ResearchSPAce

TITLE: Children with typical development or Down syndrome benefit from testing vs. restudy of arbitrary event sequences across long-term delays and in the face of sleep disturbance

Authors and affiliations:

Hughes, KMO, Bath Spa University, Bath *corresponding author

Sakhon, S, University of Arizona, Tucson

Reichsfeld, A, University of Arizona, Tucson

Luongo, A, University of Arizona, Tucson

Barness, B, University of Arizona , Tucson

Bottrill, K, University of Arizona , Tucson

Lee, NR, Drexel University, Philadelphia

Abbeduto, L, University of California Davis, Davis

Thurman, AJ, University of California Davis, Davis

Edgin, JO, Virginia Tech University, Blacksburg

Corresponding Author: KMO Hughes

Abstract:

Improved memory after prompted retrieval, also known as the testing effect, is evidenced in adults to support long term memory, but rarely examined in children in pre-school or with intellectual disabilities, such as Down syndrome (DS). This study examined episodic memory across one-month, manipulating the presentation of episodic information to compare testing versus restudy of arbitrary event sequences, and the effect of sleep variables. Retrieval rates were compared at 5-minute and one-month delays in 52 children with DS (aged 6-18, 27 male, 24 White) compared to 59 children with typical development (aged 3-7, 23 male, 35 White). A single test improved recall in the DS group over long term delays, this is a novel finding and relevant to real-life and classroom experiences.

Children with typical development or Down syndrome benefit from testing vs. repetition of
arbitrary event sequences across long-term delays and in the face of sleep disturbance

Memory and Learning – the effect of testing and restudying

The creation of memories and their storage over time is a key developmental process. Various learning strategies have been studied in adults with typical development (TD), with two of the most used strategies being retrieval practice/testing and restudying. Retrieval practice, referred to throughout as testing, requires the participant to recall and produce information, either immediately after encoding or following a delay; real world examples include exams or class tests. Restudying provides multiple opportunities to encode information; real world examples include flashcard learning. Recent research has suggested that retrieval practice may serve to strengthen and integrate cortically based neural representations, suggesting a structural basis for a functional change (Guran et al., 2022; Zhuang et al., 2022).

Compared to restudying, retrieval results in significantly better long-term recall of the information in individuals with TD, a benefit referred to as the "testing effect" (Bartlett & Tulving, 1974; Carrier & Pashler, 1992). Roediger and Karpicke (2006) found that participants who studied a prose passage and were tested after, opposed to restudying the passage for a second time, recalled more prose after a two-day delay. This benefit was also present when participants were assessed again after a 1-week interval. Moreover, testing effects across the long-term seem to be present even when the initial recollection is unsuccessful, and when no explicit feedback is provided (Smith et al., 2016).

Although less is known about the testing effect in young children, there have been a few studies that have found a benefit of testing when preschool-age children learn novel object labels (Fritz et al., 2007), and when 9-year-olds learn novel words and definitions

(Goossens et al., 2014). Little is known about the testing effect, however, in individuals with Down syndrome (DS), caused by a third copy of some or all of chromosome 21. As individuals with DS have uneven memory development across the lifespan (Hughes, KMO, 2018; Silverman, 2007), understanding memory strengths and weaknesses, and methods to improve memory function in this population is the focus of the present study.

Memory and sleep – hippocampal consolidation pathways

Despite increasing awareness in the field of the importance of sleep in both health and disease, the influence of sleep on the testing effect has not been adequately investigated. One study by Bauml, Holterman, and Abel (2014) found for adult participants who either had testing or restudy prior to a period of sleep vs wake, sleep improved the performance in the restudy condition, so that memory performance for both conditions was comparable after a night of sleep and significantly better than the restudy wake group (but see Racsmany, Conway, & Demeter, 2010; Tucker & Fishbein, 2008). Interestingly, this was caused by the restudy condition leading to better performance after sleep, rather than the testing group performing less well, so that memory performance for both conditions was comparable after a night of sleep and significantly better than the restudy wake group. This finding suggests that the benefit of sleep overpowers the benefit of testing and is consistent with the sleep-dependent learning literature that also demonstrates a memory benefit of sleep vs wake (Ellenbogen, et al., 2006; Gais, Lucas, & Born, 2006; Gomez, Bootzin, & Nadel, 2006; Wilhelm, Diekelmann, & Born, 2008).

Further evidence that sleep processes are involved in the consolidation of memory comes from studies showing that information learned during waking is reactivated during sleep as a form of internal retrieval practice (Born & Wilhelm, 2012). Both retrieval and sleep independently reduce forgetting over time, as information becomes consolidated into

long-term storage and less dependent on the hippocampus. The results from the aforementioned study by Bauml, Holterman, and Abel (2014) demonstrate how testing and retrieval practice could allow information to be consolidated into long-term storage immediately, bypassing the need for consolidation during sleep. This also suggests that the consolidation method associated with testing may not require the hippocampus, as sleep consolidation appears to involve hippocampus-cortex-hippocampus loops and reactivation patterns.

If consolidation via retrieval practice is less reliant on the hippocampus, this route of learning may be particularly beneficial for individuals with DS. Given the impaired hippocampal function in this population, it is unclear whether the same patterns of performance in the testing effect and restudying would occur, although recent work has suggested that retrieval practice also facilitates learning in adults with DS (Starling, Moreira, & Jaeger, 2019).

The development of episodic memory

Episodic memory is the autobiographical memory of sequences of events. One method used to measure episodic memory is the deferred imitation (DI) task, which requires participants to reproduce action sequences in the same temporal order as the sequences were observed immediately and/or after a delay. The DI task is considered a gold standard task for episodic memory measurement over long term delays (Milojevich & Lukowski, 2016).

Typical development

Evidence for memory function derived from DI assessments has been seen as young as 6 months of age, with infants with TD able to recall a single action over a delay of 24 hours (Barr, Dowden & Hayne, 1996; Collie & Hayne, 1999). By the age of 29 months, children with TD can recall multiple actions in temporal order for delays as long as 8 months

(Bauer, Hertsgaard, & Dow, 1994). Some studies have compared the effect of restudying in individuals with TD across early childhood. For example, infants aged 16 to 20 months recall over a 3-month delay significantly more information from sequences modelled two or three times compared to a single exposure, showing the benefit of repetition (Bauer & Levonton, 2013).

The development of language over the first few years of life results in a vast increase in the possible memory assessments available to researchers, which is likely why most research using DI has been conducted in verbal children. However, the importance of these findings is limited by the lack of longitudinal or cross-sectional studies into late childhood and adolescence to provide long-term correlations in performance.

Down syndrome

The few studies using DI show that individuals with DS between the ages of 20 to 43 months who were demonstrated a single target action with an object, show retention of this action over a delay of 5 minutes (Rast & Meltzoff, 1995). Children with DS aged 22 to 49 months demonstrated better recall of actions in a three-step sequence if tested compared to no test, over a 1-month delay, supporting the presence of the testing effect in the DS population by this age (Milojevich & Lukowski, 2016). After a one month delay all participants were then presented with the same items from the learned sequences, as well as novel items from previously unseen sequences. However, at the 1-month delay, the DS group in that study did not recall more pairs of actions on familiar sequences relative to novel sequences, suggesting that either episodic memory loss occurred, or that the sequences lack a significant degree of arbitrariness.

Individuals with DS aged 3 to 5 years have been found to recall significantly more actions of a three-step sequence than mental age (MA) matched controls with TD over a 24-hour delay (Roberts & Richmond, 2015). A limitation of this study is that the sequence was

not truly arbitrary- it was creating a toy rattle where the three actions were: putting a ball in a container, the attaching a handle, shaking the rattle. It is likely that the DS group outperformed as they benefitted from the social cues present in the non-arbitrary event sequence performance. Furthermore, the dependent variable in this study is actions- not pairs in sequence, suggesting it is less episodic memory as it misses the order component. These results provide insights into the initial emergence of episodic memory in the DS population, but its development has yet to be adequately explored.

As people with DS develop, their memory profile diverges from those with TD, often impacting their educational development, therefore understanding of this, and methods to support learning, would be highly valuable. Furthermore, no studies have examined the effect of sleep on hippocampal-dependent learning tasks such as DI in the DS population across childhood and adolescence.

The current study

The current study utilized a DI paradigm to assess the testing effect in individuals with DS, compare repeated exposure vs retrieval memory, and test both action memory and episodic memory for a temporal sequence (Milojevich & Lukowski, 2016).

We addressed the following aims and hypotheses:

1. We aimed to determine whether individuals with DS demonstrate the testing effect. The hypothesis is that individuals with DS will produce more temporally correct pairs (pairs of actions in the correct order) after a 1-month delay for sequences that were immediately tested compared to sequences that were not immediately tested.

2. We aimed to compare recall in tested vs. repeated sequences at one month, hypothesizing that both DS and TD groups will recall more information in tested sequences than repeated sequences.

3. Recall at one month is compared to actions produced in novel sequences, to examine if we are measuring true memory as opposed to chance actions. The hypothesis was that there will be more actions recalled vs produced in novel sequences, demonstrating the arbitrary nature of the sequences and the validity of the recall at one month.

4. To examine how the hippocampus and sleep dependent consolidation influence information processing in both testing vs repetition conditions. The hypothesis was that temporal order performance differences across the two time points in the test condition will not be correlated with sleep measures, and that sequences in the repetition condition will benefit from sleep. Those with DS with worse sleep quality may have poorer long-term performance in the repetition condition (Bauml, Holterman, & Abel, 2014).

Method

Participants

Three university sites participated, and each obtained IRB approval. Fifty-two children and adolescents with DS and fifty-nine children with TD were recruited for the study, with age range for each group selected to facilitate comparable cognition at the group level, see Table 1. Participants were recruited through local and parent organizations and advertisement. The exclusion criteria for both groups included: no history of uncorrected vision or hearing impairments, past head injury or brain trauma, or unstable anxiety disorder symptoms. The presence of trisomy 21 in participants with DS was confirmed by medical records. The DS and TD groups differed significantly on nonverbal raw scores ($t(94)$ =-2.36, $p= .020$) and verbal raw scores on the KBIT-2 ($t(100)=-3.03$, $p=.003$; Table 1). It is important to note that there are multiple methodological and conceptual challenges associated with

matching groups with differing skill levels through post hoc participant selection (Kover & Atwood, 2013; Mervis & Klein-Tasman, 2004). In the present study, inclusion of the sample of children with TD was done to replicate previous findings, thereby establishing the validity of the task and experimental manipulations for use in the participants with DS; thus, the whole sample was included despite the lack of "matched" groups.

Table 1. Demographic characteristics and KBIT-2 performance of final sample.

Group	N (male)	Age in months (Range)	KBIT-2 Verbal Raw mean (Range)	KBIT-2 Non-Verbal Raw mean (Range)
DS	52 (27)	140 (72-215)	22.1 (1-47)	12.8 (1-20)
TD	59 (23)	58 (33-81)	29.2 (4-49)	15.0 (1-32)

Procedure and stimulus materials

Participants in the current study were recruited into a larger study in which they came into the laboratory for multiple sessions, returning one month after session 1.

Delayed imitation task: The paradigm used for this study was adapted from Lukowski & Milojevich (2016) and was adjusted for school-age children with TD and individuals with DS. Each sequence included three target actions (example Figure 1).

Action 1: Velcro the gem onto the side of the block.
Action 2: Balance the plank on the top of the block.
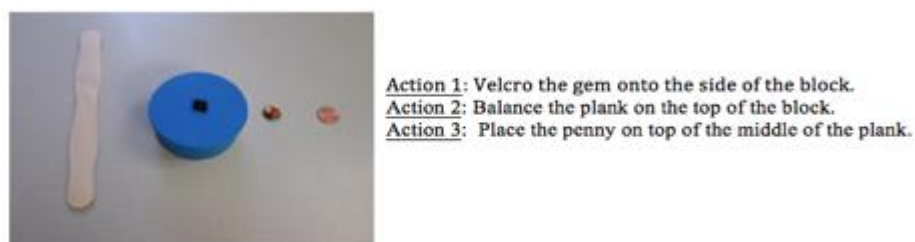Action 3: Place the penny on top of the middle of the plank.

Figure 1. Example sequence.

There were three conditions that differed, to examine how different presentations affect memory encoding of the information.. Each participant saw a total of six sequences, two in each of the three conditions: 1) Model, Model, Model (MMM); 2) Model, Model, 20 second pause (MMP); and 3) Model, Model, Test (MMT). In the MMM condition, the examiner modelled the sequences three times to the participant. In the MMP condition, the examiner modelled the sequences twice to the participant and then removed the objects from the table and sat quietly for 20 seconds to equate the amount of processing time to the other conditions. In the MMT condition, the examiner modelled the three-action sequences twice for the participant and then asked the participant to use the objects to perform all three actions in the same order demonstrated. If the participant imitated the sequence correctly the items were removed, if not they were given 90 seconds to attempt the sequence before the items were removed, with the experimenter offering praise, e.g., "that's a neat idea" at regular intervals. During session 1, all participants began with a two-action sequence warm-up, with the sequence modelled twice and then tested. After the warm-up, the full test was administered, see Figure 2. Once all the original six sequences were completed, a 5-minute gap elapsed where participants were given a filler task of drawing to complete. After these, memory was tested for three sequences, one in each condition. There was then a full month interval, after which participants returned and memory for all sequences were assessed. Participants were also presented with the items for a previously unseen sequence, to allow comparison for recalled vs random actions/pairs. For full description of all sequences see Appendix, Table 3.
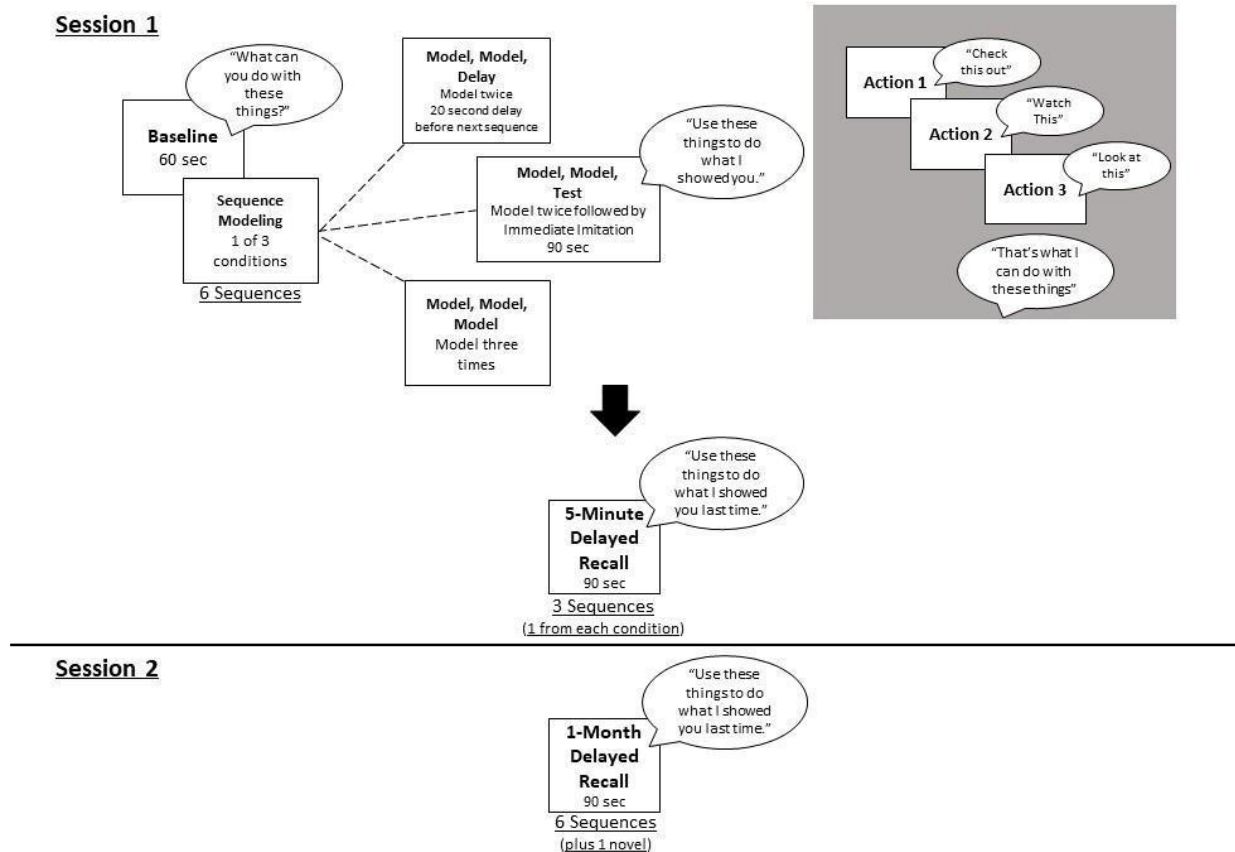
Figure 2. Study design

In terms of scoring actions and action pairs at each time point were coded. Actions were coded for the N of target actions that occurred ranging from 0-3. Only the first instance of each action was assessed when coding for pairs to account for actions produced by change and trial-and-error. Possible pairs included: pair 1: (action 1 followed by action 2) and pair 2:(action 2 followed by action 3) for a maximum of 2 pairs. For example, a participant who performed the following actions in this order 3, 3, 3, 1, 2, would receive 3 action (for completing all the actions) and 1 pair (action 1 followed by action 2). Whereas a participant who performed the actions: 3, 2, 1, 2, would be scored 3 actions, 0 pairs, as only the first instance of each action is included in pair scoring. Mean number of pairs was used to assess memory for temporal order based on Milojevich & Lukowski (2016). Sequences that were temporally correct at baseline were excluded from all calculations. A total of nine sequences

(Appendix Table 3) were counterbalanced between the three sequence modelling conditions, 5-minutes test, and novel sequences at 1-month, into four versions (see Appendix, Table 2).

Kaufman Brief Intelligence Test-Second Edition (KBIT-2): Each participant's verbal and non-verbal intelligence was assessed using the KBIT-2. Raw scores were used to assess between-group differences, due to the psychometric limitations associated with MA scores. Two participants with DS did not complete the KBIT-2 (described below) due to behavioural issues. A further participant completed the verbal but not non-verbal portion.

Actigraphy/ Sleep Diary: Actiwatch-2 devices were used to collect sleep data. Participants wore the watch for seven days prior to or following session 1 and again for session 2. Concurrently, parents also filled out a sleep diary. Based on the recommendations from previous reliability analyses in children (Acebo et al., 1999), all participants wore the Actigraph on the nondominant wrist for a minimum of 5 consecutive days, and the mean sleep scores were computed across this period. Data were collected in 30-s epochs and analysed using commercially available software (Respironics Actiware 5.71.0, Bend, OR). Actigraphy data were scored at the medium sensitivity threshold (activity counts = 40/min), with sleep onset and sleep end marked by a period of 3 and 5 min of immobility or more, respectively (Meltzer, Montgomery-Downs, Insana, & Walsh, 2012). Each epoch of data from the Actiwatch was assessed as sleep or wake, based on whether the activity score exceeded the medium threshold. Variables of interest included sleep efficiency (percentage of time in bed spent asleep) and total sleep duration.

Statistical analysis plan: To examine baseline and immediate imitation, actions and pairs were compared between groups with t-tests for each measure (N actions, N pairs). Hypothesis

1 was addressed by running two (actions and pairs) mixed model ANOVAs with condition (3: MMM, MMT, MMP) and "immediately tested" (2: yes or no) as within group factors and group as a between group factor, with post-hoc follow ups to investigate the effect of condition. Hypothesis 2 was addressed by running further two-way mixed ANOVAS with condition (MMT and MMM) as within group and group as between group. Hypothesis 3 was addressed by comparing the mean number of pairs for each sequence performed at 1 month for novel and familiar sequences within each group by t-test. Hypothesis 4 was addressed by averaging sleep data across time points (first visit and one-month delay), and correlating sleep efficiency and sleep duration with a memory change score, calculated by subtracting memory performance at one month from memory at 5 minutes.

Results

There were no significant differences between groups in actions or pairs at baseline suggesting that groups had similar naturalistic behaviours with the items prior to the task (baseline actions: $t(109)=-0.96$, $p=.34$, baseline pairs: $t(109)=-0.46$, $p=.647$). However, in immediate performances there were group differences and violation of equality of variances due to floor and low performances in the DS group compared to the controls, and thus Levenes values are reported (Immediate actions: $t(82)=-2.08$, $p=.04$; Immediate Pairs: $t(89)=-2.73$, $p=.008$), with the TD group performing significantly better at immediate imitation.

Immediate Testing effect evaluated at 1-month

<u>Actions</u>

At 1 month, actions recalled in sequences that were/were not immediately tested were compared by condition between groups. The main effect of immediate testing was significant ($F(1,170)=43.9$, $p<.001$) but did not interact with group ($p > 0.5$), suggesting the TD and DS

groups benefited comparably from immediate testing, supporting the hypothesis that testing supports long-term memory in DS. The main effect of condition was also significant ($F(2,170)=7.06$, $p=.001$), although condition effects did not interact with group ($p>0.05$). The main effect of condition was driven by significantly better recall in MMT than in either MMP ($t(2)=0.20$, $p=.010$) or MMM ($t(2)=0.27$, $p=.001$). Importantly, there was a significant interaction between condition, group, and whether tested immediately vs not tested immediately ($F(2,170)=4.9$, $p=.010$). This interaction is depicted in Figure 3, which shows that the testing effect occurred in the DS group in the MMP and MMM but not MMT conditions, whereas in the TD group the test effect was present in the MMM condition, but less pronounced in either MMP or MMT.
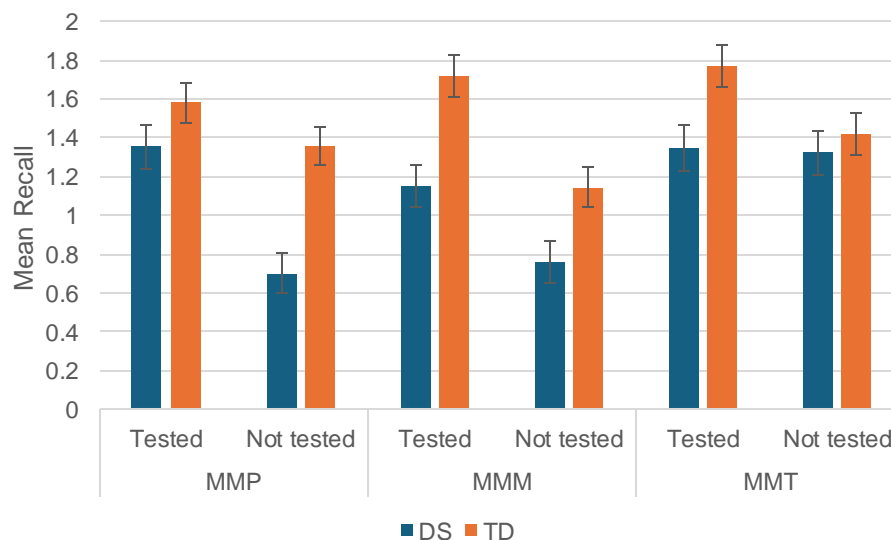

Figure 3. Recall by condition based on whether sequences were immediately tested or not, by group (MMP= pause, MMM= restudy, MMT= test). (Error bars= SEM)

Pairs

At 1 month, pairs recalled in sequences that were/were not immediately tested were compared by condition between groups. Between group results need to be interpreted with caution due to the immediate differences in group performance.

The main effect of testing was significant ($F(1,166)=21.7$, $p<.001$) but did not interact with group ($p>0.5$), meaning the TD and DS groups benefitted comparably from immediate testing. The main effect of condition was also significant ($F(2,166)=5.9$, $p=.003$), although this did not interact with group ($p>0.05$). Importantly, there was not a significant interaction between condition, group, and whether tested immediately vs not tested immediately ($F(2,166)=1.6$, $p=.198$).

Tested vs repeated sequences

Comparing actions at 1-month performance in sequences that were repeated vs tested, the main effect of condition was significant ($F(1,96)=13.8$, $p<.001$), such that tested sequences had higher mean actions recalled than repeated sequences (2.22 vs 1.96). In addition, condition significantly interacted with group ($F(1,96)=3.96$, $p=.049$). This finding was driven by a greater difference between MMM and MMT (MMT better recalled) performance in the DS than TD group, further indicating the importance of testing for individuals with DS for support long term retention.

Comparing pairs 1-month performances in sequences that were repeated vs tested, the main effect of condition was significant ($F(1,96)=12.3$, $p=.001$), such that tested sequences had a higher mean number of pairs recalled than repeated sequences (0.73 vs 0.49). This did not significantly interact with group ($F(1,96)=3.0$, $p=.086$).

Arbitrary pairs vs memory

To address the concerns raised in Milojevich & Lukowski, (2016), we compared pairs in novel sequences to pairs in familiar sequences across conditions at 1-month delay. In both the DS ($t(45)$=-7.4, $p<.001$) and TD groups ($t(51)$=-9.2, $p<.001$), significantly more pairs were performed in the familiar sequences than in novel sequences.

Sleep effects

In the TD group, there were no significant correlations between sleep measures and memory changes across measures and conditions. In the DS group, sleep efficiency was associated with retention of actions ($r(31)$=-0.398, $p=.013$) in the MMM condition, see Appendix, Table 1.

Discussion

Consistent with the first hypothesis, testing benefited long-term memory for the DS group. Whereas there were group differences based on testing and condition in the recall of actions, there was not the same three-way interaction in the comparison of pairs. This suggests that both the TD and DS groups benefit similarly from testing when recalling temporal information. The recall of actions differed between the DS and TD groups based on the condition and if they were tested, whereas this was not seen in pairs, although with a significance value of $p=.086$ it is possible a larger sample size may reveal a different result.

The second hypothesis was that long term episodic memory benefits from testing compared to restudying. Results showed that, at one month, all sequences in the MMT condition (i.e., that had immediate imitation opportunities) were better recalled than those in MMM or MMP. The nonsignificant interaction with group suggests that both individuals with DS and TD benefit comparably from testing; thus, encoding techniques may be more

similar across groups than previously suspected. However, the significant interaction of group with condition with testing suggests that the DS group benefits from single testing at five minutes (MMM, MMP) more than the TD group, suggesting immediate testing is particularly supportive for this group compared to individuals with TD. The lack of group differences in the MMT condition suggests double testing (MMT + five-minute test) does not provide a significant improvement, supporting a single immediate test to improve long term memory outcomes. These findings are consistent with previous studies suggesting restudying benefits short-term retention and testing supports long-term retention (Roediger & Karpicke, 2006; Roediger, Putnam, & Smith, 2011).

Our third hypothesis addressed a limitation of the Milojevich & Lukowski (2016) study, which reported no significant difference in production of pairs in novel and familiar sequences at one month, suggesting that either memory was impaired or that the sequences were not arbitrary. In our data, both groups performed significantly more pairs in familiar than in novel sequences at one month, showing not only that memory was functioning and assessed, but also that we have successfully designed arbitrary sequences.

Our fourth hypothesis addressed the evidence for sleep supporting weaker methods of learning over delays, but not benefitting information that has been tested. Sleep measures were not correlated with performance change across time in the MMT condition but were correlated with the MMM condition only in the DS group indicating that consolidation between study and test for the repeated sequences is related to sleep in the DS group. Bauml, Holterman, and Abel (2014) found sleep supported restudy but not retrieval condition memory, further suggesting that sleep supports weaker forms of memory. This evidence suggests that more low engagement tasks may benefit more from sleep consolidation, specifically in people with DS, potentially due to impaired hippocampal function and therefore reduced capacity for online storage.

Overall, the DS group demonstrated better temporal order memory and memory for individual actions for the MMT condition compared to the other 2 conditions after a 1-month delay. However, the TD group did not show a difference between the three conditions at 1-month. This is inconsistent with previous literature that found a benefit of retrieval practice in young children (Fritz et al., 2007; Goossens et al., 2014; Sheffield & Hudson, 2006). A common criterion associated with retrieval practice benefit is that retrieval must be effortful to benefit (Bjork & Bjork, 2011). The amount of effortful retrieval might be different between the two groups. However, comparing performance on immediate test, which can only be observed in the MMT condition, did not show a difference between the two groups, suggesting imitation abilities are not significantly worse in individuals with DS compared to individuals with TD at a similar cognitive level.

This study has clinical implications for people with DS as well as other intellectual disabilities. First, memory is not uniformly impaired across encoding presentations and time scales in individuals with DS, but they show dramatically improved long-term retention when given the chance to retrieve after learning. This finding suggests that educational strategies should give more frequent retrieval opportunities than may be present in current practice. These results also suggest that more research is needed to unpack these memory dynamics across conditions, as we will not understand fully how to optimize learning until these nuances, and the role sleep disturbance plays, are better considered. Finally, in this study we used a gold standard memory assessment from developmental studies to measure memory but given the language limitations in DS more work should understand how memory dynamics relate to verbal learning, such as word learning or phrase memory.

Beyond the need to expand investigations to different content domains, there are some other potential limitations in this study. The initial limitation is the difficultly "matching" or creating comparison groups between groups with TD and DS. There is controversy regarding

the method or validity of matching two groups with such different developmental trajectories. Matching as a concept is important methodologically for studies comparing groups on numerous intellectual developmental faculties. However, in studies of individuals with intellectual and developmental disabilities multiple authors have drawn attention to the issues both theoretical and methodological of matching or claiming to match samples with non-overlapping developmental pathways (Kover & Atwood, 2013; Mervis & John, 2008; Mervis & Klein-Tasman, 2004). The samples in this study were not matched on any measure, the KBIT-2 was included as a measure of cognitive capacity, and there were significant differences in the verbal and nonverbal raw scores, whilst this can suggest caution in comparing the two groups, the major focus of this paper is on the change in memory performance across time, across memory conditions within groups, not between. The TD comparison group was included largely to replicate previous findings, thereby establishing the validity of the task and experimental manipulations for use in the participants with DS; thus, the lack of "matching" or comparability in IQ scores is not a major cause for concern.

However, overall, the study is the first to show the difference in episodic memory function over long term delays in children with and without DS. These data have implications for the education and support of the DS population. The evidence that sleep has implications in the long-term retention of memory function, specifically the recall of non-tested memory, is also important for the DS population. These findings emphasize the importance of holistic support and a developmental approach to learning outcomes in DS.

Appendix

Table 1. The correlation matrix for sleep variables and memory change score in DS group

Change Score		Mean Duration	Mean Efficiency
MMP pairs	Correlation Coefficient	-.013	.281
	Sig. (1-tailed)	.472	.056
	N	33	33
MMM pairs	Correlation Coefficient	-.079	-.064
	Sig. (1-tailed)	.337	.367
	N	31	31
MMT pairs	Correlation Coefficient	-.153	.011
	Sig. (1-tailed)	.198	.475
	N	33	33
MMP actions	Correlation Coefficient	-.035	-.020
	Sig. (1-tailed)	.424	.457
	N	33	33
MMM actions	Correlation Coefficient	-.169	**-.398**[*]
	Sig. (1-tailed)	.181	**.013**
	N	31	**31**
MMT actions	Correlation	-.094	.143

	Coefficient		
	Sig. (1-tailed)	.302	.213
	N	33	33

Table 2. sequences in each condition at each time point across the four packets of Deferred Imitation task. Each four packets were equally counterbalanced across the participants and groups.

		Packet 1	Packet 2	Packet 3	Packet 4
Immediate	MMP	F I	F I	C D	C D
	MMM	A E	A E	G H	G H
	MMT	C G	C G	B F	B F
5 minute	MMP	F	I	D	C
	MMM	E	A	H	G
	MMT	C	G	F	B
1 month	Baseline	B	D	A	I

Table 3. Target actions of the nine possible sequences, in addition to the warmup sequence

Sequence	Action 1	Action 2	Action 3
Warm Up	Roll ball on the table towards the tape	Put ball in the top of the tape	N/A
A	Velcro the gem onto the side of the block	Balance the plank on the top of the block	Place the penny on the middle of the plank

B	Place the pole into the base	Hook flag to the string	Pull string
C	Place the popsicle stick vertically on the centre Velcro at the bottom of the paper	Fan out the ovals of the multi-coloured shape	Place the multi-coloured shape on the Velcro at the top right of the paper
D	Place the head onto the large spring	Place the small spring into the larger spring arms perpendicularly	Place the large spring into the base
E	Velcro on the sun and cloud to left Velcro	Slip the background into the slits of the base	Velcro tree on the right side of the plank
F	Flip the paper over to the side with the polka dots	Place the fish on the paper	Use magnet on string to pick up the fish
G	Velcro star on top of the cap	Put cap on flashlight	Shine flashlight onto paper
H	Flip base so two bars are at the bottom and four bars are facing the examiner	Balance ruler across the base	Hook plane onto chain attached to the ruler
I	Slip container in the blue cover	Balance the straw across the rim of the container	Hang the fish over the straw

Appendix Methods:

Each sequence started with a baseline assessment. During baseline, participants had 60 seconds to manipulate the objects to determine what actions and the temporal order of the actions they produced prior to the sequence modelling phase, enabling the measurement, and controlling for naturalistic or non-arbitrary sequences. Participants were permitted to interact with the items naturalistically, with the instructions, "what can you do with these things?". Non-specific praise and encouragement were given.

After each baseline assessment, the examiner performed the actions associated with one of the sequences consistent with the conditions by either modelling three times (MMM), modelling two times and pausing (MMP), or modelling two times and testing the participant (MMT) (Figure 2). The condition order was randomized so that the same condition was not performed twice in a row.

The demonstration of sequences was tightly constrained, with items placed on the table in a specific layout, from left to right. On demonstrating the first action, the administrator said, "check this out", the second action "watch this", and the third action "look at this", followed by "that's what I can do with these things". Between each action, the experimenter removed their hands from the table to clearly delineate the order of actions. The assembled items were then removed from the table (and participants view) before being disassembled. In the immediate imitation condition, and in all following tests of imitation, the items were then placed on the table and pushed across the table towards the participants with the prompt "use these things to do what I showed you". Non-specific phrases were used to motivate the participant such as "That's a neat idea", "good job!", or "cool!". If the participant correctly imitated the sequence, they were praised "well done!", and the period was terminated to prevent further actions being performed. Timings of imitation were

recorded to indicate how quickly participants carried out imitation, the maximum time permitted for imitation was 90 seconds. The pause condition was included as a measure of additional processing time, not an additional presentation, as recent research has suggested that intervals of being awake (compared to asleep) may help support memory processing (Spano et al., 2018). Further, some studies have shown reduced levels of performance initially because of the massed practice of multiple presentations, so this condition provided a manipulation of presentation number as a comparison.

Following a five-minute interval, during which an unrelated colouring task was carried out, each participant was tested on three sequences, one from each condition ('5-minute delayed recall'; see Figure 2). The same prompts and discontinue rules apply. After a one-month delay, participants returned for session 2 and the same prompted-imitation procedure was used as in the previous session. The participant was instructed first to do the two-step warm up without demonstration. The participants were then tested on the six sequences they saw at time 1, with the same discontinue rules ('1-month delayed recall'; see Figure 2). Additionally, a novel set of items was presented with the cue "What can you do with these things?" to measure the spontaneous increase in actions of each participant across development.

References:

Acebo, C., Sadeh, A., Seifer, R., Tzischinsky, O., Wolfson, A. R., Hafer, A., & Carskadon, M. A. (1999). Estimating sleep patterns with activity monitoring in children and adolescents: how many nights are necessary for reliable measures? *Sleep*, *22*(1), 95–103. doi: 10.1093/sleep/22.1.95

Barr, R., Dowden, A., & Hayne, H. (1996). Developmental changes in deferred imitation by 6- to 24-month-old infants. *Infant Behavior and Development*, *19*(2), 159–170. doi: 10.1016/S0163-6383(96)90015-6

Bartlett, J. C., & Tulving, E. (1974). Effects of temporal and semantic encoding in immediate recall upon subsequent retrieval. *Journal of Verbal Learning and Verbal Behavior*, *13*(3), 297–309. doi: 10.1016/S0022-5371(74)80066-6

Bauer, P J, Hertsgaard, L. A., & Dow, G. A. (1994). After 8 months have passed: long-term recall of events by 1- to 2-year-old children. *Memory*, *2*(4), 353–382. doi: 10.1080/09658219408258955

Bauer, P J, Wenner, J. A., Dropik, P. L., & Wewerka, S. S. (2000). Parameters of remembering and forgetting in the transition from infancy to early childhood. *Monographs of the Society for Research in Child Development*, *65*(4), i–vi, 1.

Bauer, P J, Wiebe, S. A., Waters, J. M., & Bangston, S. K. (2001). Reexposure breeds recall: effects of experience on 9-month-olds' ordered recall. *Journal of Experimental Child Psychology*, *80*(2), 174–200. doi: 10.1006/jecp.2000.2628

Bauer, Patricia J., & Leventon, J. S. (2013). Memory for One-Time Experiences in the Second Year of Life: Implications for the Status of Episodic Memory. *Infancy: The Official Journal of the International Society on Infant Studies*, *18*(5), 755–781. doi: 10.1111/infa.12005

Bäuml, K.-H. T., Holterman, C., & Abel, M. (2014). Sleep can reduce the testing effect: it enhances recall of restudied items but can leave recall of retrieved items unaffected. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *40*(6), 1568–1581. doi: 10.1037/xlm0000025

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society*, *2*(59-68).

Born, J., & Wilhelm, I. (2012). System consolidation of memory during sleep. *Psychological Research*, *76*(2), 192–203. doi: 10.1007/s00426-011-0335-6

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*(6), 633–642. doi: 10.3758/BF03202713

Carver, L J, & Bauer, P. J. (1999). When the event is more than the sum of its parts: 9-month-olds' long-term ordered recall. *Memory*, *7*(2), 147–174. doi: 10.1080/741944070

Carver, Leslie J., & Bauer, P. J. (2001). The dawning of a past: The emergence of long-term explicit memory in infancy. *Journal of Experimental Psychology: General*, *130*(4), 726–745. doi: 10.1037/0096-3445.130.4.726

Collie, R., & Hayne, H. (1999). Deferred imitation by 6- and 9-month-old Infants: More evidence for declarative memory. *Developmental Psychobiology*.

Ellenbogen, J. M., Hulbert, J. C., Stickgold, R., Dinges, D. F., & Thompson-Schill, S. L. (2006). Interfering with theories of sleep and memory: sleep, declarative memory, and associative interference. *Current Biology*, *16*(13), 1290–1294. doi: 10.1016/j.cub.2006.05.024

Fritz, C. O., Morris, P. E., Nolan, D., & Singleton, J. (2007). Expanding retrieval practice: an effective aid to preschool children's learning. *Quarterly Journal of Experimental Psychology*, *60*(7), 991–1004. doi: 10.1080/17470210600823595

Gais, S., Lucas, B., & Born, J. (2006). Sleep after learning aids memory recall. *Learning & Memory*, *13*(3), 259–262. doi: 10.1101/lm.132106

Gómez, R. L., Bootzin, R. R., & Nadel, L. (2006). Naps promote abstraction in language-learning infants. *Psychological Science*, *17*(8), 670–674. doi: 10.1111/j.1467-9280.2006.01764.x

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H. K., & Zwaan, R. A. (2014). The benefit of retrieval practice over elaborative restudy in primary school vocabulary learning. *Journal of Applied Research in Memory and Cognition*, *3*(3), 177–182. doi: 10.1016/j.jarmac.2014.05.003

Guran, C.-N. A., Deuker, L., Göttlich, M., Axmacher, N., & Bunzeck, N. (2022). Benefit from retrieval practice is linked to temporal and frontal activity in healthy young and older humans. *Cerebral Cortex Communications*, *3*(1), tgac009. doi: 10.1093/texcom/tgac009

Hughes, K. M. (2018). *The uneven profile of memory development in Down Syndrome* (Doctoral dissertation, Birkbeck, University of London).

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*(6018), 772–775. doi: 10.1126/science.1199327

Klein, P. J., & Meltzoff, A. N. (1999). Long-term memory, forgetting, and deferred imitation in 12-month-old infants. *Developmental Science*, *2*(1), 102–113. doi: 10.1111/1467-7687.00060

Kover, S. T., & Atwoo, A. K. (2013, January). *Establishing equivalence: Methodological progress in group-matching design and analysis*. American journal on intellectual and developmental

disabilities. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5656059/

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4–5), 494–513. doi: 10.1080/09541440701326154

Meltzer, L. J., Montgomery-Downs, H. E., Insana, S. P., & Walsh, C. M. (2012). Use of actigraphy for assessment in pediatric sleep research. *Sleep Medicine Reviews*, *16*(5), 463–475. doi: 10.1016/j.smrv.2011.10.002

Mervis, C. B., & John, A. E. (2008, August). *Vocabulary abilities of children with Williams syndrome: Strengths, weaknesses, and relation to visuospatial construction ability.* Journal of speech, language, and hearing research: JSLHR. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2562689/

Mervis, C. B., & Klein-Tasman, B. P. (2004). Methodological Issues in Group-Matching Designs: α Levels for Control Variable Comparisons and Measurement Characteristics of Control and Target Variables. *Journal of Autism and Developmental Disorders, 34*(1), 7–17. https://doi.org/10.1023/B:JADD.0000018069.69562.b8

Milojevich, H., & Lukowski, A. (2016). Recall memory in children with Down syndrome and typically developing peers matched on developmental age. *Journal of Intellectual Disability Research*, *60*(1), 89–100. doi: 10.1111/jir.12242

Mulligan, N., Susser, J., & Smith, S. A. (2016, April 8). *The testing effect is moderated by experimental design.* Journal of Memory and Language. https://www.sciencedirect.com/science/article/pii/S0749596X16300018?casa_token=IQoTXhVHoQEAAAAA%3AuOWuuZwwsW3VIN1io8cS_D1Ora5FzOzl2C2iGIkqMsICbSq6lvlMytVQWsZcDNOguSDQq22gARw

Racsmány, M., Conway, M. A., & Demeter, G. (2010). Consolidation of episodic memories during sleep: long-term effects of retrieval practice. *Psychological Science*, *21*(1), 80–85. doi: 10.1177/0956797609354074

Rast, M., & Meltzoff, A. N. (1995). Memory and representation in young children with Down syndrome: Exploring deferred imitation and object permanence. *Development and Psychopathology*, *7*(3), 393–407. doi: 10.1017/S0954579400006593

Roberts, L. V., & Richmond, J. L. (2015). Preschoolers with Down syndrome do not yet show the learning and memory impairments seen in adults with Down syndrome. *Developmental Science*, *18*(3), 404–419. doi: 10.1111/desc.12225

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. doi: 10.1111/j.1467-9280.2006.01693.x

Roediger III, H. L., Putnam, A. L., & Smith, M. A. (2011). *Ten Benefits of Testing and Their Applications to Educational Practice*. In *Psychology of Learning and Motivation*: *Vol. 55* (pp. 1–36). Elsevier. doi: 10.1016/B978-0-12-387691-1.00001-6

Sheffield, E. G., & Hudson, J. A. (2006). You Must Remember This: Effects of Video and Photograph Reminders on 18-Month-Olds' Event Memory. *Journal of Cognition and Development*, *7*(1), 73–93. doi: 10.1207/s15327647jcd0701_4

Silverman, W. (2007). Down syndrome: cognitive phenotype. *Mental retardation and developmental disabilities research reviews*, *13*(3), 228-236.

Spanò, G., Gómez, R. L., Demara, B. I., Alt, M., Cowen, S. L., & Edgin, J. O. (2018). REM sleep in naps differentially relates to memory consolidation in typical preschoolers and children with Down syndrome. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(46), 11844–11849. doi: 10.1073/pnas.1811488115

Starling, D. S. V., Moreira, B. F. T., & Jaeger, A. (2019). Retrieval practice as a learning strategy for individuals with Down syndrome A preliminary study. *Dementia & Neuropsychologia*, *13*(1), 104–110. doi: 10.1590/1980-57642018dn13-010012

Tucker, M. A., & Fishbein, W. (2008). Enhancement of declarative memory performance following a daytime nap is contingent on strength of initial task acquisition. *Sleep*, *31*(2), 197–203. doi: 10.1093/sleep/31.2.197

Wilhelm, I., Diekelmann, S., & Born, J. (2008). Sleep in children improves memory performance on declarative but not procedural tasks. *Learning & Memory*, *15*(5), 373–377. doi: 10.1101/lm.803708

Zhuang, L., Wang, J., Xiong, B., Bian, C., Hao, L., Bayley, P. J., & Qin, S. (2022). Rapid neural reorganization during retrieval practice predicts subsequent long-term retention and false memory. *Nature Human Behaviour*, *6*(1), 134-145.