRL and both grammar and reading. Once mental age was controlled in our pre-registered regressions, no relationships between HRL and any language or reading measure emerged in either group. Exploratory regressions that did not control for mental age indicated that HRL predicted vocabulary in the ID group and reading in the TD group (even after controlling for VSTM), but these findings should be treated with caution.

In relation to reliability, the initial analyses were conducted on HRL residuals using combined data from verbal and visual materials. Two methods of HRL scoring were used, ordinal and Levenshtein. The test-retest reliabilities between Session 1 and 2 using ordinal and Levenshtein scoring respectively were .39 and .43. These correlations, when adjusted using the Spearman-Brown formula, gave higher split-half reliabilities (.56 and .60). Although these are moderate values, they are equivalent or higher than previous reports in children (West et al., 2018); and exceed values reported for adults (Bogaerts et al., 2018). It should be noted that more traditional and less complex scoring or analysis approaches all led to poor reliability correlations. For example, test-retest reliabilities for the differences in improvements from the first to the second half of the Hebb experiment for repeated Hebb versus non-repeated filler trials were very low ($r_{tt} = .03$ to $r_{tt} = .10$ n.s.; (for details please see supplementary materials, Section 3, Table S4). Similarly, test-retest and cross-modal correlations were low for gradient scores for each individual based on linear models ($r_{tt} = -.15$ to $r_{tt} = .16$ n.s.) and also based on generalized linear models ($r_{tt} = -.06$ to $r_{tt-} = .29$**; see supplementary materials, Section 3, Tables S4 and S5). This suggests that our method of analysis with GLMMs may have provided a better summary of HRL by taking into account more aspects of performance. However, further research is required to obtain better reliability in HRL tasks. In terms of different HRL scoring methods, correlations were low and non-significant between ordinal scoring of the combined visual and verbal stimuli calculated for Session 1 and for session 2 and corresponding Levenshtein residuals. Thus, despite there being some test-retest reliability, the two scoring systems were not significantly correlated with each other for combined stimuli, suggesting that one or both were not accurate and valid methods of assessing HRL.

To explore these findings, analyses were carried out on the relationships between the combined ordinal HRL residuals and the eight sets of individual scores used to calculate the combined residuals (two materials × two sessions × two scoring systems; Tables 2 and 3). The *combined* ordinal HRL residual scores for Sessions 1 and for Session 2 were generally not related to the *verbal* ordinal scores for Sessions 1 and 2, and were significantly negatively related to the *visual* ordinal scores for sessions 1 and 2, even though a positive relationship would have been expected. A similar pattern of correlations occurred between these two combined ordinal scores and the individual Levenshtein scores. In contrast, all the correlations which did not involve the *combined ordinal scores* were positive and significant, irrespective of whether the correlations involved: Sessions 1 or 2; visual or verbal materials; ordinal or Levenshtein scores.

At first sight, the findings suggest that combined ordinal scores might not always produce valid assessments of HRL. However, ordinal scoring was not inherently problematic, because there were significant correlations between all four ordinal measures involving visual and verbal measures in the same and different sessions (and between these four ordinal scores and the corresponding Levenshtein scores). Importantly, the *combined* ordinal HRL scores produced models with the most complex random effects structure. While Levenshtein models needed only one random intercept to capture individual differences in how well participants remembered the sequences of items, ordinal models needed two intercepts (two separation parameters) to capture individual differences for item memory and items in correct position memory. This difference resulted in the combined ordinal model being more complex than the other models because of the amplification caused by the additional parameter in relation to a relatively large number of other effects.[1]

These findings suggest caution when working with HRL residuals from ordinal scoring if the model to derive those residuals has a complex random effects structure. Furthermore, computing HRL scores based on Levenshtein scoring is recommended because: (1) Levenshtein scores give partial credit for remembering items with minor errors in item position (Kalm & Norris, 2016) and at the trial level, ordinal and Levenshtein scores were highly correlated ($r = .94$; Henry et al., 2022); (2) ordinal GLMMs are less often used in Psychology than standard binomial GLMMs, so analyses of Levenshtein scores with standard GLMMs are more accessible; and (3) the less complex models on Levenshtein scores consistently produced correlated HRL scores for different slices of the data, thus were more reliable and valid.

The second set of analyses concerned whether HRL was related to language and reading in adolescents with ID and children with TD of similar mental age. The pre-registered analyses, controlling for mental age as the first step in all regression models, showed that neither HRL nor VSTM were significant predictors of language or reading in either group. Mental age was a good predictor of vocabulary, grammar and reading, in all but one model, that this may have obscured more subtle predictive relationships involving HRL and VSTM. Mental age is based on both verbal and non-verbal reasoning abilities, and it is likely to contain or be related to abilities which form the basis of HRL and VSTM. Exploratory regressions which excluded mental age revealed: (1) modest relationships between HRL and vocabulary in adolescents with ID; and (2) modest relationships between HRL and reading in children with TD. These findings suggest that HRL may play a small but important role in vocabulary acquisition for adolescents with ID. They also support previous research that HRL is related to reading acquisition in primary school-age children with TD (Attout, Ordonez Magro et al., 2020; Bogaerts et al., 2016; Smalle et al., 2018). Importantly, although the reported relationships are exploratory, they are based on measures of HRL that are likely to be more reliable than those used in previous research. Finally, although HRL also showed a modest relationship with both grammar and vocabulary in the TD group when it was the only predictor, once VSTM was included in the models, HRL was no longer significant.

One interesting feature of the exploratory findings was that both HRL and VSTM contributed to the prediction of vocabulary in the ID group. The HRL findings support previous arguments (e.g., Page & Norris, 2009) that difficulty with serial-order HRL could negatively affect the long-term acquisition of phonological sequences, which are, in turn, important for acquiring new vocabulary. Further, the results suggest that HRL and passive serial order short-term verbal storage capacity were not entirely overlapping skills (Mosse & Jarrold, 2008; Yanaoka et al., 2019), at least in this group. Although HRL related to vocabulary in children with TD when it was the only predictor, supporting previous findings (Archibald & Joanisse, 2013; Mosse & Jarrold, 2008; Smalle et al., 2018), once VSTM was included in the model, this relationship was no longer significant. This suggests that for young school-age children with TD, HRL is less important for vocabulary acquisition than VSTM (Avons et al., 1998; Gathercole et al., 1992, 1997; Michas & Henry, 1994). However, it is important to note that in the TD group, HRL and VSTM showed a moderate correlation ($r = .40$; suggesting some degree of overlap, and supporting previous findings, Mosse & Jarrold, 2008, $r = .35$), whereas no such correlation emerged in the ID group ($r = .11$). There are key differences between HRL and VSTM, so

---

[1] All models needed a random effect for the interaction between list type (Hebb *vs* Filler) and trial position to model individual differences in HRL, and additionally a random slope for trial position (in case this improved model fit) to disentangle individual differences in HRL from individual differences in reacting to trial position due to fatigue or interference effects. If models were based on verbal and visual materials, then combined models needed an additional random effect for material to capture individual differences in better memory for verbal versus visual material. Therefore, setting up models with ordinal scoring and combining both materials leads to a complex random effects structure with 5 random effects variances and 10 covariances between the random effects. The material-specific ordinal models and the combined Levenshtein model were still quite complex as they had 4 random effects and 6 covariances, but like the material-specific Levenshtein models with an even smaller random effect structure, these models produced HRL scores with convergent validity.

relationships are not necessarily expected. For example, although both tasks involved serial order recall, VSTM required the participant to generate and output the full word list verbally, whereas HRL tasks presented the full item set at recall and required nonverbal serial order reconstruction of list items. Furthermore, HRL is a measure of the extent of serial order learning over repeated trials of the same list, whereas VSTM assesses how well individuals can recall individual lists that are different on every trial. It is also possible that methodological differences between studies of HRL in children could be important. Some studies, like the present one, titrate HRL list lengths to span level, usually adding at least one item to span level (e.g., Archibald & Joanisse, 2013; Hsu & Bishop, 2014; West et al., 2018), whereas, more commonly, fixed list lengths for all participants are used, which could lead to floor and ceiling effects (e.g., Attout, Grégoire, et al., 2020; Bogaerts et al., 2018; Mosse & Jarrold, 2008, 2019; Smalle et al., 2016, 2018; Yanaoka et al., 2019). The correlation between HRL and VSTM in the TD group, therefore, could have reflected less successful titration.

One factor that may play a part in HRL significantly predicting vocabulary only in the ID group, is that vocabulary acquisition may have been extended over a longer period, with more repetitions of vocabulary items, allowing those with better HRL to benefit more from this extended experience. The children with TD were between 5 and 10 years old, so vocabulary acquisition was likely to occur with fast mapping (Dockrell & Messer, 2004) and need limited exposure to new items, which might explain why HRL was not a good predictor of vocabulary for the TD group. By contrast, HRL did predict reading for the children with TD, even after VSTM had been included, suggesting that the ability to form long-term memory representations for sequential information could be important for word and nonword decoding in primary school-age children. Interestingly, once HRL had been entered as a predictor, VSTM was not a significant predictor of reading. This suggests that for early and developing readers with TD (Hulme & Snowling, 2009), HRL involving repeated regularities in text and speech could be more important than VSTM abilities.

VSTM was, as expected, a significant predictor of some language and reading skills even when HRL had been controlled: these included vocabulary and grammar in the ID group; and grammar in the TD group. The relationships between VSTM and grammar represented the only consistent finding across the two groups, suggesting that at the cognitive developmental level of our current samples, grammatical skills were related to the ability to hold verbal information passively in short-term memory for brief periods. Although HRL is often lower in younger children (Yanaoka et al., 2019), and can be lower in groups with neurodevelopmental differences (Hsu & Bishop, 2014), little is known about the relationship between HRL and grammar; and previous work has failed to find relationships between HRL and grammatical abilities (Hsu & Bishop, 2014). Furthermore, it may be that in the primary school years, learning of grammar is less likely to involve hearing new grammatical forms repeated in speech within a short space of time, and may depend more on the ability to retain information about the relationship between and positioning of words in complex multi-word sentences.

Overall, there were fewer correlations between the memory, language, and reading variables in the ID group than in the TD group. Given the large age range in the TD group (5–10 years), significant correlations between cognitive variables are expected; but the fewer significant correlations in the ID group could suggest a lower level of coherence between cognitive abilities. However, in the ID group, four of the six non-significant correlations involved reading abilities. Reading abilities are likely to be an important focus of teaching but given the variability in support for students with ID in the UK, some students may receive more support than others, randomly increasing variability in reading and weakening correlations between reading and other abilities. In contrast, children with TD were more likely to receive similar reading support given recent standardization of phonics approaches to teaching in England (Department for Education, 2023).

There are some limitations to this research that should be acknowledged. The choice to use auditory and visual presentation in the verbal HRL was motivated by wanting to have the same picture pointing response method for the verbal and visual HRL tasks. However, this may have led to incidental matching of auditory labels and pictures, perhaps introducing the possibility of semantic processing strategies. Such strategies could have reduced the size of the HRL effect. Further, as participants needed to select images to reconstruct serial order at recall, this may have introduced additional sensorimotor and spatial components to the task. We attempted to reduce the impact of this by randomizing the order of items as they appeared in the response sets across trials. However, visual search and selection processes may still have been involved in the HRL tasks, leading to a less direct and pure measure of HRL. Finally, although we made every effort to titrate HRL task difficulty to span levels, we only had two HRL tasks, so this was not as precise as it could have been; also, the standardized measures available to measure verbal and visual short-term memory were structured differently, leading to possible differences in scoring comparability.

5. Conclusion

This research has provided new and valuable insights into the role of Hebb repetition learning (HRL) in adolescents with intellectual disabilities and its association with language and reading skills. The current findings suggested that several procedures can increase the reliability of HRL assessments, which should be of help when assessing this important neurodevelopmental process, however, further work is required in this area. Both ordinal and Levenshtein scoring methods were found to be correlated, but Levenshtein scoring may be preferable for more complex analyses. HRL was found to predict language and reading skills in both ID and the TD groups. However, the influence of mental age on these relationships cannot be overlooked, as its adjustment meant that HRL was no longer a significant predictor. The findings offer promising avenues for targeted educational interventions, and further research is needed to better understand these complex neurodevelopmental relationships.

Funding

This work was supported by The Baily Thomas Charitable Fund [grant reference number 4768-7692]. The funder had no role in study design; collection, analysis and interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

Data statement

The stimuli, data, models, outputs, scripts and tasks for this study are available on the Open Science Framework https://osf.io/9g5w2/ and on the Gorilla Open Materials page https://app.gorilla.sc/openmaterials/782819.

Open practices

The study in this article has earned Open Data, Open Materials and Preregistered badges for transparent practices. The data, materials and preregistered studies are available at: https://osf.io/gkpwh/.

CRediT authorship contribution statement

Lucy A. Henry: Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **David J. Messer:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization. **Sebastian Poloczek:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Rachel Dennan:** Project administration, Methodology, Investigation, Data curation. **Elisa Mattiauda:** Project administration, Methodology, Investigation, Data curation. **Henrik Danielsson:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

None.

Acknowledgements

We would like to extend our warm gratitude to the children, young people and families who took part in this study, and to the participating schools. We would also like to extend our deep appreciation for the support and input from Mencap and their Research Champions in preparing our study materials and at all stages of this project.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cortex.2024.05.012.

REFERENCES

American Psychiatric Association. (2013). Diagnostic and Statistical Manual of Mental Disorders (5th edition). Arlington, VA: American Psychiatric Association.

Archibald, L. M. D., & Joanisse, M. F. (2013). Domain-specific and domain-general constraints on word learning. *Memory & Cognition, 41*, 268–280. https://doi.org/10.3758/s13421-012-0259-4

Attout, L., Grégoire, C., & Majerus, S. (2020). How robust is the link between working memory for serial order and lexical skills in children? *Cognitive Development,* 53, 100854. https://doi.org/10.1016/j.cogdev.2020.100854

Attout, L., Ordonez Magro, L., Szmalec, A., & Majerus, S. (2020). The developmental neural substrates of Hebb repetition learning and their link with reading ability. *Human Brain Mapping, 41*(14), 3956–3969. https://doi.org/10.1002/hbm.25099. https://onlinelibrary.wiley.com/doi/full/10.1002/hbm.25099

Avons, S., Wragg, C., Cupples, L., & Lovegrove, W. J. (1998). Measures of phonological short-term memory and their relationship to vocabulary development. *Applied Psycholinguistics, 19*, 583–601. https://doi.org/10.1017/S0142716400010377

Bishop, D. V. M. (1997). Cognitive neuropsychology and developmental disorders: Uncomfortable bedfellows. *The Quarterly Journal of Experimental Psychology Section A,* 50(4), 899–923. https://doi.org/10.1080/713755740

Bishop, D. V. M. (2003). *Test for reception of grammar* (2nd ed.). London: Harcourt Assessment.

Bogaerts, L., Siegelman, N., Ben-Porat, T., & Frost, R. (2018). Is the Hebb repetition task a reliable measure of individual difference in sequence learning? *Quarterly Journal of Experimental Psychology, 71*(4), 892–905.

Bogaerts, L., Szmalec, A., De Maeyer, M., Page, M. P. A., & Duyck, W. (2016). The involvement of long-term serial-order memory in reading development: A longitudinal study. *Journal of Experimental Child Psychology, 145*, 139–156. https://doi.org/10.1016/j.jecp.2015.12.008

Bogaerts, L., Szmalec, A., Hachmann, W. M., Page, M. P. A., & Duyck, W. (2015). Linking memory and language: Evidence for a serial-order learning impairment in dyslexia. *Research in Developmental Disabilities, 43*–44, 106–122. https://doi.org/10.1016/j.ridd.2015.06.012

Burack, J. A., Evans, D. W., Russo, N., Napolean, J.-S., Goldman, K. J., & Iarocci, G. (2021). Developmental perspectives on the study of persons with intellectual disability. *Annual Review of Clinical Psychology, 17*, 13.1–13.25. https://doi.org/10.1146/annurev-clinpsy-081219-090532.

Charlton, C., Rasbash, J., Browne, W. J., Healy, M., & Cameron, B. (2018). *MLwiN version 3.02*. Centre for Multilevel Modelling, University of Bristol.

Couture, M., & Tremblay, S. (2006). Exploring the characteristics of the visuospatial Hebb repetition effect. *Memory & Cognition, 34*(8), 1720–1729.

Cunningham, A. J., Burgess, A. P., Witton, C., Talcott, J. B., & Shapiro, L. R. (2021). Dynamic relationships between phonological memory and reading: A five-year longitudinal study from age 4 to 9. *Developmental Science, 24*(1), e12986. https://onlinelibrary.wiley.com/doi/full/10.1111/desc.12986.

Department for Education. (2023). Guidance: Choosing a phonics teaching programme. https://www.gov.uk/government/publications/choosing-a-phonics-teaching-programme/list-of-phonics-teaching-programmes.

Dockrell, J., & Messer, D. (2004). Lexical acquisition in the school years. In R. Berman (Ed.), *Language development: Psycholinguistic and typological perspectives*. New York: John Benjamins.

Dunn, L. M., Dunn, D. M., & Styles, B. (2009). *British Picture Vocabulary Scale* (3rd ed.). London, UK: GL Assessment.

Gathercole, S. E., Hitch, G. J., Service, E., & Martin, A. (1997). Phonological short-term memory and new word learning in children. *Developmental Psychology, 33*, 966–979. https://doi.org/10.1037/0012-1649.33.6.966

Gathercole, S. E., Willis, C. S., Emslie, H., & Baddeley, A. D. (1992). Phonological memory and vocabulary development during the early school years: A longitudinal study. *Developmental Psychology, 28*, 887–898. https://doi.org/10.1037/0012-1649.28.5.887

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research, 26*(3), 499–510. https://doi.org/10.1207/s15327906mbr2603_7

Hebb, D. O. (1949). *The organization of behavior; a neuropsychological theory*. Wiley.

Hebb, D. O. (1961). Distinctive Features of Learning in the Higher Animal. In J. F. Delafresnaye (Ed.), Brain mechanisms and learning (pp. 37–46). Oxford: Blackwell.

Henry, L. A., Poloczek, S., Messer, D. J., Dennan, R., Mattiauda, E., & Danielsson, H. (2022). Hebb repetition learning in adolescents with intellectual disabilities. *Research in Developmental Disabilities, 125*(June 2022), 104219. https://doi.org/10.1016/j.ridd.2022.104219

Hsu, J. H., & Bishop, D. V. M. (2014). Sequence-specific procedural learning deficits in children with specific language impairment. *Developmental Science, 17*(3), 352–365.

Hulme, C., & Snowling, M. J. (2009). *Developmental disorders of language learning and cognition*. Wiley Blackwell.

Johnson, A. J., Dygacz, A., & Miles, C. (2017). Hebb repetition effects for non-verbal visual sequences: Determinants of sequence acquisition. *Memory, 25*(9), 1279–1293. https://doi.org/10.1080/09658211.2017.1293692

Kalm, K., & Norris, D. (2016). Recall is not necessary for verbal sequence learning. *Memory & Cognition, 44*, 104–113. https://doi.org/10.3758/s13421-015-0544-0

Kuppuraj, S., Rao, P., & Bishop, D. V. (2016). Declarative capacity does not trade-off with procedural capacity in children with specific language impairment. *Autism & Developmental Language Impairments, 1*, 239694151667441. https://doi.org/10.1177/2396941516674416

Leclercq, A.-L., & Majerus, S. (2010). Serial-order short-term memory predicts vocabulary development: Evidence from a longitudinal study. *Developmental Psychology, 46*(2), 417–427. https://doi.org/10.1037/a0018540

Lemons, C. J., Zigmond, N., Kloo, A. M., Hill, D. R., Mrachko, A. A., Paterra, M. F., Bost, T. J., & Davis, S. M. (2013). Performance of students with significant cognitive disabilities on early-grade curriculum-based measures of word and passage reading fluency. *Exceptional Children, 79*(4), 408–426. https://doi.org/10.1177/001440291307900402

Loo, C., Lee, A. C. H., & Buchsbaum, B. R. (2021). Multivariate FMRI Signatures of Learning in a Hebb Repetition Paradigm With Tone Sequences. *Frontiers in Neurology, 12*, 674275. https://doi.org/10.3389/fneur.2021.674275

Lum, J. A. G., Conti-Ramsden, G., Morgan, A. T., & Ullman, M. T. (2014). Procedural learning deficits in specific language impairment (SLI): A meta-analysis of serial reaction time task performance. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior, 51*, 1–10. https://doi.org/10.1016/j.cortex.2013.10.011

Lum, J. A. G., Conti-Ramsden, G., Page, D., & Ullman, M. T. (2012). Working, declarative and procedural memory in specific language impairment. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior, 48*(9), 1138–1154. https://doi.org/10.1016/j.cortex.2011.06.001

Majerus, S., & Boukebza, C. (2013). Short-term memory for serial order supports vocabulary development: New evidence from a novel word learning paradigm. *Journal of Experimental Child Psychology, 116*, 811–828. https://doi.org/10.1016/j.jecp.2013.07.014

Majerus, S., Leclercq, A.-L., Grossmann, A., Billard, C., Touzin, M., Van der Linden, M., & Poncelet, M. (2009). Serial order short-term memory capacities and specific language impairment: No evidence for a causal association. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior, 45*, 708–720. https://doi.org/10.1016/j.cortex.2008.10.006

McKenzie, K., Milton, M., Smith, G., & Ouellette-Kuntz, H. (2016). Systematic review of the prevalence and incidence of intellectual disabilities: Current trends and issues. *Current Developmental Disorders Reports, 3*, 104–115. https://doi.org/10.1007/s40474-016-0085-7

Michas, I., & Henry, L. A. (1994). The link between phonological memory and vocabulary acquisition. *British Journal of Developmental Psychology, 12*, 147–163.

Moraleda-Sepúlveda, E., & López-Resa, P. (2022). Morphological difficulties in people with developmental language disorder. *Children, 9*(2), 125. https://doi.org/10.3390/children9020125

Mosse, E. K., & Jarrold, C. (2008). Hebb learning, VSTM, and the acquisition of phonological forms in children. *Quarterly Journal of Experimental Psychology, 61*, 505–514.

Mosse, E. K., & Jarrold, C. (2010). Searching for the Hebb effect in Down syndrome: Evidence for a dissociation between VSTM and domain-general learning of serial order. *Journal of Intellectual Disability Research, 54*(4), 295–307. https://doi.org/10.1111/j.1365-2788.2010.01257.x

Musfeld, P., Souza, A. S., & Oberauer, K. (2023). Repetition learning is neither a continuous nor an implicit process. *Psychological and Cognitive Sciences, 20*(16), e2218042120. https://doi.org/10.1073/pnas.2218042120.

Nilsson, K., Danielsson, H., Elwér, Å., Messer, D., Henry, L., & Samuelsson, S. (2021). Decoding abilities in adolescents with intellectual disabilities: The contribution of cognition, language, and home literacy. *Journal of Cognition, 4*(1), 58. https://doi.org/10.5334/joc.191, 1–16.

Nilsson, K., Danielsson, H., Elwér, A., Messer, D., Henry, L., & Samuelsson, S. (2021). Investigating reading comprehension in adolescents with intellectual disabilities: Evaluating the simple view of reading. *Journal of Cognition, 4*(1), 56. https://doi.org/10.5334/joc.188

Norris, D. (2017). Short-term memory and long-term memory are still different. *Psychological Bulletin, 143*(9), 992–1009. https://doi.org/10.1037/bul0000108.

Norris, D., Page, M. P. A., & Hall, J. (2018). Learning nonwords: The Hebb repetition effect as a model of word learning. *Memory, 26*(6), 852–857. https://doi.org/10.1080/09658211.2017.1416639

Page, M. P. A., & Norris, D. (2008). Is there a common mechanism underlying word-form learning and the Hebb repetition effect? Experimental data and a modelling framework. In A. Thorn, & M. P. A. Page (Eds.), *Interactions between short-term and long-term memory in the verbal domain* (pp. 136–155). Hove: Psychology Press.

Page, M. P. A., & Norris, D. (2009). A model linking immediate serial recall, the Hebb repetition effect and the learning of phonological word forms. *Philosophical Transactions of the Royal Society of London, 364*, 3737–3753. https://doi.org/10.1098/rstb.2009.0173

Patel, D. R., Cabral, M. D., Ho, A., & Merrick, J. (2020). A clinical primer on intellectual disability. *Translational Pediatrics, 9*(Suppl 1), S23–S35. https://doi.org/10.21037/tp.2020.02.02

Pickering, S., & Gathercole, S. E. (2001). *Working Memory Test Battery for Children*. London, UK: Pearson Assessment.

Ratz, C., & Lenhard, W. (2013). Reading skills among students with intellectual disabilities. *Research in Developmental Disabilities, 34*(5), 1740–1748. https://doi.org/10.1016/j.ridd.2013.01.021

Reynolds, C., & Voress, J. K. (2007). *Test of memory and learning* (2nd ed.). Austin, TX: Pro-Ed.

Rice, M. L., & Wexler, K. (1996). A phenotype of specific language impairment: Extended optional infinitives. In M. L. Rice (Ed.), *Toward a genetics of language* (pp. 215–237). Mahwah, NJ: Lawrence Erlbaum.

Roid, G. H. (2003). *Stanford-Binet Intelligence Scales* (5th ed.). Rolling Meadows, IL: Riverside Publishing.

Siegelman, N., Bogaerts, L., & Frost, R. (2017). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavior Research Methods, 49*, 418–432. https://doi.org/10.3758/s13428-016-0719-z

Smalle, E. H. M., Bogaerts, L., Simonis, M., Duyck, W., Page, M. P. A., Edwards, M., & Szmalec, A. (2016). Can chunk size differences explain developmental changes in lexical learning? *Frontiers in Psychology, 6*, 1925. https://doi.org/10.3389/fpsyg.2015.01925

Smalle, E. H. M., Page, M. P. A., Duyck, W., Edwards, M., & Szmalec, A. (2018). Children retain implicitly learned phonological sequences better than adults: A longitudinal study. *Developmental Science, 21*(5). https://doi.org/10.1111/desc.12634

Sparrow, S. S., Cicchetti, D. V., & Saulnier, C. A. (2016). *Vineland Adaptive Behavior Scales* (3rd ed.). Bloomington, MN: Pearson.

Sridhar, S., Khamaj, A., & Asthana, M. K. (2023). Cognitive neuroscience perspective on memory: overview and summary. *Frontiers in Human Neuroscience, 17*, 1217093. https://www.frontiersin.org/articles/10.3389/fnhum.2023.1217093.

Szmalec, A., Loncke, M., Page, M. P. A., & Duyck, W. (2011). Order or disorder? Impaired Hebb learning in dyslexia. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(5), 1270–1279. https://psycnet.apa.org/doi/10.1037/a0023820.

Tabachnick, B. G., & Fidell, L. S. (2014). *Using multivariate statistics* (6th ed.). Harlow, Essex: Pearson.

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (2012). *Test of Word Reading Efficiency* (2nd ed.). New York: Pearson.

Ullman, M. T., & Pierpont, E. I. (2005). Specific language impairment is not specific to language: The procedural deficit hypothesis. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior, 41*(3), 399–433. https://doi.org/10.1016/S0010-9452(08)70276-4

van der Schuit, M., Segers, E., van Balkom, H., & Verhoeven, L. (2011). How cognitive factors affect language development in children with intellectual disabilities. *Research in Developmental Disabilities, 32*(5), 1884–1894. https://doi.org/10.1016/j.ridd.2011.03.015

Vulchanova, M., Foyn, C. H., Nilsen, R. A., & Sigmundsson, H. (2014). Links between phonological memory, first language competence and second language competence in 10-year-old children. *Learning and Individual Differences, 35*, 87–95. https://doi.org/10.1016/j.lindif.2014.07.016

West, G., Vadillo, M. A., Shanks, D. R., & Hulme, C. (2018). The procedural learning deficit hypothesis of language learning disorders: We see some problems. *Developmental Science, 21*(2), e12552. https://doi.org/10.1111/desc.12552

Yanaoka, K., Nakayama, M., Jarrold, C., & Saito, S. (2019). Determining the developmental requirements for Hebb repetition learning in young children: Grouping, short-term memory, and their interaction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(4), 573–590. https://doi.org/10.1037/xlm0000606

Zhang, Z., Parker, R. M. A., Charlton, C. M. J., Leckie, G., & Browne, W. J. (2016). R2MLwiN: A package to run MLwiN from within R. *Journal of Statistical Software, 72*(10), 1–43. https://doi.org/10.18637/jss.v072.i10