JSLHR

Research Article

# After Initial Retrieval Practice, More Retrieval Produces Better Retention Than More Study in the Word Learning of Children With Developmental Language Disorder

Laurence B. Leonard,[a] Patricia Deevy,[a] Jeffrey D. Karpicke,[a] Sharon L. Christ,[a] and Justin B. Kueser[a]

**Purpose:** Children with developmental language disorder (DLD) often have difficulty with word learning. Recent studies have shown that incorporating retrieval practice provides a significant benefit to this learning. However, we have not yet discovered the best balance between the amount of retrieval and the amount of study (hearing the word in the presence of the referent) that is provided. In this investigation, we compared a word learning procedure using more retrieval and less study with a procedure that used more study and less retrieval.

**Method:** Participants were 13 children with DLD and 13 same-age peers with typical language development (TD). Both groups ranged in age from 4 to 6 years. The children learned two sets of novel words, with each set taught in two sessions. During an initial criterion period, the children had the opportunity to retrieve all of the words. Following this period, the words were either retrieved without further study or studied without additional retrieval. Recall and recognition testing immediately followed the second learning session and was repeated 1 week later. Testing assessed the children's retention of both the word forms and their meanings.

**Results:** Better recall both immediately after learning and after 1 week was seen for the more retrieval/less study condition. This was seen for both groups of children for word form recall and for children with DLD for meaning. Group differences were not found.

**Conclusion:** This study served as a stringent test of the benefits of retrieval to children's word learning. Continued retrieval after initial retrieval practice appeared to be helpful even when further study was discontinued and when the comparison study condition had also provided retrieval practice in the initial stages. Further refinement of retrieval procedures might lead to the development of useful clinical tools to promote word learning.

Imagine the following scenario. Two actors, Actor 1 and Actor 2, prepare for their roles in a theatrical production. After they study their lines for a key dialogue, they recite their respective parts until they manage a mistake-free rehearsal. After that initial success—1 week before opening night—Actor 1 develops laryngitis and confines his additional preparation to watching his understudy rehearse with Actor 2. The morning of the production's opening, Actor 1's voice returns. However, the director is nervous about the recovered actor going on stage, fearing he might not remember all of his lines.

We can easily recognize the reason for the director's concern. We understand that, for an activity such as a stage performance, one successful rehearsal is not likely to be enough. Continued active recall of the lines seems necessary. Only Actor 2 met this standard. By only passively studying the lines after he lost his voice, Actor 1 could well draw a blank at crucial moments of the performance and put the entire production in jeopardy.

The general premise of this article is that there are other learning activities for which limited successful "rehearsal" followed by passive study is likely to produce less than satisfactory results. Here, we report a study in which this notion is applied to children's word learning. We replace "rehearse" and "recite from memory" with "retrieve

[a]Purdue University, West Lafayette, IN

Correspondence to Laurence B. Leonard: xdxl@purdue.edu

Editor-in-Chief: Stephen M. Camarata
Editor: Mary Alt

Received March 4, 2020
Revision received April 21, 2020
Accepted May 13, 2020
https://doi.org/10.1044/2020_JSLHR-20-00105

*Disclosure: The authors have declared that no competing interests existed at the time of publication.*

from memory," but in other respects, the word learning situation mirrors our acting scenario. Specifically, in this study, after successfully retrieving a set of new words, children either continued to receive retrieval practice or continued to study the words without additional attempts at recall.

Our study focused on preschool-age children with developmental language disorder (DLD), often referred to as children with specific language impairment. These children have a significant and persistent impairment in language ability that cannot be explained by weaknesses in sensory, motor, or cognitive functioning. Word learning is one of the vulnerable areas in these children (see reviews in Kan & Windsor, 2010; Leonard, 2014). The goal of this study is to provide a more stringent test of our earlier findings showing that children with DLD learn words more successfully when they engage in repeated retrieval than when learning is restricted to repeated study of the words. In our earlier work, the repeated retrieval condition provided children not only with continuous retrieval practice but also the same number of study trials as in the repeated study condition. In this study, we reduce the number of study trials in the repeated retrieval condition and provide early retrieval opportunities for words in the study condition. In essence, this is a comparison between more retrieval with less study and more study with less retrieval. We begin with a review of the earlier work and its theoretical underpinnings.

### *The Value of Retrieval Practice*

Our work has been influenced by a longstanding finding in the memory literature that, when people attempt to retrieve information during the learning period, their recall of that information is significantly improved. Retrieval is more than a test of what has already been learned; it is itself a form of learning. One traditional way this has been measured is to compare the learning that occurs from jointly studying and retrieving information with the learning that occurs from studying alone. In the vast majority of experiments of this type, repeated study with retrieval produces better learning (see Rowland, 2014). The nature of the material to be learned has varied widely, and in recent years, children, like adults, have been found to benefit from retrieval practice (see Fazio & Marsh, 2019).

Two procedural details in the retrieval literature have been especially helpful to our work. First, studies have shown that retrieval is most effective when it occurs frequently during the learning phase (e.g., Roediger & Karpicke, 2006). Second, retrieval produces better results when it is somewhat effortful—a condition often created by inserting other items between a study trial and the retrieval trial for the same item ("spaced retrieval"; e.g., Karpicke & Roediger, 2007).

The precise mechanisms responsible for these effects are not yet clear. One account that provides a plausible explanation for these effects is the "episodic context" account of Karpicke et al. (2014). According to this account, during encoding, features of the context are registered along with the material to be learned. During successful retrieval, these contextual features are reinstated and combined with the present context to form a composite. With each subsequent retrieval, additional contextual features are included in the composite, rendering the composite increasingly distinct from its competitors, which, in turn, aids the memory search. Spaced retrieval amplifies this distinctiveness because, when other items intervene between a study trial and a retrieval trial, the context changes to a greater degree, adding to the uniqueness of the composite of contextual features.

In everyday learning situations, features of the context include things such as the physical setting, the time of day, and any other people who may have been present. However, in experimental studies, contextual features are usually subtle and include details such as the particular list in which an item appeared and the order of the item on the list. These subtle contextual details have been shown to be accessible to retrieval even when they were not brought to the attention of the learner during the initial study period (e.g., Whiffen & Karpicke, 2017).

### *Repeated Spaced Retrieval and Word Learning in DLD*

By combining the procedural details of repeated retrieval and spaced retrieval, we have been able to apply retrieval practice to children's word learning. Such practice seems to have special relevance for children with DLD. The vocabularies of these children lag behind those of their peers with typical language development (TD) from the preschool years into adulthood, with the gap between the two groups becoming larger over time (Rice & Hoffman, 2015). When asked to learn a set of novel words, children with DLD require more exposure to these words than their age mates with TD to meet the same criterion level (e.g., Alt, 2011; Gray, 2004; McGregor et al., 2013). Findings such as these have led researchers to explore the possibility that these children might benefit from procedures that include retrieval.

To our knowledge, the first studies of DLD to incorporate retrieval into novel word learning procedures were those of Chen and Liu (2014) and McGregor et al. (2017). Chen and Liu focused on children of preschool age, whereas McGregor et al. studied young adults. In both studies, the individuals with DLD made greater gains when retrieval was included in the protocol. Comparison conditions involved learning without retrieval. This work was soon followed by a series of studies of preschool-age children with DLD by our research group. Typically developing children matched for age served as a comparison group. All studies employed repeated spaced retrieval as one of the learning conditions. In our first study, children were asked to learn the novel names of exotic plants and animals (e.g., /fɪm/)—a measure of word form—and what each plant or animal "liked" (e.g., rain)—a measure of meaning (Leonard, Karpicke, et al., 2019). For each child, half of the words were learned in a repeated spaced retrieval condition, and half were learned in a repeated study condition. (We use the term "study" as this term is common in the retrieval literature. Here, it refers to children being asked to learn new words they repeatedly hear while viewing a photo of the corresponding exotic plant or animal on a computer screen.) The children participated in two learning sessions, held on consecutive days. Recall testing ("What's this called?" "What does this one like?") and recognition testing (e.g., "Where's the /fɪm/?") were conducted immediately after the second learning session and 1 week later. Key findings were that both groups of children showed greater recall of the novel words if they were learned in the repeated spaced retrieval condition than if they were learned in the repeated study condition. This was true for both word form (e.g., /fɪm/) and meaning (e.g., likes rain), though effect sizes were larger for word form. For both groups, recall was as good after 1 week as it was immediately after learning. A trend toward higher scores for the children with TD than the children with DLD was not statistically significant. Recognition testing proved to be less sensitive, in large part due to ceiling effects. The task required only an imprecise representation of the word—just detailed enough to allow children to recognize the word relative to the (quite phonetically distinct) alternative words.

In a second study involving novel names of plants and animals, the comparison condition was repeated immediate retrieval rather than repeated study (Haebig et al., 2019). By using immediate retrieval as a comparison condition, we could determine whether the spacing of retrieval trials was beneficial over and beyond retrieval in general. In other respects, the procedure used in the second study was the same as in our first study. Again, we found that repeated spaced retrieval produced the greatest gains in recall, with effect sizes larger for word form than for meaning. Recall was stable over the 1-week period. Unlike in our first study, the children with TD showed greater recall than the children with DLD. They also scored higher on the recognition test. Only the children with DLD showed the advantage for repeated spaced retrieval on the recognition test; the children with TD were at ceiling levels.

The third study went back to a comparison between repeated spaced retrieval and repeated study but employed novel adjectives that referred to unusual attributes associated with common objects, as in "a /taɪmɪk/ pencil" and "a /taɪmɪk/ toothbrush" (Leonard, Deevy, et al., 2019). This study allowed us to determine if children could not only learn novel adjectives but also apply them to objects that were never used during the learning period. An advantage for repeated spaced retrieval was again seen for both groups. Yet, regardless of learning condition, if children were able, during testing, to apply a novel adjective correctly to an object that had been used during the learning period, they were also able to successfully apply it to a new object (e.g., "a /taɪmɪk/ flower"). Again, recall was stable over 1 week. The two groups of children did not differ in their recall scores. On the recognition test, the children with DLD were more accurate on words in the repeated spaced retrieval condition than on words in the repeated study condition. The children with TD were at ceiling for words in both conditions.

### This Study

An unanswered question regarding retrieval effects on word learning is the degree to which additional retrieval after initial successful retrieval helps children retain the words relative to continued study without additional retrieval. In our acting scenario, we assumed that Actor 2 would be more likely to remember the material than Actor 1 because of continued rehearsal rather than only continued study after one successful rehearsal. Will this be true for children's word learning?

We can use as a guide to answering this question an experiment by Karpicke and Roediger (2008). These investigators asked college student participants to learn the English translations of a set of Swahili words. Participants were assigned to different learning conditions. Two of the conditions are especially relevant to this study. In both conditions, participants began by studying and retrieving each item. In one condition, after an item was successfully retrieved, participants no longer studied that item but continued to receive retrieval trials for the item. The converse was true for the second condition; after an item was successfully retrieved, it continued to appear in study trials but no longer appeared in a retrieval trial. Testing 1 week later showed that the condition involving repeated retrieval (only) after the first successful retrieval resulted in recall scores that were more than twice as large as those seen for the condition involving repeated study (only) after the first successful retrieval.

In this study, we created conditions that resembled these two conditions in the Karpicke and Roediger (2008) study and applied them to children's novel word learning. We refer to them here as the "more retrieval/less study" and "more study/less retrieval" conditions, respectively. We hypothesized that preschool-age children with DLD would show greater learning and retention of novel words in the "more retrieval/less study" condition. As a basis of comparison, we also recruited a group of children with TD matched for age. Although we expected the children with TD to show better retention overall, our main interest was in determining if retrieval processes in children with DLD functioned in the same way as in TD, despite any differences seen in language ability.

## Method

### Participants

The research procedures used here were approved by the first author's institutional review board. Informed written consent was obtained from the parents, and verbal assent was given by the children. Twenty-six children participated in the study. Thirteen children (seven boys, six girls) met our selection criteria for the DLD group, and 13 children (six boys, seven girls) displayed TD and so were in the comparison group. Efforts were made to ensure a representative number of girls in the DLD group; otherwise, children were included regardless of sex. The group with TD included one Asian/Pacific Islander; all other children were identified by their parents as White (non-Hispanic).

The children in the DLD group ranged in age from 48 to 71 months (*M* = 56.69, *SD* = 6.50). All children were enrolled in a language intervention program or were scheduled to be enrolled in such a program. The children went through the following selection process. Children met the selection criterion for language if their standard scores on the Structured Photographic Expressive Language Test–Preschool 2 (SPELT-P 2; Dawson et al., 2005) were below 87, the cutoff determined by Greenslade et al. (2009) to show acceptable levels of sensitivity and specificity. If their scores on the SPELT-P 2 were just above the cutoff score, additional language assessment occurred. This was true for three children who scored 87 on the SPELT-P 2. For these children, we computed their finite verb morphology composite scores (Goffman & Leonard, 2000) and developmental sentence score (Lee, 1974) based on a spontaneous speech sample obtained during the assessment session. All three children scored below the cutoff score (89%) for adequate sensitivity and specificity on the finite verb morphology composite based on the data of Souto et al. (2014). In addition, these three children scored below the 10th percentile on developmental sentence scoring (Lee, 1974). We applied these additional criteria because they were already candidates for language intervention and the fact that, in the local participant recruitment area, SPELT-P 2 scores skew higher for both the DLD and TD populations. For the entire DLD group, the mean SPELT-P 2 standard score was 77.15 (*SD* = 11.89). Their standard scores on the Kaufman Assessment Battery for Children, Second Edition (Kaufman & Kaufman, 2004) ranged from 83 to 130 (*M* = 103.62, *SD* = 13.52). All children passed a pure-tone hearing screening at a level of 20 dB HL at 500, 1000, 2000, and 4000 Hz. They also scored in the “minimal to no symptoms of autism spectrum disorder” range (between 15 and 29.5) on the Childhood Autism Rating Scale–Second Edition (CARS-2; Schopler et al., 2010). The mothers’ years of education averaged 16.54 years (*SD* = 2.67).

Standardized tests of vocabulary have not shown adequate levels of sensitivity and specificity (Gray et al., 1999; Spaulding et al., 2013) and were therefore not used as part of our selection criteria. However, for descriptive purposes, we administered both the Expressive Vocabulary Test, Second Edition (EVT-2; Williams, 2007) and the Peabody Picture Vocabulary Test–Fourth Edition (Dunn & Dunn, 2007). The standard scores of the children with DLD on these tests averaged 99.31 (*SD* = 9.47) and 103.77 (*SD* = 13.64), respectively. We did not require vocabulary test scores to exceed a certain level; indeed, the DLD group’s scores were very similar to those we have reported in our previous studies, and it is not uncommon to find age-appropriate vocabulary test scores in other studies of children with DLD (e.g., McGregor et al., 2012). Furthermore, some earlier studies, including our own, have found no relationship between vocabulary test scores and novel word learning performance (e.g., Gray, 2003; Haebig et al., 2019; Leonard, Karpicke, et al., 2019).

The children with TD were selected to resemble the children with DLD in age, *t*(24) = 0.44, *p* = .667. They, too, ranged in age from 48 to 71 months (*M* = 57.80, *SD* = 6.47). Six were boys, and seven were girls. All children scored above 87 on the SPELT-P 2 (*M* = 113.46, *SD* = 11.11). Scores on the Kaufman Assessment Battery for Children, Second Edition, averaged 112.00 (*SD* = 8.51). Years of mothers’ education averaged 16.15 years (*SD* = 2.34). All children passed a hearing screening. Children in the group with TD were not administered the CARS-2, as all potential research participants for this group were prescreened for parental concerns and were reported to have no language or cognitive difficulties.

For comparison purposes, the children with TD were also given both the EVT-2 (*M* = 114.92, *SD* = 11.52) and the Peabody Picture Vocabulary Test–Fourth Edition (*M* = 121.85, *SD* = 7.26). Not surprisingly, the children with TD had significantly higher scores on these tests than the children with DLD, *t*s(24) ≥ 3.78, *p*s < .001. A summary of the test scores of both participant groups can be seen in Table 1.

To gain a picture of how the children’s productions of the novel words might relate to their phonological characteristics, we also administered an 18-item short-sentence repetition task. The items on this task tested the same consonants and vowels, in the same word positions, as the novel words. For example, corresponding to the novel word /jʌt/ (see below), there were items testing word-initial /j/ (“I like *you*”), medial /ʌ/ (“Yuck, a *bug*!”), and word-final /t/ (“Wear a *hat*!”). Although developmental errors were expected, of special interest was the identification of any unusual productions (e.g., labial or velar assimilation) that might help us better interpret the children’s novel word productions. The items used in this sentence repetition task are provided in the Appendix.

### Materials and Procedure

The children learned six novel words, divided into two sets of three words. In each set, two words were monosyllabic consonant–vowel–consonant (CVC) words and one was a disyllabic word (CVCV or CVCVC). Together, monosyllabic and disyllabic words constitute approximately 90% of the words that children hear between the ages of 2 and 6 years (Roark & Demuth, 2000). The novel words were /fumi/, /jʌt/, /nɛp/, /tɛkət/, /bog/, and /paɪb/. The novel words served as the names of exotic animals and plants depicted in color photographs. These photographs, originally used by McGregor (2014), were among those used in two of our previous studies (Haebig et al., 2019; Leonard, Karpicke, et al., 2019). Each set was presented in a different learning condition. The words in the two sets were matched on number of syllables, phonotactic probability

**Table 1.** Mean standard scores (and standard deviations) on the standardized tests administered to the children with developmental language disorder (DLD) and the children with typical language development (TD).

| | DLD<br>*N* = 13 (7 boys) | | TD<br>*N* = 13 (6 boys) | | |
|---|---|---|---|---|---|
| Variable | *M* | *SD* | *M* | *SD* | Group comparisons |
| Age (months) | 56.69 | 6.50 | 57.80 | 6.47 | *p* = .667 |
| Maternal education (years) | 16.54 | 2.67 | 16.15 | 2.34 | *p* = .699 |
| SPELT-P 2 | 77.15 | 11.89 | 113.46 | 11.11 | *p* < .001 |
| KABC-II | 103.62 | 13.52 | 112.00 | 8.51 | *p* = .071 |
| EVT-2 | 99.31 | 9.47 | 114.92 | 11.52 | *p* < .001 |
| PPVT-4 | 103.77 | 13.64 | 121.85 | 7.26 | *p* < .001 |

*Note.* SPELT-P 2 = Structured Photographic Expressive Language Test–Preschool 2; KABC-II = Kaufman Assessment Battery for Children, Second Edition; EVT-2 = Expressive Vocabulary Test, Second Edition; PPVT-4 = Peabody Picture Vocabulary Test–Fourth Edition.

(average biphone frequency), and neighborhood density based on the Storkel and Hoover (2010) database.

Each child participated in two learning conditions, one providing more retrieval and less study (the "more retrieval/less study" condition), the other providing more study and less retrieval (the "more study/less retrieval" condition). A different set of words was used for each condition. For the first 12 children in each group, the two sets were counterbalanced both for the condition to which they were assigned and for the order in which they were presented. For the remaining child in each group, the condition assignment and order of presentation were selected at random.

### Learning Phase

Each set of words was learned in two sessions, each approximately 20 min in duration, held on consecutive days. The first session was devoted entirely to the learning phase with short breaks. The second session began with a continuation of the learning phase for 10 min with short breaks, followed by a longer (5-min) break and then testing, which required 5 min or less (see below). The pictures of the words to be learned, along with digitally recorded study and retrieval trials, were presented via laptop computer. The children were told they were going to learn some new words for funny plants and animals. At the beginning of each session, the children were given a sticker page featuring familiar cartoon dogs about to embark on a path toward a dish of treats. At regular intervals during the session, a picture of dogs appeared on the computer screen, signaling time for a sticker break. During these breaks, the children could fill in segments of the path and see their progress. In addition, during the retrieval-only and study-only phases of the task, 5-s animated video clips were interspersed with experimental trials to keep the children engaged.

Table 2 provides an example of the first session for one of the sets. Although each set of words constituted a different learning condition, the first session for each set began in the same way, with a familiarization period followed by the initial criterion period. During familiarization, each word was initially introduced in a study trial. The child saw the picture of the animal or plant and heard both its name and what it "likes." The association between an animal or plant and what it likes was completely arbitrary, and there was no visual information in the picture that could serve as a cue to this association. An example is "This is a /fumi/. It's a /fumi/. A /fumi/ likes snow." We refer to the word itself as the "word form" and what the referent likes as the "meaning." Following the study trial, a retrieval trial was presented, in which the picture reappeared on the screen and the child heard the requests "What's this called? What do we call this?" After the child's response, the meaning was requested, as in "And what does this one like? What does it like?" Regardless of the accuracy of the child's response, a study trial followed, which was identical to the initial study trial.

After each word in the set proceeded through this study trial–retrieval trial–study trial sequence, the initial criterion period began. For this period, each word appeared in a retrieval trial–study trial–retrieval trial sequence, with the words alternating in random order except that the sequence for a word never appeared immediately after the sequence for the same word. With this arrangement, each sequence began with a spaced retrieval trial since one or two other words intervened between the initial retrieval trial of the sequence and the most recent study trial for the same word (which appeared in an earlier sequence; see Table 2). The second retrieval trial in each sequence was an immediate retrieval trial, because it directly followed a study trial. We established the criterion that a child had to correctly retrieve the *word form* of each word for four *immediate retrieval trials*. This was a criterion that all children could meet. As just described, we also included spaced retrieval trials during this initial criterion period because our earlier work had found that it led to better retention than immediate retrieval. Pilot work made it clear that not all word forms would be correctly retrieved in a spaced retrieval trial during this initial criterion period without significantly extending this period with the risk of causing fatigue or frustration on the part of the child. For this reason, we used the number of different word forms correctly produced on a

**Table 2.** An example of the first session.

**I. Practice and familiarization**

| Novel word | Trial type |
|---|---|
| tiger[a] | Study–retrieve |
| /nɛp/ | Study–retrieve–study |
| /fumi/ | Study–retrieve–study |
| /jʌt/ | Study–retrieve–study |

**II. Initial criterion period**

| Novel word | Trial type: ”spaced”–study–“immediate” sequence | Spaced retrieval | Immediate retrieval |
|---|---|---|---|
| /fumi/ | Retrieve[b]–study–retrieve[c] | | + |
| /jʌt/ | Retrieve–study–retrieve | | + |
| /nɛp/ | Retrieve–study–retrieve | | + |
| /jʌt/ | Retrieve–study–retrieve | | + |
| /nɛp/ | Retrieve–study–retrieve | + | + |
| /fumi/ | Retrieve–study–retrieve | | + |
| /jʌt/ | Retrieve–study–retrieve | | + |
| /fumi/ | Retrieve–study–retrieve | | + |
| /nɛp/ | Retrieve–study–retrieve | + | + |
| /fumi/ | Retrieve–study–retrieve | + | + |
| /nɛp/ | Retrieve–study–retrieve | + | + |
| /jʌt/ | Retrieve–study–retrieve | | + |
| | | Practice score = 2 | Criterion met |

**III. Short break and refresher**

| Novel word | Trial type: ”spaced”–study–“immediate” sequence |
|---|---|
| /fumi/ | Retrieve–study–retrieve |
| /jʌt/ | Retrieve–study–retrieve |
| /nɛp/ | Retrieve–study–retrieve |

**IV. Division into separate conditions: more retrieval or more study**

| Novel word | Trial type |
|---|---|
| /fumi/ | Retrieve only or study only |
| /jʌt/ | Retrieve only or study only |
| /nɛp/ | Retrieve only or study only |
| /jʌt/ | Retrieve only or study only |
| /nɛp/ | Retrieve only or study only |
| /fumi/ | Retrieve only or study only |
| /jʌt/ | Retrieve only or study only |
| /fumi/ | Retrieve only or study only |
| /nɛp/ | Retrieve only or study only |
| *Continue…* | |

*Note.* In this example, in the initial criterion period, the child met the criterion of four correct retrieval trials for each novel word in an immediate retrieval trial and could proceed to the retrieval-only or study-only phase. The child also successfully retrieved two different novel words (/nɛp/ and /fumi/) in a spaced retrieval trial during the initial criterion period. This number (2) served as a covariate in the data analysis. + = correct.

[a] A real word was used as the first item to ensure that the child understood the task. [b] A spaced retrieval trial during the initial criterion period. The number of different novel words retrieved during this period served as a covariate. [c] An immediate retrieval trial during the initial criterion period. Each novel word had to be correctly retrieved 4 times on this type of trial before proceeding to the retrieval-only or study-only phase.

spaced retrieval trial during this initial criterion period as a covariate for the word form recall test, the meaning recall test, and the recognition test. This covariate was a measure of the children's early success with recalling the word forms before the words appeared in separate conditions and was therefore needed as a control in order to more accurately gauge the differences between more retrieval/less study and more study/less retrieval. (During the initial criterion period, the children were also asked for each word's meaning in both spaced and immediate retrieval trials. However, even for spaced retrieval trials, errors on meaning were extremely rare [97% correct]. With an extremely narrow range, these scores were not statistically associated with the children's meaning recall test scores and were therefore not suitable as a covariate.) Because each set represented a different learning condition, the number of word forms correct on a spaced retrieval trial in the initial criterion period was computed separately for each set. The statistical approach used to compare learning conditions accommodated separate covariate scores for each set.

Once the child met the criterion of four correct word form responses in immediate retrieval trials for each word, a brief break was taken. Then one additional retrieval trial–study trial–retrieval trial sequence for each word was completed as a "refresher" before continuing. The remainder of the session was devoted to either study-only trials or retrieval-only trials, depending on the condition that had been assigned to that set. These trials were identical to the study and retrieval trials used in the initial criterion period. There were four study-only or retrieval-only trials for each word on the first day. Again, words appeared in random order, except that the same word never appeared consecutively.

The session held on the second day began with a "refresher"—a single retrieval trial–study trial–retrieval trial sequence for each word. Then, each word appeared in five study-only or retrieval-only trials, depending on the condition. In total, counting all trials, including the familiarization period, the initial criterion period, and the "refresher" period, all words regardless of condition had at least eight study trials, seven immediate retrieval trials, and six spaced retrieval trials for each word. The words in the two conditions differed in either having nine additional study (only) trials or nine additional retrieval (only) trials per word.

#### Testing

The child had a 5-min break after the final study-only or retrieval-only trial of the second session and then was given a recall test. We refer to this test as the "5-min" recall test. Each word was tested twice in random order, though the same word was never tested in two consecutive items. The items were identical to the retrieval trials, with digitally presented audio and video presented via laptop computer. The children's responses were audio-recorded.

One week later, the child was again administered the recall test (referred to as the "1-week" recall test). The child was then given a form-referent link recognition test (referred to here as simply the "recognition test"). For this test, each item consisted of three pictures on the laptop computer screen (the target picture and the pictures of the other two referents in the set), and the child heard the audio-recorded request "Which one is the (e.g., /fumi/)? Where's the (e.g., /fumi/)?" Each word form was tested twice with no word appearing twice in a row. The recognition test was administered after the recall test to avoid having the children hear each novel word (as occurs during recognition testing) prior to being asked to recall it.

### *Scoring and Reliability*

Scoring of the children's word form responses on the 5-min and 1-week tests involved several steps. First, the response could not resemble a real word that could be a reasonable (though incorrect) real name for the referent (e.g., "squirrel"). Second, the response met the subjective criterion of being a plausible attempt at the correct word form. In making this judgment, we also consulted the children's productions in the 18-item short-sentence repetition task in case there were unusual errors that we should take into consideration when interpreting the children's novel word productions (e.g., labial assimilation, metathesis). With the exception of one child who exhibited considerable initial consonant omission, unusual errors were not seen. Next, if the word appeared to be a plausible attempt at the novel word, we applied the scoring method created by Edwards et al. (2004). For each consonant, 1 point each was credited for correct place, manner, and voicing. For each vowel, 1 point each was credited for correct height, length, and backness. One additional point was given for correct syllable shape (e.g., CVC). For example, the response /fupi/ for the correct form /fumi/ would be scored as 3 + 3 + 1 + 3 + 1 = 11, as the second consonant received only 1 point for place. This score was then compared to the score earned if we assumed that the response was actually an attempt at one of the other word forms in the study. For example, if treated as an attempt at /nɛp/, the response /fupi/ would earn a total score of 3, with points awarded only because the second consonant, /p/, matched that of the correct word form. If the points awarded to the presumed correct response were higher than the points awarded to all other novel word forms, it was considered correct.

The scoring of meaning did not require a phonetic scoring system. Although responses were sometimes phonetically inaccurate (e.g., /wen/ for "rain"), they were readily interpretable. However, during the learning period, two children in the DLD group unexpectedly seemed confused by the questions "And what does this one like? What does it like?" (even in immediate retrieval trials) and provided the word form instead of what the plant or animal liked. It seemed possible that they had difficulty with "does" in the questions and interpreted the questions as "What is this one like? What is it like?" (In two previous studies employing the same questions, we had never encountered this issue.) For this reason, data from these two children were excluded from the analysis for meaning recall. Their data were included in all other analyses. Correct responses were credited on the recognition test if the child pointed to the correct picture. Self-corrections were allowed when they were immediate.

To assess interjudge reliability, two of the investigators independently scored the 5-min and 1-week word form recall test responses of eight children—four from each participant group. Scoring of each response was correct or incorrect using the criteria described above. For responses that were phonetically inaccurate but plausible attempts at the correct novel word, the independent scorers applied the Edwards et al. (2004) scoring method to determine if the response met the criteria for being regarded as a correct response. Item-to-item interscorer agreement for judging responses as correct or incorrect was 100%.

### *Data Analysis*

Separate analyses were conducted for the number of correct items on the word form recall, meaning recall, and recognition tests. For each of these measures, a series of mixed-effects models was estimated, using a random intercept at the child level with repeated measures nested within a child. There were four repeated observations per child totaling 104 observations for word form recall and 96 observations for meaning recall due to two fewer participants. There were two repeated observations per child totaling 52 observations for the recognition analysis. Random slopes for learning condition and time were included if they did not approximate zero. For word form recall and meaning recall outcomes, models included participant group (DLD vs. TD), learning condition (more retrieval/less study vs. more study/less retrieval), and time (5 min vs. 1 week). For recognition, only participant group and learning condition were included, as this test was administered only at the 1-week mark. For all outcomes, we employed models with and without the covariates of maternal education, EVT-2 standard scores, and practice scores (number of word forms successfully retrieved in a spaced retrieval trial during the initial criterion period; range: 0–3 for each set). For the analyses for meaning recall and recognition, bootstrapped standard errors (with 500 replicates) were used because the outcomes for these measures were highly left-skewed. For each outcome, the random slope for learning condition was relevant in all models, suggesting that the learning condition effect varied across children. We estimated main effects models and models with two- and three-way interactions among the three primary study variables of participant group, learning condition, and time (with and without covariates). We present here main effects models and models that include interactions that were statistically and substantively relevant. Both the random effects and the Level 1 variance–covariance structures were independent for all models. Effect sizes are reported as partially standardized beta coefficients ($b_{\text{std}}$), which are comparable to Cohen's *d,* except they represent conditional, standardized mean differences conditioned on the other variables in the model.

Although the order of sets was counterbalanced, we ran preliminary analyses to ensure that set order was neither significant nor interacting with the other factors. This proved to be the case. All *p*s were nonsignificant.

## Results

A summary of the children's accuracy on the word form recall, meaning recall, and recognition tests can be seen in Table 3. Values are unconditional means (with standard deviations) for the number of items correct.

**Table 3.** The unconditional means (and standard deviations) for word form recall, meaning recall, and recognition by the children with developmental language disorder (DLD) and those with typical language development (TD) in the more retrieval/less study condition and the more study/less retrieval condition.

| | More retrieval/ less study | | More study/ less retrieval | |
|---|---|---|---|---|
| **Variable** | **5 min** | **1 week** | **5 min** | **1 week** |
| Form | | | | |
| DLD | 3.23<br>(2.20) | 3.77<br>(2.13) | 1.77<br>(1.92) | 2.31<br>(1.84) |
| TD | 3.92<br>(1.75) | 3.62<br>(1.89) | 2.85<br>(1.68) | 3.62<br>(1.66) |
| Meaning | | | | |
| DLD | 5.55<br>(0.82) | 5.36<br>(1.29) | 4.64<br>(1.86) | 4.82<br>(2.04) |
| TD | 5.69<br>(0.75) | 5.77<br>(0.83) | 5.54<br>(1.39) | 5.77<br>(0.60) |
| Recognition | | | | |
| DLD | | 7.62<br>(2.33) | | 7.23<br>(2.17) |
| TD | | 8.54<br>(1.39) | | 8.08<br>(1.66) |

### *Word Form Recall*

The most appropriate model for word form recall was the main effects model with covariates, shown in Table 4. A moderate to large effect was seen for learning condition, such that test scores for the more retrieval/less study condition were 1.09 points higher, on average, than those for more study/less retrieval ($p$ = .001). (Mean percentages correct for the two conditions were 60.57% and 43.91%, respectively.) A small effect was also seen for time, with scores at 1-week testing 0.38 points higher than at 5 min ($p$ = .030). (Mean percentages correct were 55.45% and 49.00% correct, respectively.) Participant group showed no effect in the model with covariates. In the model without covariates, the children with TD had 1.12 ($p$ = .07) higher value, on average, compared to the children with DLD. There were no statistically significant two-way interactions; likewise, the three-way interaction was not statistically significant. In Table 4, it can be seen that the covariate of practice score was statistically significant. This score is the number of words

**Table 4.** The main effects model for word form recall with the covariates included (26 participants and 104 observations).

| Fixed effects | $b$ | 95% CI | | $b_{\text{std}}$ | $p$ |
|---|---|---|---|---|---|
| Group (DLD vs. TD) | −0.03 | −1.33 | 1.26 | −0.02 | .962 |
| Condition (MR vs. MS) | 1.09 | 0.45 | 1.74 | 0.56 | .001 |
| Time (1 wk vs. 5 min) | 0.38 | 0.04 | 0.73 | 0.20 | .030 |
| Covariates | | | | | |
| EVT-2 | 0.00 | −0.05 | 0.04 | 0.00 | .878 |
| Mother's education | −0.03 | −0.27 | 0.21 | −0.01 | .814 |
| Practice score | 1.23 | 0.80 | 1.66 | 0.63 | .000 |
| Intercept | 1.58 | −3.86 | 7.02 | −0.79 | .570 |
| **Random effects** | $\sigma^2$ | | | | |
| Condition | 2.00 | 0.87 | 4.61 | | |
| Intercept | 1.05 | 0.45 | 2.48 | | |
| Level 1 residual | 0.82 | 0.53 | 1.27 | | |

*Note.* The unconditional model random effect variance was 1.35 [0.59, 3.09], and Level 1 residual variance was 2.56 [1.87, 3.50]. $b_{\text{std}}$ are partially standardized coefficients where the outcome is in standard deviation units. CI = confidence interval; DLD = children with developmental language disorder; TD = age-matched children with typical language development; MR = more retrieval/less study condition; MS = more study/less retrieval condition; 1 wk = recall test administered 1 week after the second learning session; 5 min = recall test administered 5 min after the second learning session; EVT-2 = Expressive Vocabulary Test–Second Edition.

correct on a spaced retrieval trial in the initial criterion period. This means that the practice scores and later recall test scores were related (the higher the practice score, the better the outcome). However, and importantly, this relationship had no bearing on the differences between the two learning conditions, as this comparison was conditioned on the covariate scores. Figure 1 provides an illustration of the key findings. (We note here that both learning condition [$b_{\text{std}} = 0.51$, $p = .02$] and time [$b_{\text{std}} = 0.20$, $p = .03$] also showed an effect in the model without the covariates.)

The scoring system for word form allowed for phonetic errors in the children's recall productions. For example, the phonetically inaccurate productions /funi/ and /fupi/ would both be regarded as a correct attempt at the novel word /fumi/, yet /funi/ would earn a score of 12, whereas the score for /fupi/ would be 11. As a supplemental analysis, we computed the children's mean percentage of possible points earned for recall test responses that were judged to be correct attempts at the novel word. On this measure of phonetic accuracy, there was a trend toward higher percentages for the children with TD ($M = 93.54$, $SD = 7.63$) than for the children with DLD ($M = 84.96$, $SD = 18.26$), though this difference was not significant, and no other differences were seen for this measure.

**Table 5.** The main effects model for meaning recall with the covariates included (24 participants and 96 repeated observations).

| Fixed effects | $b$ | 95% CI | | $b_{\text{std}}$ | $p$ |
|---|---|---|---|---|---|
| Group (DLD vs. TD) | −0.16 | −1.29 | 0.98 | −0.12 | .787 |
| Condition (MR vs. MS) | 0.42 | 0.03 | 0.81 | 0.32 | .037 |
| Time (1 wk vs. 5 min) | 0.08 | −0.25 | 0.42 | 0.06 | .624 |
| Covariates | | | | | |
| EVT-2 | 0.00 | −0.04 | 0.04 | 0.00 | .887 |
| Mother's education | −0.04 | −0.23 | 0.15 | −0.03 | .675 |
| Practice score | 0.99 | 0.22 | 1.77 | 0.77 | .012 |
| Intercept | 2.83 | −1.68 | 7.34 | | .219 |
| **Random effects** | $\sigma^2$ | | | | |
| Condition | 0.36 | 0.04 | 3.60 | | |
| Intercept | 0.41 | 0.10 | 1.69 | | |
| Level 1 residual | 0.64 | 0.36 | 1.15 | | |

*Note.* The unconditional model random effect variance was 0.77 [0.32, 1.87], and Level 1 residual variance was 0.91 [0.43, 1.93]. $b_{\text{std}}$ are partially standardized coefficients where the outcome is in standard deviation units. CI = confidence interval; DLD = children with developmental language disorder; TD = age-matched children with typical language development; MR = more retrieval/less study condition; MS = more study/less retrieval condition; 1 wk = recall test administered 1 week after the second learning session; 5 min = recall test administered 5 min after the second learning session; EVT-2 = Expressive Vocabulary Test, Second Edition.

## *Meaning Recall*

Table 5 shows the main effects model with covariates for meaning recall. There was a moderate effect for learning condition, where scores for the more retrieval/less study condition were 0.42 points higher than scores for the more study/less retrieval condition ($p = .037$). (Corresponding percentages correct were 93.40% and 87.15%, respectively.) There were no other main effects. Again, the covariate of practice score was statistically significant but did not influence the learning condition difference. (The main effects model without covariates also showed an effect for learning condition [$b_{\text{std}} = 0.29$, $p = .04$]. Although participant group

**Figure 1.** The (unconditional) mean word form recall scores showing differences according to learning condition and time. More retrieval = more retrieval/less study condition; More study = more study/less retrieval condition; 5 min = recall test administered 5 min after the second learning session; 1 week = recall test administered 1 week after the second learning session. Error bars are standard errors.

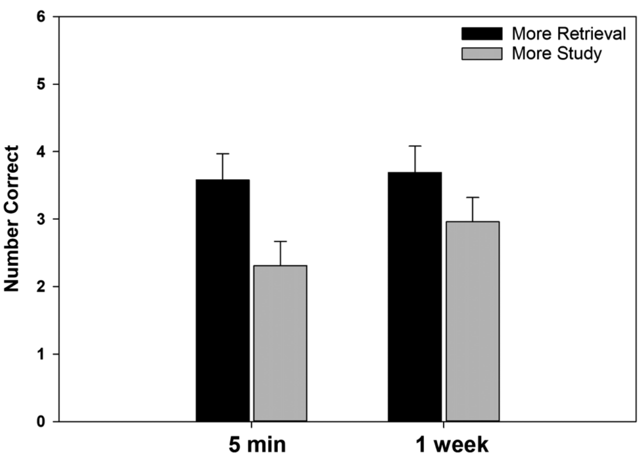


**Table 6.** The model for the group by learning condition interaction for meaning recall with the covariates included (24 participants and 96 repeated observations).

| Fixed effects | $b$ | 95% CI | | $b_{\text{std}}$ | $p$ |
|---|---|---|---|---|---|
| Group (DLD vs. TD) | −0.41 | −1.58 | 0.76 | | .493 |
| Condition (MR vs. MS) | 0.08 | −0.42 | 0.58 | | .762 |
| Time (1 wk vs. 5 min) | 0.08 | −0.26 | 0.43 | | .639 |
| Two-way interaction | | | | | |
| Group × Cond | 0.74 | −0.03 | 1.51 | 0.58 | .060 |
| Covariates | | | | | |
| EVT-2 | 0.00 | −0.04 | 0.04 | | .905 |
| Mother's education | −0.03 | −0.21 | 0.15 | | .744 |
| Practice score | 1.00 | 0.22 | 1.78 | | .012 |
| Intercept | 2.80 | −1.73 | 7.34 | | .226 |
| **Random effects** | $\sigma^2$ | | | | |
| Condition | 0.24 | 0.01 | 5.56 | | |
| Intercept | 0.36 | 0.09 | 1.55 | | |
| Level 1 residual | 0.67 | 0.11 | 4.07 | | |

*Note.* The unconditional model random effect variance was 0.77 [0.32, 1.87], and Level 1 residual variance was 0.91 [0.43, 1.93]. $b_{\text{std}}$ are partially standardized coefficients where the outcome is in standard deviation units. CI = confidence interval; DLD = children with developmental language disorder; TD = age-matched children with typical language development; MR = more retrieval/less study condition; MS = more study/less retrieval condition; 1 wk = recall test administered 1 week after the second learning session; 5 min = recall test administered 5 min after the second learning session; Cond = learning condition; EVT-2 = Expressive Vocabulary Test, Second Edition.

showed no effect in the model with covariates, there was an effect before covariates were applied [$b_{std} = 0.55$, $p = .01$].)

Although the main effects model with covariates was suitable, we also include the model shown in Table 6 (with covariates), which reveals that one of the two-way interactions—that of participant group by learning condition—showed a moderate to large effect ($b_{std} = 0.58$), in spite of $p = .060$. No other interactions were statistically significant. In Table 7, it can be seen that the participant group by learning condition effect was largely driven by the children with DLD. For the children with DLD, the mean percentage correct for the more retrieval/less study condition was 91%, whereas it was 79% for the more study/less retrieval condition. For the children with TD, these means were 96% and 94%, respectively. This pattern is illustrated in Figure 2.

### *Recognition*

Recognition scores were relatively high for both learning conditions (more retrieval/less study $M = 88\%$; more study/less retrieval $M = 85\%$), and no difference was found in any model. Likewise, learning condition did not interact with participant group. Although participant group showed a difference in the main effects model without covariates (TD $M = 92\%$; DLD $M = 80\%$; $b_{std} = 0.58$, $p = .01$), this effect was reduced and less statistically reliable in the model that included the covariates ($b_{std} = 0.43$, $p = .18$).

## Discussion

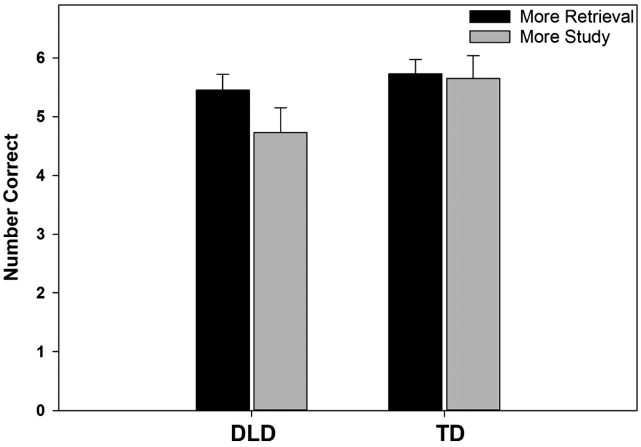
Before discussing the major questions behind this study, we note here those details of the results that corroborate the findings of our previous studies and those of other investigators. An especially reliable finding—replicated in this study—is that retention from 5-min testing to 1-week testing was extremely stable in both groups of children. This was seen in each of our three previous studies (Haebig et al., 2019; Leonard, Deevy, et al., 2019; Leonard, Karpicke, et al., 2019) and in the study of young adults by McGregor et al. (2017). Indeed, in this study, word form recall was actually better at the later time point. Because this stability has been seen in all learning conditions evaluated in these studies, we do not ascribe it to repeated spaced retrieval in particular. Rather, we believe it is a characteristic of the children—DLD and TD. Using procedures such as we have used, if children can recall the information immediately after the second learning session, they can retain it over the next week. We find no evidence of forgetting within that time frame.

**Table 7.** The simple effects for the group by learning condition interaction for meaning recall.

| Interaction | $b$ | 95% CI | | $b_{std}$ | $p$ |
|---|---|---|---|---|---|
| DLD vs. TD for MS condition | −0.41 | −1.58 | 0.76 | −0.32 | .493 |
| DLD vs. TD for MR condition | 0.33 | −0.59 | 1.26 | 0.26 | .481 |
| MR vs. MS for TD group | 0.08 | −0.42 | 0.58 | 0.06 | .762 |
| MR vs. MS for DLD group | 0.82 | 0.24 | 1.40 | 0.64 | .005 |

*Note.* $b_{std}$ are partially standardized coefficients where the outcome is in standard deviation units. CI = confidence interval; DLD = children with developmental language disorder; TD = age-matched children with typical language development; MR = more retrieval/less study condition; MS = more study/less retrieval condition.

**Figure 2.** The (unconditional) mean meaning recall scores showing differences according to learning condition for the children with developmental language disorder (DLD) but not the children with typical language development (TD). More retrieval = more retrieval/less study condition; More study = more study/less retrieval condition; Error bars are standard errors.



Another consistent finding corroborated in this study is that children can acquire and retain the "meanings" of novel words (e.g., likes rain) more readily than the word forms themselves (e.g., /fumi/). These meanings were arbitrarily assigned to the referents with no visual cues as to their association. However, the fact that the meanings were known words (e.g., rain, birds) probably made the association less difficult. Note especially that the association to be learned was that between the meaning and the visual referent. This association could be learned without knowing the word form.

A third finding consistent with our earlier work is that the learning condition effect size for meaning (see Table 5) was smaller than the learning condition effect size for word form (see Table 4). This could be due in part to the overall higher scores for meaning, constraining the magnitude of the differences between the two conditions.

Also consistent with much of our earlier work was the absence of a relationship between the children's standardized vocabulary test scores and their scores on the word form recall, meaning recall, and recognition tests. The one exception was seen in our earlier study on adjective learning where a $p$ level of .050 was seen (Leonard, Deevy, et al., 2019). There are probably several interacting reasons why this relationship is relatively weak. Vocabulary tests tend to reflect meaning-based knowledge accumulated over considerable stretches of time with no control over the degree to which performance is affected by degree of prior experience. In contrast, in our studies, children had a tightly controlled degree of exposure over a more concentrated time with equal emphasis on word form and meaning.

As in our previous studies, we expected generally higher scores for the children with TD than for the children with DLD, yet once again such differences were limited. For word form recall, no group differences were seen in any model. For meaning recall and recognition, differences favoring the children with TD were seen, but only before the covariates were applied. Why were there so few differences between these two groups of children?

We considered participant selection as one possible explanation for finding relatively few group differences. For example, it could be argued that, because our participants with DLD averaged within ± 1 *SD* of the mean on the standardized vocabulary tests, they represented a "mild" form of DLD, at least in vocabulary skill. This is possible. We did not use standardized vocabulary test scores as part of the selection criteria. Children were not required to score above or below a certain level on these tests to be included in the DLD group. In principle, children with DLD with much lower vocabulary test scores could have been included. Note also that our children with TD in this and other studies tended to average well above the mean on these standardized vocabulary tests. Accordingly, the high standardized vocabulary test scores of the children with TD might have been expected to ensure a group difference in recall even if the children with DLD showed standardized test scores in the average range.

Regarding word form recall, in our three previous studies, only a single comparison—of 5-min recall scores—revealed better recall by the children with TD than by the children with DLD (Haebig et al., 2019). We can only speculate as to why group differences in word form recall have been in such short supply. First, in all of our studies, the phonetic accuracy of words judged to have been recalled correctly has been somewhat greater for the children with TD. Thus, our word form recall measures might have reflected whether the children had sufficient recall of the correct word forms to attempt their production, but they were too imprecise to capture finer grain differences between the two groups of children.

Another possibility is that our inclusion of what each word's referent "likes" prompted some of the children to focus more on meaning than on form, and this tendency was more prevalent in the children with TD. Such a tendency might result in word form recall scores that underestimated these children's actual ability.

With regard to meaning recall and recognition, in one of our two previous studies employing these measures, the children with TD had higher scores than the children with DLD (Haebig et al., 2019). The design of that study did not require practice scores as covariates. In the other study employing these measures, our failure to find group differences appeared to be due to ceiling effects; scores were high for both groups of children (Leonard, Karpicke, et al., 2019). If we consider the role played by the practice scores in this study, our findings for meaning recall and recognition might be more in line with those of Haebig et al. (2019) than those of Leonard, Karpicke, et al. (2019). Because the major focus of this study was the comparison between more retrieval/less study and more study/less retrieval, it was crucial to place these two conditions on equal footing before the retrieval-only versus study-only phases began. The practice score covariate served this purpose quite well. However, this covariate also served to take into account a potential difference between the two groups of children prior to the experimental manipulation. Indeed, in the models without the covariate, the meaning recall and recognition scores of the children with TD were higher than those of the children with DLD. Our finding of no group difference was the result of the application of the covariates.

Despite the limited findings of group differences, we do not believe the two groups were similar in their word learning skills. First, studies by other investigators point to word encoding as an area of DLD vulnerability (Alt & Plante, 2006; Bishop & Hsu, 2015; McGregor et al., 2017). This aspect of word learning was not a focus of this study. The numerical difference we found between the two groups in phonetic accuracy suggests that the children with DLD had weaker encoding skills than their peers, but our principal measures were learning over 2 days and 1 week later—durations that may have permitted processes that were more intact in these children to have some compensatory effects. Second, the number of novel words to be learned was based on our pilot data, suggesting that two sets of three to six words each were sufficient to show differences according to learning condition. The number of words was not chosen with an eye toward finding differences according to participant group. If instead we had used, say, eight to 10 novel words per set, the increased number of words might have been a greater burden for children with DLD than for children with TD, resulting in significant group differences.

The principal finding of this study was that word learning and retention were facilitated when there were additional retrieval opportunities after earlier successful retrieval. Furthermore, this advantage occurred even when no additional study trials were presented; only the participants in the comparison condition were able to hear the words after the initial period. This represents a stringent test of the advantages of additional retrieval.

Although the more retrieval/less study condition resulted in better recall than the more study/less retrieval condition, we suspect that the magnitude of this difference would have been even greater if each child had successfully retrieved all words in the spaced retrieval trials during the initial criterion period. In the Karpicke and Roediger (2008) study with college student participants, all words were eventually retrieved successfully during the initial criterion period. In that study, 1-week recall scores were more than twice as high in the more retrieval/less study condition than in the more study/less retrieval condition. In our study, the words not retrieved successfully in spaced retrieval during the initial criterion period created an advantage for the more study/less retrieval condition. Specifically, words in the more retrieval/less study condition that had not been successfully retrieved in spaced retrieval trials during the initial criterion period were treated no differently than words that had been retrieved; they were never heard again. Some of these words were nevertheless correct during the subsequent retrieval-only trials, no doubt aided by the fact that they had been retrieved 4 times in the initial immediate retrieval trials. However, this was not true for other words in this condition. In contrast, words in the more study/less retrieval condition continued to be heard during the study-only trials even if they had never been successfully retrieved in a spaced retrieval trial during the initial criterion period. Note that our covariate could not control for this difference; it could only control for any differences between the conditions in the number of words correctly retrieved in the spaced retrieval trials during the initial criterion period.

The results of this study provided a possible explanation for an observation we have consistently made in our earlier studies (Haebig et al., 2019; Leonard, Deevy, et al., 2019; Leonard, Karpicke, et al., 2019). In those studies, we noted that words correctly retrieved early in the learning period were almost always recalled on the final tests. For words correctly retrieved only toward the end of the learning period, correct recall on the final tests was much less certain. This difference gave the impression that the final outcome hinged on whether a child “got it” from the very beginning. However, this view ignored the fact that words correctly recalled in early trials continued to appear in retrieval trials throughout the learning period. Therefore, they were retrieved much more frequently than the words that were correctly retrieved only toward the end of the learning period. It seems likely that the process of repeated retrieval further strengthened the earlier retrieved words, making their successful outcome even more likely.

This study served primarily as a “proof of concept.” Previous studies have compared repeated retrieval with repeated study, and in those studies, the repeated retrieval condition provided not only multiple retrieval opportunities but also the same number of study trials as the repeated study condition. In this study, we found that repeated retrieval can be relatively successful even with a reduction in the number of study trials. In fact, this was true even when the comparison (more study/less retrieval) condition provided children with some initial practice with spaced retrieval. We were especially encouraged that these findings applied to both groups of children. In fact, for meaning recall, the children with DLD appeared to benefit more from the more retrieval/less study condition than did the children with TD. These findings suggest that, despite their likely weaker lexical skills, children with DLD can learn new words with the same balance of retrieval and study opportunities as their typically developing peers.

Although these findings are encouraging, the procedures used were not ideal. We had hoped that successfully retrieving each word 4 times in initial immediate retrieval trials would create an equivalent starting point before the words branched into either the retrieval-only phase or the study-only phase. Knowing that spaced retrieval provides even greater benefits than immediate retrieval, we included these trials as well during the initial criterion period for added preparation. Unfortunately, success with the spaced retrieval trials during the initial criterion period was not uniform. Our use of the covariate statistically controlled for any differences between the two conditions in this regard. Nevertheless, in both conditions, there were words that were never successfully retrieved during the spaced retrieval trials of the initial criterion period, and these were the least likely to be recalled during the final tests. We need to develop procedures that ensure that all words can be retrieved in spaced retrieval at an earlier point in the learning process.

The findings of this study join those of earlier studies in providing motivation to continue careful study of the contributions of repeated spaced retrieval to children’s word learning. Through further refinement in the laboratory, we hope to reach the point of translating the procedures into more practical learning activities, with the eventual goal of providing children with a more effective and efficient means of learning new words.

## Acknowledgments

This research was supported in part by Research Grant R01 DC014708 awarded to Laurence B. Leonard.

## References

**Alt, M.** (2011). Phonological working memory impairments in children with specific language impairment: Where does the problem lie? *Journal of Communication Disorders, 44*(2), 173–185. https://doi.org/10.1016/j.jcomdis.2010.09.003

**Alt, M., & Plante, E.** (2006). Factors that influence lexical and semantic fast mapping of young children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 49*(5), 941–954. https://doi.org/10.1044/1092-4388(2006/068)

**Bishop, D. V. M., & Hsu, H. J.** (2015). The declarative system in children with specific language impairment: A comparison of meaningful and meaningless auditory–visual paired associate learning. *BMC Psychology, 3,* 3. https://doi.org/10.1186/s40359-015-0062-7

**Chen, Y., & Liu, H.-M.** (2014). Novel-word learning deficits in Mandarin-speaking preschool children with specific language impairments. *Research in Developmental Disabilities, 35*(1), 10–20. https://doi.org/10.1016/j.ridd.2013.10.010

**Dawson, J., Stout, C., Eyer, J., Tattersall, P., Fonkalsrud, J., & Croley, K.** (2005). *Structured Photographic Expressive Language Test–Preschool 2 (SPELT-P 2)*. Janelle Publications.

**Dunn, L., & Dunn, D.** (2007). *Peabody Picture Vocabulary Test–Fourth Edition (PPVT-4)* [Database record]. APA PsycTests. https://doi.org/10.1037/t15144-000

**Edwards, J., Beckham, M. E., & Munson, B.** (2004). The interaction between vocabulary size and phonotactic probability effects on children’s production accuracy and fluency in nonword repetition. *Journal of Speech, Language, and Hearing Research, 47*(2), 421–436. https://doi.org/10.1044/1092-4388(2004/034)

**Fazio, L., & Marsh, E. J.** (2019). Retrieval-based learning in children. *Current Directions in Psychological Science, 28*(2), 111–116. https://doi.org/10.1177/0963721418806673

**Goffman, L., & Leonard, J.** (2000). Growth of language skills in preschool children with specific language impairment: Implications for assessment and intervention. *American Journal of Speech-Language Pathology, 9*(2), 151–161. https://doi.org/10.1044/1058-0360.0902.151

**Gray, S.** (2003). Word learning by preschoolers with specific language impairment: What predicts success? *Journal of Speech, Language, and Hearing Research, 46*(1), 56–67. https://doi.org/10.1044/1092-4388(2003/005)

**Gray, S.** (2004). Word learning by preschoolers with specific language impairment: Predictors and poor learners. *Journal of Speech, Language, and Hearing Research, 47*(5), 1117–1132. https://doi.org/10.1044/1092-4388(2004/083)

**Gray, S., Plante, E., Vance, R., & Henrichsen, M.** (1999). The diagnostic accuracy of four vocabulary tests administered to preschool-age children. *Language, Speech, and Hearing Services in Schools, 30*(2), 196–206. https://doi.org/10.1044/0161-1461.3002.196

**Greenslade, K., Plante, E., & Vance, R.** (2009). The diagnostic accuracy and construct validity of the Structured Photographic Expressive Language Test–Preschool 2. *Language, Speech, and Hearing Services in Schools, 40*(2), 150–160. https://doi.org/10.1044/0161-1461(2008/07-0049)

**Haebig, E., Leonard, L., Deevy, P., Karpicke, J., Christ, S., Usler, E., Kueser, J. B., Souto, S., Krok, W., & Weber, C.** (2019). Retrieval-based word learning in young typically developing children and children with developmental language disorder. II: A comparison of retrieval schedules. *Journal of Speech, Language, and Hearing Research, 62*(4), 944–964. https://doi.org/10.1044/2018_JSLHR-L-18-0071

**Kan, P. F., & Windsor, J.** (2010). Word learning in children with primary language impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research, 53*(3), 739–756. https://doi.org/10.1044/1092-4388(2009/08-0248)

**Karpicke, J. D., Lehman, M., & Aue, W.** (2014). Retrieval-based learning: An episodic context account. In B. Ross (Ed.), *Psychology of learning and motivation* (Vol. 61, pp. 238–284). Elsevier.

**Karpicke, J. D., & Roediger, H. L., III.** (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(4), 704–719. https://doi.org/10.1037/0278-7393.33.4.704

**Karpicke, J. D., & Roediger, H. L., III.** (2008). The critical importance of retrieval for learning. *Science, 319*(5865), 966–968. https://doi.org/10.1126/science.1152408

**Kaufman, A., & Kaufman, N.** (2004). *Kaufman Assessment Battery for Children, Second Edition (KABC-II)*. AGS.

**Lee, L.** (1974). *Developmental sentence analysis*. Northwestern University Press.

**Leonard, L.** (2014). *Children with specific language impairment* (2nd ed.). MIT Press. https://doi.org/10.7551/mitpress/9152.001.0001

**Leonard, L., Deevy, P., Karpicke, J. D., Christ, S., Weber, C., Kueser, J., & Haebig, E.** (2019). Adjective learning in young typically developing children and children with developmental language disorder: A retrieval-based approach. *Journal of Speech, Language, and Hearing Research, 62*(12), 4433–4449. https://doi.org/10.1044/2019_JSLHR-L-19-0221

**Leonard, L., Karpicke, J., Deevy, P., Weber, C., Christ, S., Haebig, E., Souto, S., Kueser, J. B., & Krok, W.** (2019). Retrieval-based word learning in young typically developing children and children with developmental language disorder. I: The benefits of repeated retrieval. *Journal of Speech, Language, and Hearing Research, 62*(4), 944–964. https://doi.org/10.1044/2018_JSLHR-L-18-0070

**McGregor, K.** (2014). *Deficits in word form encoding characterize developmental learning disability*. Paper presented at the Symposium on Research in Child Language Disorders, Madison, WI, United States.

**McGregor, K., Berns, A., Owen, A., Michels, S., Duff, D., Bahnsen, A., & Lloyd, M.** (2012). Association between syntax and the lexicon among children with or without ASD and language impairment. *Journal of Autism and Developmental Disorders, 42,* 35–47. https://doi.org/10.1007/s10803-011-1210-4

**McGregor, K., Gordon, K., Eden, N., Arbisi-Kelm, T., & Oleson, J.** (2017). Encoding deficits impede word learning and memory in adults with developmental language disorders. *Journal of Speech, Language, and Hearing Research, 60*(10), 2891–2905. https://doi.org/10.1044/2017_JSLHR-L-17-0031

**McGregor, K., Licandro, U., Arenas, R., Eden, N., Stiles, D., Bean, A., & Walker, E.** (2013). Why words are hard for adults with developmental language impairments. *Journal of Speech, Language, and Hearing Research, 56*(6), 1845–1856. https://doi.org/10.1044/1092-4388(2013/12-0233)

**Rice, M., & Hoffman, L.** (2015). Predicting vocabulary growth in children with and without specific language impairment: A longitudinal study from 2;6 to 21 years of age. *Journal of Speech, Language, and Hearing Research, 58*(2), 345–359. https://doi.org/10.1044/2015_JSLHR-L-14-0150

**Roark, B., & Demuth, K.** (2000). Prosodic constraints and the learner's environment: A corpus study. In S. Howell, S. Fish, & T. Keith-Lucas (Eds.), *BUCLD 24: Proceedings of the 24th Annual Boston University Conference on Language Development* (pp. 597–608). Cascadilla Press.

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

**Rowland, C. A.** (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432–1463. https://doi.org/10.1037/a0037559

**Schopler, E., Van Bourgondien, M., Wellman, G., & Love, S.** (2010). *Childhood Autism Rating Scale–Second Edition*. Western Psychological Services.

**Souto, S., Leonard, L., & Deevy, P.** (2014). Identifying risk for specific language impairment with narrow and global measures of grammar. *Clinical Linguistics and Phonetics, 28*(10), 741–756. https://doi.org/10.3109/02699206.2014.893372

**Spaulding, T., Hosmer, S., & Schechtman, C.** (2013). Investigating the interchangeability and diagnostic utility of the PPVT-III and PPVT-IV for children with and without SLI. *International Journal of Speech-Language Pathology, 15*(5), 453–462. https://doi.org/10.3109/17549507.2012.762042

**Storkel, H., & Hoover, J.** (2010). An online calculator to compute phonotactic probability and neighborhood density on the basis of child corpora of spoken American English. *Behavior Research Methods, 42,* 497–506. https://doi.org/10.3758/BRM.42.2.497

**Whiffen, J., & Karpicke, J.** (2017). The role of episodic context in retrieval practice effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(7), 1036–1046. https://doi.org/10.1037/xlm0000379

**Williams, K. T.** (2007). *Expressive Vocabulary Test, Second Edition (EVT-2)* [Database record]. APA PsycTests. https://doi.org/10.1037/t15094-000

## Appendix

Sentence Repetition Task

Underlined segments reflect the principal phoneme of interest.
I like you.
Wear a hat.
See my nose?
A treasure map.
That's funny.
My mommy.
Fly a kite.
Use the soap.
Pop a bubble.
A spider has a web.
I have one dollar.
See my tummy?
I hear the geese.
Yuck, a bug.
Eat a cookie.
The pig is fat.
I have a pocket.
I hear the music.
Mommy wears makeup.