The contributions of immediate retrieval and spaced retrieval to word learning in preschoolers with developmental language disorder

Laurence B. Leonard [iD] [1], **Justin B. Kueser**[1], **Patricia Deevy**[1], **Eileen Haebig** [iD] [2], **Jeffrey D. Karpicke**[1] **and Christine Weber**[1]

[1]Purdue University, West Lafayette, IN, USA
[2]Louisiana State University, Baton Rouge, LA, USA

Abstract

Background and Aims: Children with developmental language disorder (DLD) benefit from word learning procedures that include a mix of immediate retrieval and spaced retrieval trials. In this study, we examine the relative contribution of these two types of retrieval.

Methods: We examine data from Haebig et al. (2019) in their study that compared an immediate retrieval condition and a condition of spaced retrieval that also included immediate retrieval trials. Participants were 4- and 5-year old children with DLD and same-age peers with typical language development. Each child learned novel (made-up) words referring to unusual plants and animals in both conditions. We examined the phonetic accuracy of the novel words used during the final learning trial and during recall tests 5 min and 1 week after learning.

Results: On the final learning trial, the children were more phonetically accurate in using the novel words learned in the immediate retrieval condition. However, recall tests after the learning trials revealed a decrease in accuracy, especially for the children with DLD. After one week, accuracy was much lower for words in the immediate retrieval condition than for words in the mixed spaced-plus-immediate retrieval condition. For words learned in the mixed spaced-plus-immediate retrieval condition, accuracy was very stable across time for both groups.

Conclusions: Immediate retrieval boosts the phonetic accuracy of new words in the short term but spaced retrieval promotes stability and increases the likelihood that short-term gains are maintained.

Implications: When novel word learning is assessed at the level of phonetic accuracy, children with DLD can show declines over time not characteristic of children with typical language development. Spaced retrieval procedures augmented by immediate retrieval opportunities during learning appear to prevent such declines, leading to longer-lasting gains.

Keywords

Developmental language disorder, specific language impairment, retrieval, word learning, word recall

Introduction

Many children with developmental language disorder (DLD) show significant deficits in word learning. Relative to their same-age peers with typical language development (TD), they know fewer words and, for the words they do know, their understanding of these words is relatively shallow (McGregor et al., 2013). These deficits are not easily attributable to the children's linguistic experience. Even when new words are taught in systematic ways, many children with DLD have only limited success (e.g., McGregor et al., 2021; Storkel et al., 2019). This is not a

Corresponding author:
Laurence B. Leonard, Department of Speech, Language, and Hearing Sciences, Purdue University, 715 Clinic Drive, Lyles-Porter Hall, West Lafayette, IN 47907 USA.
Email: xdxl@purdue.edu

short-term weakness; the vocabulary skills of children with DLD fall further behind those of their peers from preschool age through adolescence (Rice & Hoffman, 2015).

To gain a better understanding of the word learning deficits of children with DLD, many investigators have employed a paradigm in which the children are asked to learn a set of "novel" (made-up) words referring to unfamiliar referents, actions, or attributes. (In keeping with the current literature, we employ the descriptor "novel" to refer to such made-up words. An alternative such as "nonword" is often used for made-up phonological sequences presented with no referent, such as in "nonword repetition" tasks.) In general, novel word learning studies reveal that children with DLD require more exposures to the novel words than their age mates to reach the same learning criterion level (Kan & Windsor, 2010). Novel word studies often provide children with information about both the word form (e.g., "pibe" /paɪb/) and the meaning assigned to the novel word (e.g., "a plant that likes butterflies"). Although children with DLD can struggle with both form and meaning, learning word forms seems to be the most difficult (Gray, 2004; McGregor et al., 2017a). Even when words are recalled, these children's phonetic productions tend to be less precise and consistent than their peers with TD.

Most novel word learning studies have included measures of both the children's encoding and longer-term retention of the words. Encoding refers to the process of forming temporary representations of the word. With increasing experience with the word during the learning period, these representations can be gradually refined. This may occur for word forms (e.g., refined phonetic representations) and for word meanings (e.g., more precise, detailed semantic representations). Longer-term retention is assessed in these studies using tests of word form and meaning several days or one week after the learning period (e.g., Leonard et al., 2019b; McGregor et al., 2017b).

To date, evidence suggests that encoding is the weakest aspect of word learning in DLD (Bishop & Hsu, 2015; Gordon et al., 2020; Jackson et al., 2021; McGregor et al., 2017b). Once individuals with DLD show an ability to consistently recall a word in the short term, their ability to retain the word over longer stretches of time is less impaired (Leonard et al., 2021; McGregor et al., 2017b; McGregor et al., 2020).

Among the more recent novel word learning studies are those exploring the possible contributions of retrieval practice (Chen & Liu, 2014; Leonard et al., 2020; McGregor et al., 2017b). Thus far, findings from such studies seem to hold promise for application to the clinic (see Gordon, 2020 for a recent review). This work has borrowed two major principles from the memory literature in psychology. The first is the value of including repeated retrieval trials during the learning period. When learners are asked to retrieve material they have studied on multiple occasions throughout the study period, their retention of the material is dramatically better than when learners continue to study the material without including any attempts at retrieval (e.g., Roediger & Karpicke, 2006). The second principle is the benefit obtained by spacing the retrieval trials during study periods in a manner that makes retrieval effortful but usually successful. In many studies, this is done by inserting other items between successive retrieval attempts of a given item (e.g., placing trials for other words between the first and second time a given word has been seen). Relative to immediate retrieval (in which there are no intervening items between a retrieval trial and the preceding study trial), spaced retrieval usually leads to much greater long-term retention (e.g., Karpicke & Roediger, 2007).

The two principles of repeated retrieval and spaced retrieval have been combined in several studies of children's novel word learning by Leonard and his colleagues. These investigators found that preschool-age children with DLD and same-age typically developing peers learned and retained novel words referring to nouns (Leonard et al., 2019b) and adjectives (Leonard et al., 2019a) strikingly better when repeated spaced retrieval (RSR) trials were inserted during learning than when the children had the same exposure to the words during study trials without the opportunity for retrieval. This difference held not only during recall testing 5 min after the learning period but also 1 week later.

In the studies just described, the investigators increased the likelihood of the children's early success with RSR by including immediate retrieval trials at the beginning of learning and at least one point later during the learning period. An example of the RSR schedule used by Leonard et al. (2019b) appears in Figure 1. The sequence in the figure shows, for the novel word /nɛp/, one immediate retrieval trial (labelled "0" because there are no words that intervene between the study trial and the retrieval trial) and two spaced retrieval trials (labelled "3" because there are three intervening words between a retrieval trial and the preceding study trial for that word). This sequence occurred four times over a two-session learning period and thus consisted of a total of four immediate retrieval trials and eight spaced retrieval trials for each word. It can also be seen from Figure 1 that both immediate retrieval and spaced retrieval trials were followed by study trials. Insertion of these study trials ensured that the words in this condition were heard as often as the words in the repeated study condition. However, these study trials could also serve as feedback, as they provided the "correct" answer to the retrieval prompt. (The children were not explicitly told whether their retrieval attempt was correct.)

Although these studies provided strong support for the use of repeated retrieval, the *spacing* of retrieval was not
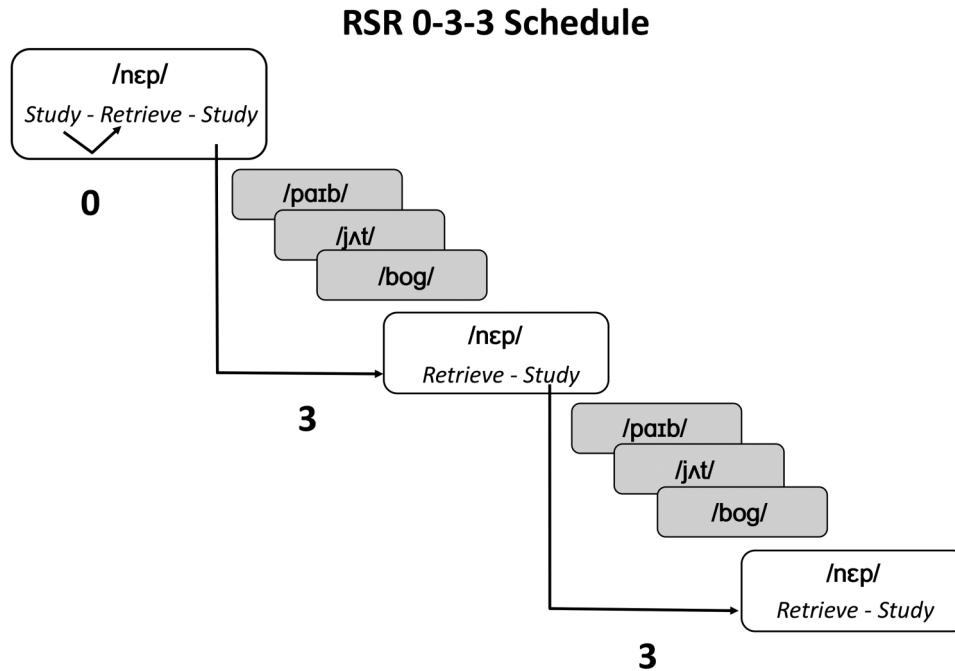
RSR 0-3-3 Schedule

Figure 1. An example of the first block of the learning period in the study of Leonard et al. (2019b). In this example, the novel word /nɛp/ is assigned to the repeated spaced retrieval (RSR) condition. In this block, /nɛp/ is retrieved in three instances. Retrieval is immediate in the first retrieval trial. This is designated "0" because there are no words intervening between the retrieval trial and the preceding study trial. For the second and third retrieval trials for /nɛp/, three other words intervened between the retrieval trial and the preceding study trial. For this reason, these retrieval trials are designated "3." Two other words in the sequence (/paɪb/ and /bog/) were in the repeated study condition and did not have retrieval trials. Three additional blocks, identical to the first were used for all words. Thus, each word in the RSR condition has four immediate retrieval trials and eight repeated spaced retrieval trials.

put to the test in these studies. In the comparison condition of repeated study, there were no retrieval trials of any kind. However, in a study by Haebig et al. (2019), spaced retrieval in particular was examined by comparing a condition that included both immediate and spaced retrieval trials with a condition consisting of immediate retrieval trials only. Specifically, across the two-session learning period, there were two immediate retrieval trials and four spaced retrieval trials for each word in the spaced retrieval condition, and six immediate retrieval trials for each word in the immediate retrieval condition. Children heard each word 24 times. A strength of this design was the fact that the two conditions provided not only equal exposure to the words (24 exposures per word) but also the same number of retrieval opportunities (6 retrieval opportunities per word). Only the spacing of the retrieval trials distinguished the two conditions. Haebig et al., found that better learning and retention was produced by the condition that included spacing; this advantage was seen at both five-minute and one-week recall testing, for both children with DLD and same-age peers with typical language development. Haebig et al. (2021) replicated this finding in another study of children with typical language skills.

One of the notable findings of the Haebig et al. (2019, 2021) studies is that the immediate retrieval condition

was not as successful even though, during the learning period, the children produced the novel words in the immediate retrieval condition more frequently than they produced the novel words in the spaced retrieval condition. Many of the children's retrieval attempts on spaced retrieval trials were unsuccessful, especially early in the learning period, whereas retrieval attempts on immediate retrieval trials were usually correct. It appeared that the reduced retrieval demands of the immediate retrieval condition were responsible for this difference. Along with providing the children with more production practice on the words, the immediate retrieval condition provided more opportunities for encoding. That is, the productions were made in a meaningful context – each immediate retrieval attempt was made directly after a study trial and in the presence of the novel word's referent. Furthermore, as immediate retrieval continued across the learning period, the phonetic representation of each novel word had the possibility of being gradually refined. The likelihood of the representation becoming stronger seems even more likely given that feedback was provided after each retrieval attempt.

Yet these multiple encoding opportunities were not sufficient to bring the children's later recall of these words up to the level seen for the words in the spaced retrieval condition. The most obvious explanation for this finding –

consistent with the broader memory literature – is that the more effortful retrieval required in the spaced retrieval condition enabled the words in this condition to be retained for a longer duration.

But two details in the spaced retrieval procedure used by Haebig et al. (2019, 2021) suggest that additional factors may have been at work in rendering spaced retrieval more effective. First, feedback was provided after retrieval attempts. Initial work has demonstrated that young children benefit from receiving feedback after retrieval prompts relative to no feedback and relative to learning conditions that include repeated study and elaborative learning strategies (Ma et al., 2020). Importantly, though, feedback is most effective as an aid to learning when retrieval attempts are unsuccessful (e.g., Rowland and DeLosh, 2015) or when retrieval is successful but the learner has low confidence in whether the attempt was, in fact, correct (Butler et al., 2008). As just noted, during the learning period in the Haebig et al., studies, retrieval was often unsuccessful in the spaced retrieval condition. In addition, it seems likely that in many instances, when the children did produce a correct response in the spaced retrieval condition, they were unsure of its accuracy until feedback confirmed that their response was, in fact, correct.

A second potentially important detail in the Haebig et al. (2019, 2021) studies was that immediate retrieval trials were included in the spaced retrieval condition. The assumption was that the early appearance of an immediate retrieval trial might result in early retrieval success, and any later appearances of an immediate retrieval trial might serve as a much-needed "refresher" for those words that were not yet being retrieved successfully. Although not numerous, these immediate retrieval trials might have enabled the children to produce a response that approximated the correct form, which, in turn, could have increased the likelihood of a best guess at the correct word during a subsequent spaced retrieval trial. In such a circumstance, feedback is especially effective. Of course, without frequent re-appearances of immediate retrieval trials in the spaced retrieval condition, there were fewer instances in which the child's retrieval attempt could closely follow a study trial. Therefore, the phonetic representation of the word might not have had as much opportunity to be refined. However, the spacing had the compensatory effect of creating a more stable representation that would endure for at least one week after the learning period.

This possible facilitating role played by immediate retrieval in the spaced retrieval condition receives support from a recent study by Kueser et al. (2021) that examined the trial-by-trial data from the spaced retrieval conditions in the studies by Haebig et al. (2019); Leonard et al. (2019a), and Leonard et al. (2019b). As noted earlier, in each of these studies, immediate retrieval trials were included along with the larger number of spaced retrieval trials. Kueser et al., found that the children's success on an immediate retrieval trial increased the likelihood of success on subsequent spaced retrieval trials. By themselves, successful immediate retrieval trials did not predict the children's final longer-term recall at 5 min or 1 week after learning; only success on spaced retrieval trials proved predictive. In that study, a retrieval attempt was deemed successful even if the word form produced was not a perfect match to the target form (e.g., /mɛp/ for /nɛp/). However, an unexplored possibility is that immediate retrieval might have had a mediating effect on longer-term recall by improving the fine-grained phonetic representation of the word which, in turn, could be rendered more stable through successful retrieval on spaced retrieval trials.

In the present study, we take a closer look at the relative contributions of immediate retrieval and spaced retrieval to phonetic accuracy. We do so by focusing on both the (repeated) immediate retrieval condition (hereafter, the IR condition) and the (repeated) spaced retrieval condition (hereafter, the RSR condition) in the study of Haebig et al. (2019) and by using a more discerning means of scoring accuracy. In each of the three studies already described – including that of Haebig et al., – a binary "correct-incorrect" scoring system was devised using as its basis the method developed by Edwards et al. (2004). In the Edwards et al., system, each consonant is awarded one point for each of place, manner, and voicing. Each vowel is credited with one point for each of backness, length, and height. An additional point is given if the child's production maintains the correct syllable shape (e.g., consonant-vowel-consonant; CVC).

In the three retrieval studies adopting this system, for a production to be scored as "correct," it had to subjectively seem like a true attempt at the novel word and the total points awarded the production had to be higher than the points that would be credited if the production was actually an attempt at one of the other novel words. For example, a fully accurate production of the novel word /pobɪk/ would earn 16 points $(3 + 3 + 3 + 3 + 3 + 1)$. If a child's apparent attempt at /pobɪk/ were produced as /topɪk/, the scored production would earn 14 points $(2 + 3 + 2 + 3 + 3 + 1)$. (The place error of the initial consonant and voicing error of the medial consonant reduced the score from 16 to 14). To ensure that /topɪk/ was more likely an attempt at /pobɪk/ than an attempt at one of the other novel words, the similarity of /topɪk/ to the other novel words would also be tested. For example, scoring /topɪk/ as an attempt at the novel word /kodəm/ would produce a total of only 9 points $(2 + 3 + 1 + 2 + 0 + 1)$. Assuming that tests of the similarity of other novel words to the child's production likewise resulted in lower scores, /topɪk/ would be scored as "correct" for the word /pobɪk/. To take another example, if a child appeared to be referring to the novel word /fumi/ with the production /fupi/, the score would be 11 $(3 + 3 + 1 + 3 + 1)$, with points deducted for the manner and voicing errors of the second consonant. If we

test an alternative assumption that /fupi/ was instead an attempt at the novel word /nɛp/, the resulting score would be 3, with only the second consonant matching (namely, /p/ in /fupi/ and /nɛp/). On the other hand, if the child produced /nup/ when the referent for /fumi/ was presented, the production would not pass the test as a likely attempt at /fumi/, as it would earn only 4 points (3 points for the first vowel and 1 point for the second consonant), whereas if taken as an attempt at the (wrong) novel word /nɛp/, points would be deducted only for the vowel, with full credit for both consonants plus the extra point for syllable shape for a total of 7 points. Thus, /nup/ would be scored as incorrect in the context of the referent for /fumi/. This system of scoring a production as "correct" or "incorrect" yielded very high inter-judge reliability and provided an objective means of testing subjective judgments of the novel words the children were actually attempting.

However, this binary adaptation did not take full advantage of the Edwards et al. (2004) system. Even when a child's multiple productions of the same word were all deemed "correct" by binary scoring, these productions might have differed in their phonetic precision. For example, according to the scoring system used, both /topɪk/ and /tobɪk/ would be scored as equally "correct" attempts at /pobɪk/, but /tobɪk/ would actually earn more points (15) than /topɪk/ (14). In the present study, we use as the dependent measure the actual point totals of the productions originally meeting the criteria as "correct." It was important to consider only those productions meeting the initial criteria of "correct" to avoid serious distortions in the data. In particular, we needed a basis for excluding productions such as /nup/ in response to the referent for /fumi/, which on both subjective and objective grounds could have been an attempt at a different novel word.

In previous studies with older individuals with DLD, the more detailed application of the Edwards et al., system has proven to be quite useful (Gordon et al., 2020; McGregor et al., 2017b). For example, Gordon et al., found that the phonological precision of the participants' word productions during the learning period was related to their precision on the same words on a free recall test administered 24 h later.

By adopting this level of scoring detail, we could pursue questions concerning the relative contributions of immediate retrieval and spaced retrieval that might not otherwise be possible. Our hypotheses follow from the Kueser et al. (2021) finding of apparent beneficial effects of immediate retrieval on subsequent spaced retrieval trials but not on longer-term recall. We hypothesize three outcomes, the first two of which, at first blush, may appear contradictory:

1. Phonetic accuracy at the end of the learning period on the final learning trial will be greater for words in the IR condition than for words in the RSR condition;

2. Phonetic accuracy will show greater decline from the end of the learning period to the 5-min testing point, and then again to the 1-week testing point, for words in the IR condition than for words in the RSR condition;
3. The decline described in (2) will be most apparent in the DLD group.

Our rationale for Hypothesis 1 is that immediate retrieval would seem to provide the most benefit to encoding and thus phonetic accuracy during the learning period itself. Because the IR condition contained the most immediate retrieval trials, there was more opportunity for phonetic encoding to continue to be enhanced and therefore by the final trial in the learning period words in this condition should show greater phonetic accuracy. Although words in the RSR condition also benefit from immediate retrieval trials, there were fewer of these trials, and final trials in this condition were always spaced retrieval trials. Further benefits to encoding would therefore be more limited.

Hypothesis 2 is based on the finding of Kueser et al. (2021) that success on immediate retrieval trials did not predict success on longer-term recall at 5 min or 1 week, even though it appeared to facilitate success on subsequent spaced retrieval trials during the learning period. This suggests that without the effortful retrieval involved in spaced retrieval, encoding effects on the phonetic aspects of a word might have a relatively short life span. When combined with effortful retrieval, the enhanced phonetic details might be more likely to be preserved. It is true, as noted in Hypothesis 1, that words in the RSR condition had fewer opportunities for phonetic refinement due to the inclusion of a smaller number of immediate retrieval trials. However, the greater number of opportunities for effortful retrieval – especially with feedback – should render those phonetic details that have been acquired more stable. As a result, words in this condition should show less decline from the final learning trials to the 5-min and 1-week tests.

Finally, Hypothesis 3 is founded on earlier studies that point to encoding as an especially weak aspect of word learning in individuals with DLD (Bishop & Hsu, 2015; McGregor et al., 2017a). Based on these earlier studies, we take as a given that these children's productions on the final trial of the learning period will be phonetically less accurate than those of their TD peers. However, even with this lower baseline, we anticipate that the phonetic details resulting from immediate retrieval will weaken further when the children with DLD must try to recall the word 5 min and, especially, 1 week later. This decline is likely to be greater than we see in the TD group.

It is important to note that the present study is concerned only with productions that were scored as "correct" using the original scoring system. As a point of reference, for the children with DLD an average of 3.50 words in the RSR condition and 1.31 words in the IR condition were recalled on both

the 5-min and 1-week test based on the original scoring. For the children in the TD group, an average of 5.38 words in the RSR condition and 2.69 words in the IR condition were recalled at 5-min testing; the corresponding averages for the 1-week test were 4.94 and 1.88 words recalled. From Haebig et al. (2019), we already know that highly inaccurate productions were not scored as correct, and there were many instances of "I don't know" or "I don't remember." All of these could have been instances of very partial phonetic representations or complete encoding failures. Thus, even though, using the original scoring standard, fewer words were correctly recalled in the IR condition than in the RSR condition, in the present study, we examine only those that met the original "correct" standard. We hypothesize that even in this case phonetic differences will emerge.

Methods

Participants

The data used in this study came from the novel word productions of the children in the study by Haebig et al. (2019). A summary of the participant characteristics appears in Table 1. Thirty-two children participated, 16 who met the selection criteria for DLD (10 girls, six boys, M age = 59.60 months) and 16 who met the criteria for showing typical language development (10 girls, six boys, M age = 61.58 months). All children passed a pure tone hearing screening in both ears at 20 dB at 500, 1000, 2000, and 4000 Hz using a calibrated portable Beltone Model 119 audiometer, and all scored above a nonverbal cognitive assessment standard score of 85 on the Kaufman Assessment Battery for Children – Second Edition (2004).

The children in the DLD group were receiving treatment for a language deficit or were scheduled for such treatment. Fourteen of these children were selected on the additional basis of standard scores below 87 on the Structured Photographic Expressive Language Test – Preschool 2

Table 1. Summary of participant characteristics.

Test/Measure	Participant Group	
	DLD	TD
Age (in months)	59.60 (4.43)	61.58 (5.16)
Maternal Education (in years)	15.50 (1.59)	16.63 (1.75)
SPELT-P2 (standard score)	78.69 (9.41)	113.06 (9.17)
K-ABC-2 (standard score)	101.88 (8.00)	115.81 (10.06)
PPVT-4 (standard score)	103.44 (9.91)	121.06 (12.47)

Note. Values are means (standard deviations). DLD = children with developmental language disorder; TD = children with typical language development; SPELT-P2 = Structured Photographic Language Test – Preschool Second Edition; K-ABC-2 = Kaufman Assessment Battery for Children – Second Edition; PPVT-4 = Peabody Picture Vocabulary Test – Fourth Edition.

(SPELT-P; Dawson et al., 2005), the cutoff yielding good sensitivity and specificity (Greenslade et al., 2009). The two remaining children scored 89 and were included on the basis of scoring below the 10th percentile on Developmental Sentence Scoring (Lee, 1974) from a language sample. The children in the TD group scored at least 100 on the SPELT-P2. All children in the DLD group scored in the "minimal to no symptoms" range on the Childhood Autism Rating Scale – Second Edition (Schopler et al., 2010). This test was not administered to the children in the TD group.

The Peabody Picture Vocabulary Test – Fourth Edition (Dunn & Dunn, 2007) was also administered to all children. Scores on this test served as a covariate in the analysis conducted by Haebig et al. (2019) and were not used as a selection criterion.

Given that the children likely differed in their phonological skills, they were also given a speech sound test consisting of real words that included all segments in the same word position and syllable shape as the novel words (e.g., *castle, saddle,* and *bottom* for /kodəm/). In the present study, the children's scores on this test were used as a covariate in the analysis. All procedures were approved by the Purdue University Institutional Review Board. The children gave their verbal assent and parents provided informed written consent.

Novel word learning procedure

The children learned 12 novel words, divided into two sets of six words. The novel words were /bog/, /nɛp/, /paɪb/, /jʌt/, /daɪbo/, /fumi/, /gine/, /tomə/, /kodəm/, /meləp/, /pobɪk/, and /tɛkət/. The two-syllable words had syllable-initial stress. The words in the two sets were balanced in terms of syllable shape (CVC, CVCV, CVCVC). Within each set, three of the words were assigned to the IR condition and three to the RSR condition (originally referred to as the "repeated retrieval with contextual reinstatement" or RRCR condition in Haebig et al.), again balanced according to syllable shape. The words in the sets and conditions were also matched on phonotactic probability (average biphone frequency) and neighbourhood density using the Storkel and Hoover (2010) child language corpora database. The words assigned to each condition were counterbalanced across children. The novel words served as the labels for colour photos of exotic plants and animals.

The photos and recorded stimuli were presented via computer. Within each set, four blocks were used, two containing the words in the IR condition and two blocks containing the words in the RSR condition. One block from each condition was presented on each of two consecutive days. The order of the blocked learning conditions was counterbalanced across children.

Both learning conditions employed both study trials and retrieval trials. In study trials, the child saw the photo and

heard both the novel name 3 times and what the referent "liked" once as in "This is a /pobɪk/. It's a /pobɪk/. A /pobɪk/ likes snow." In retrieval trials, the photo appeared and the child heard questions probing both the name of the referent and what it likes, as in "What's this called? What do we call this?" and "What does this one like? What does it like?" In the present study, we focus only on the children's retrieval and recall of the novel words, not the familiar words (e.g., "snow") representing what the plant or animal liked.

Each word, regardless of condition, had four study trials (each trial providing 3 exposures to the word) and three retrieval trials in each block. All retrieval trials were followed by a study trial for the same word that could serve as feedback, though the children were never told explicitly whether their retrieval attempt was correct.

The two conditions differed in the type of retrieval schedule used. For words in the IR condition, the retrieval trial always occurred directly after a study trial of the same word. Thus, there were six immediate retrieval trials for each novel word, three in each block. Haebig et al. (2019) referred to this schedule as a 0-0-0 schedule, representing each word's spacing pattern in each block. For words in the RSR condition, the first retrieval trial in each block occurred directly after the study trial for the same word (thus involving immediate retrieval). The remaining retrieval trials occurred after two other words had intervened between the retrieval trial and the previous study trial for the word to be retrieved (thus involving spaced retrieval). For this condition, then, there were two immediate retrieval trials (one in each block) and four spaced retrieval trials (two in each block) for each novel word. This schedule was referred to as a 0-2-2 schedule, reflecting the spacing pattern for the words in each block for this condition.

Recall testing occurred at the end of the second learning session, after a five-minute break. The recall test used the same photos and prompts ("What's this called? What do we call this?") as the retrieval trials. Each novel word was tested twice, with the second test item for the word appearing only after all words were tested a first time. One week later, the same recall test was repeated. Beginning the next week, the second set of novel words was introduced, following the same procedures as the first set.

Scoring

For all responses on the final learning trial and all recall test responses that were originally scored as "correct" by Haebig et al. (2019), we recorded the specific numerical score earned based on the Edwards et al. (2004) scoring system. As noted earlier, the Edwards et al., system awarded one point each for correct place, manner, and voicing of consonants, one point each for correct backness, height, and length of vowels, and one additional point for correct syllable shape (e.g., CVCV). By employing a

precise value for each word judged originally as "correct" we could make more fine-grained distinctions of phonetic accuracy. For example, a completely accurate rendition of the novel word /fumi/ was entered with a score of 13 ($3 + 3 + 3 + 3 + 1$), which could then be distinguished from a production such as /fubi/ – with a value of 12 ($3 + 3 + 2 + 3 + 1$) – even though /fubi/, like /fumi/, had met the original criteria for "correct."

Given our hypotheses, we entered the numerical score for each correct production occurring on the final trial of the learning period – which was an immediate retrieval trial for the IR condition and a spaced retrieval trial for the RSR condition. We also entered the numerical score for each correct production on the 5-min and 1-week recall tests. Note that both the 5-min test and the 1-week test included two test items for each word. For our mixed-effects modelling, the numerical score for each correct production was entered. Thus, if all of a child's productions of a word on the final learning trial, the 5-min test, and the 1-week test had been judged as correct, the numerical values for five different productions of the word would be included in the data to be analysed. For consistency, we refer to the two 5-min test items and the two 1-week test items as the "5-min test trials" and "1-week test trials," respectively. This terminology seems compatible with the fact that the test items were identical to the retrieval trials during the learning period in both the photos shown and the retrieval prompts given ("What do we call this? What's this called?").

The numerical values for each word production were then converted to a percentage correct. This was necessary because completely accurate productions of CVC, CVCV, and CVCVC words yielded different total scores, owing to the different number of segments included in these syllable shapes. Thus, a score of 15 for an attempt at the novel word /pobɪk/ would be 15/16 or 94% and a score of 12 for an attempt at the novel word /fumi/ would be 12/13 or 92%.

Analytical approach

We used mixed-effects linear regression to model children's phonetic accuracy during the final learning trial, the two 5-min test trials, and the two 1-week test trials. Phonetic accuracy percentages were arcsin-square-root transformed to correct for non-normality. Independent variables were Participant Group (TD/DLD), Learning Condition (IR/ RSR), and Trial Time (final trial/5-min/1-week), and their interactions. Scaled and centred percentage correct on the pre-experiment real-word speech sound test was also included as a covariate. We included the most complex random effects structure justified by our design that would converge to reduce the likelihood of Type I error (Barr et al., 2013). For our primary model, this structure included random intercepts of participant and word with a

random slope of condition for the participant intercept and random slopes of condition and group on the word intercept. The R-style model equation was Phonetic_Accuracy ~ Group * Condition * Trial_Time + Covariate_Percentage + (Condition | Participant) + (Group + Condition | Word). (In this equation, the symbol "~" separates the dependent variable on the left from the independent variables on the right.) As the trial time variable had three levels, we used the *afex* and *lme4* R packages (Bates et al., 2015; Lenth, 2019) to present the model in an ANOVA-style table with Type-3 tests. We also provide the full model output with coefficients and standard errors in Supplementary Materials. Post-hoc tests were conducted using the *emmeans* R package (Singmann et al., 2019). When we report post-hoc means from the model, we back-transform from the arcsin-square-root scale to the percentage point scale for clarity. In addition, when describing the results, we use the "%" symbol to denote changes in phonetic accuracy as absolute changes in percentage points of accuracy using the Edwards et al. (2004) system but not percentage change relative to a baseline.

Results

A summary of the children's phonetic accuracy during the final learning trial, the 5-min test trials, and the 1-week test trials is provided in Figure 2a-2b. Two participants, both in the group with TD, were excluded from the analysis because they were at ceiling on phonetic accuracy across words. There were 627 trials distributed across 30 participants and 12 words. The linear mixed-effects model predicting phonetic accuracy from the interaction of group, condition, and trial time is presented in Table 2. (The full model output is provided in Table S1a-S1b in Supplemental Materials.) The linear mixed-effects model was a significantly better fit to the data than a model with random effects only, $X^2(12) = 146.42$, $p < .001$. In the following sections, we characterize the main effects and significant interactions, focusing specifically on the comparisons related to our three hypotheses.

Main effects

There was a significant effect of the covariate – the accuracy score on the pre-experiment real-word speech production test. With each increase in 10% on the speech production test score there was an associated increase of 3.26% (i.e., 3.26 percentage points), $SE = 0.25\%$, in phonetic accuracy on the learning and test trials. The covariate was included in all analyses reported here; thus, all effects reported were apparent even after pre-experiment speech sound test scores were taken into account. There was a significant main effect of Participant Group; the group with TD scored 5.60 percentage points higher on the phonetic accuracy

measure than the group with DLD across trial times and learning conditions, $F(1, 29.16) = 6.10$, $p = .020$. There was also a significant main effect of Trial Time, $F(2, 570.23) = 36.46$, $p < .001$, that was characterized by significantly decreasing phonetic accuracy from the final learning trial, EMM [estimated marginal mean] $= 96.5\%$, $SE = 1.56\%$, to the 5-min test trials, $EMM = 92.1\%$, $SE = 2.32\%$, $t(574) = -5.59$, $p < .001$, and from the 5-min test trials to the 1-week test trials, $EMM = 88.7\%$, $SE = 2.77\%$, $t(563) = -3.01$, $p = .003$, across groups and learning conditions. Finally, the main effect of Learning Condition was not significant, $F(1, 16.81) = 0.23$, $p = .634$.

Interactions and hypotheses

In our first hypothesis, we predicted that the final trial of the learning period would show greater phonetic accuracy for words in the IR condition than for words in the RSR condition. A significant interaction between Trial Time and Condition showed that this was the case, $F(2, 571.96) = 30.21$, $p < .001$. Specifically, words learned in the IR condition had phonetic accuracy scores that were 7.18 percentage points larger than words learned in the RSR condition in the final learning trial, $t(17.7) = -2.23$, $p = .039$. The interaction between Trial Time and Condition also supported our second hypothesis that phonetic accuracy would decline more quickly for words in the IR condition than for words in the RSR condition. Table 3 shows differences between consecutive trial times across conditions. These differences show that phonetic accuracy from the final learning trial to the 1-week test trials declined for words in the IR condition, $EMM_{Final-Week} = 13.42\%$, $SE = 2.03\%$, $t(564) = 9.97$, $p < .001$, but remained stable for words in the RSR condition, $EMM_{Final-Week} = 0.64\%$, $SE = 1.32\%$, $t(546) = 0.49$, $p = .627$. Overall, words in the IR condition initially demonstrated better phonetic accuracy than words in the RSR condition but this accuracy quickly decreased over time. In fact, on the 1-week test

Table 2. Linear mixed-effects model predicting phonetic accuracy from the interaction of group, condition, and trial time.

Effect	df	F	p
Intercept	1, 16.36	959.63	<.001
Group	1, 29.16	6.10	.020
Condition	1, 16.81	0.23	.634
Trial time	2, 570.23	35.46	<.001
Phon. covariate	1, 29.62	11.98	.002
G × C	1, 27.14	0.13	.720
G × TT	2, 569.08	6.88	.001
C × TT	2, 571.96	30.21	<.001
G × C × TT	2, 567.22	6.56	.002

Note. G = Group (DLD/TD), C = Condition (IR/RSR), TT = Trial time (final learning trial/5-min test/1-week test). *P* values were calculated using the Kenward-Roger approximation for degrees of freedom.
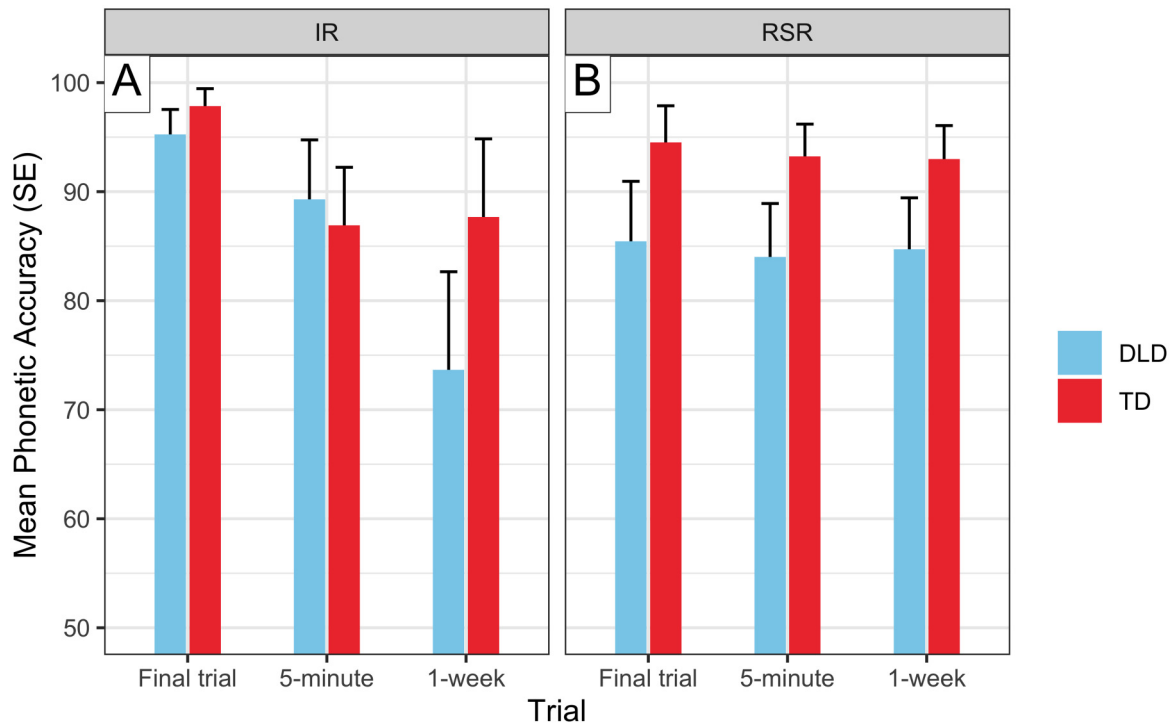
Figure 2. A. The mean phonetic accuracy of words in the immediate retrieval (IR) condition by the children with typical language development (TD) and the children with developmental language disorder (DLD) on the final learning trial, the test administered 5 min after the second learning session, and the test administered 1 week later. Error bars are standard errors. B. The mean phonetic accuracy of words in the repeated spaced retrieval (RSR) condition by the children with typical language development (TD) and the children with developmental language disorder (DLD) on the final learning trial, the test administered 5 min after the second learning session, and the test administered 1 week later. Error bars are standard errors.

trials, words in the RSR condition demonstrated slightly numerically higher phonetic accuracy than words in the IR condition, $EMM_{RSR-IR} = 5.60\%$, $SE = 5.31\%$, though this was not a significant difference, $t(20.1) = 0.97$, $p = .349$.

However, this overall pattern did not hold equally for both groups; there were significant interactions between Group and Trial Time, $F(2, 569.08) = 6.88$, $p = .001$, and between Group, Trial Time, and Condition, $F(2, 567.22) = 6.56$, $p = .002$. The interaction between Group and Trial Time showed that the difference in phonetic accuracy between groups grew over time. Specifically, the two groups demonstrated similar decrements in phonetic accuracy from the final learning trial to the 5-min test trials, $EMM_{\Delta DLD-\Delta TD} = -0.77\%$, $SE = 1.84\%$, $t(571) = -1.25$, $p = .213$, but the group with DLD demonstrated a significantly larger drop in phonetic accuracy from the 5-min test trials to the 1-week test trials, $EMM_{\Delta DLD-\Delta TD} = 9.31\%$, $SE = 2.49\%$, $t(561) = 3.68$, $p < .001$.

The interaction between Group, Trial Time, and Condition indicates that the drop in phonetic accuracy for the group with DLD was driven by words in the IR condition, supporting our third hypothesis that the effect of learning condition over time would be particularly pronounced for children with DLD. Table 4 shows the difference

between the RSR and IR learning conditions across trial times for each group. An examination of these interaction contrasts revealed two patterns. First, the group with DLD did not differ from the group with TD in the degree of change from the final learning trial to the 5-min test trials across the two learning conditions, $EMM_{\Delta DLD-\Delta TD} = 2.90\%$, $SE = 3.74\%$, $t(567) = 0.59$, $p = .556$. Put another way, both groups demonstrated a similar loss in phonetic accuracy from the final learning trial to the 5-min test trials and this loss was greater for words learned in the IR

Table 3. Comparisons of phonetic accuracy between consecutive trial times and conditions.

Condition	Trial Time	Estimate	SE	df	t	p
IR	Final − 5-min	6.56	1.24	553	7.39	<.001
	5-min − 1-week	6.86	2.10	566	3.43	.003
RSR	Final − 5-min	0.40	1.29	545	0.31	>.999
	5-min − 1-week	0.25	1.16	535	0.21	>.999

Note. Estimate refers to the difference in absolute percentage points from the first time point to the second time point in each condition. P values are corrected for multiple comparisons using the Holm method.

Table 4. Comparisons of phonetic accuracy between the RSR and IR learning conditions across trial times for the DLD and TD groups.

Group	Trial Time	Estimate	SE	df	t	p
DLD	Final − 5-min	7.93	2.92	575	3.91	<.001
	5-min − 1-week	15.50	4.03	562	3.93	<.001
TD	Final − 5-min	5.03	2.29	566	3.42	.001
	5-min − 1-week	−0.24	2.65	556	−0.07	.944

Note. Estimate refers to the difference in absolute percentage points. A positive number shows that the difference between the RSR and IR conditions increased across trial times. A negative number shows that the difference between conditions decreased. *P* values are corrected for multiple comparisons using the Holm method.

condition than in the RSR condition (compare Figure 2A and Figure 2B). This result contrasts with a second pattern. Here, the group with DLD demonstrated greater differences between learning conditions from the 5-min test trials to the 1-week test trials compared to the group with TD, $EMM_{\Delta DLD-\Delta TD} = 15.70\%$, $SE = 4.82\%$, $t(563) = 2.89$, $p = .008$. For the group with TD, the level of phonetic accuracy was stable from the 5-min test trials to the 1-week test trials for words in both the IR and RSR conditions; for the group with DLD, there was loss of phonetic accuracy from the 5-min test trials to the 1-week test trials for words in the IR condition (Figure 2A) but stability for words in the RSR condition (Figure 2B).

As noted earlier, the present study examined only those productions meeting the original criteria as "correct" in the Haebig et al. (2019) study and, in those data, more novel words by the children with TD were judged as "correct" than novel words by the children with DLD. This resulted in more data available for analysis from the TD group. However, there was nevertheless a sufficient number of "correct" productions from both groups to permit analysis. Specifically, all 16 children in the DLD group had "correct" responses on the final learning trial in both the RSR and IR condition. This was also true for 15 of the 16 children in the TD group. The remaining child had "correct" responses for the final learning trial in the RSR condition but not the IR condition. One of the 16 children in the DLD group failed to show "correct" recall of any of the novel words following the learning period; post-learning "correct" recall data were available for all other children. Nevertheless, the missing data allowed for the possibility that our results were providing a distorted picture of the real state of affairs. Although the mixed-effects models we employed are designed to handle missing data, as an extra step we ran a second model with only the data from words that were successfully recalled ("correct") by a child on all trials. These analyses appear in Supplementary Materials and largely confirmed the findings of our primary analyses.

Finally, in Supplementary Materials, we provide descriptive data showing the percentage of "correct"

productions that were judged as 100% phonetically accurate during the final learning trial, the 5-min recall trial, and the 1-week recall trial. The resulting pattern remained very much as in our original analysis. For the RSR condition, the percentage of "correct" productions with 100% phonetic accuracy in both participant groups changed relatively little across time. However, for the IR condition, the percentage of "correct" productions with 100% phonetic accuracy decreased from the final learning trial to the 5-min testing trial for both groups and continued to decline from 5 min to 1 week for the children with DLD.

Discussion

This study pursued three hypotheses concerning the relative contributions of immediate retrieval and spaced retrieval to the phonetic accuracy of novel words. All productions examined in this study met sufficient criteria to be judged by Haebig et al. (2019) to be true attempts at the correct word during testing 5 min and 1 week after the learning period. Our first hypothesis was that final trials of the learning period would show greater phonetic accuracy for words in the IR condition than for words in the RSR condition. Our rationale was that the high frequency of retrieval attempts occurring immediately after the study trial would allow encoding to be gradually refined, which should enhance the phonetic fidelity of the children's productions of words in this condition. Words in the RSR condition had fewer immediate retrieval trials, and the final trials were spaced retrieval attempts. Consequently, for the RSR condition, the phonetic accuracy of the final-trial productions tended to be based on fewer successful encoding opportunities and depended on whatever phonetic details of the words survived when retrieval occurred after other words had intervened.

The results were consistent with this first hypothesis. Although statistically reliable ($p = .039$), the absolute difference of 7.18 percentage points in accuracy between the IR and RSR conditions on the final learning trial was not large. On the other hand, given the fact that all the productions analysed in this study had already met the original criterion for "correct," we should not be surprised at the rather narrow (yet statistically reliable) differences. For example, the imprecise productions /bobɪk/ and /pobɪt/ as attempts at the novel word /pobɪk/ would each earn 15 points or a percentage correct of 94%.

Our second hypothesis was that declines from the final learning trial to the longer-term tests at 5 min and 1 week would be greater for words in the IR condition than for words in the RSR condition. This hypothesis, too, was supported by the data. Phonetic accuracy for the words in the RSR condition changed very little if at all from the final learning trial to the 1-week test trials. This was not true for words in the IR condition, which showed a decline across the same time period.

This finding might appear counterintuitive given the fact that words in the IR condition had more immediate retrieval trials than did the RSR condition. And, as noted above, phonetic accuracy was higher at the final learning trial for words in the IR condition. However, immediate retrieval placed minimal demands on memory. Although each production in this condition might have been optimized thanks to the close proximity of study and retrieval trials, the stability of these productions could not be taken for granted when the support of immediately preceding study trials was removed. Encoding, as we see it, refers to how well a phonetic representation "sticks" and becomes associated with a referent, not how well a phonetic form can be reproduced in ideal circumstances. We believe the productions arising from the RSR condition were a better reflection of the former. Whatever their accuracy, the productions in the RSR condition that did occur certainly stood the test of time; as can be seen in Figure 2, mean percentages correct at 5 min and 1 week never strayed from the mean percentages on the final learning trial by more than 2.5%.

These findings are important in documenting that spaced retrieval enables children to preserve whatever phonetic accuracy they achieved through encoding. This observation is in line with the oft-described facilitative effect of "effortful" retrieval. We believe there could be two quite specific factors contributing to this effect. First, when the children attempted to retrieve words during spaced retrieval trials, it is likely that some of those attempts were made with little confidence, even when the attempts proved to be correct. This is the scenario in which feedback is most beneficial (Butler et al., 2008). In contrast, much of the feedback provided after retrieval in the IR condition might have had less impact given that the general accuracy of the children's attempts were in little doubt. For words in the IR condition, most of the benefit probably arose from the close proximity between a study trial and the retrieval attempt, though, as already noted, this benefit did not have much staying power.

A second factor that could have contributed to the advantage of spaced retrieval was the inclusion of immediate retrieval trials within the RSR condition. It is in this context that immediate retrieval might be most useful. Recall that Kueser et al. (2021) found evidence that successful immediate retrieval trials led to greater success on subsequent spaced retrieval trials. In that study, the original correct-incorrect scoring system was used. We suspect that, in the present study with the more phonetically precise scoring system applied to productions already deemed "correct," accuracy levels were likewise boosted thanks to preceding immediate retrieval trials. In short, immediate retrieval might have facilitated phonetic accuracy in the short-term, and spaced retrieval promoted stability in these productions.

Our pursuit of the third hypothesis produced one of the most illuminating findings of this study. As can be seen in Figure 2, in the IR condition, the TD children showed some reduction in phonetic accuracy from the final learning trial to the 5-min mark but this level showed no further decline at the 1-week point. In contrast, the children with DLD continued to show a reduction in accuracy over time, with an especially steep decline from 5 min to 1 week. These findings comport with the third hypothesis.

We should note that the DLD group's significant declines in accuracy in the IR condition did not likely reflect a more general weakness in longer-term retention. In the RSR condition, the children with DLD were remarkably stable in phonetic accuracy from the final learning trial to the 1-week testing point. It is true that these children were less accurate than their typically developing age mates across the entire period, but this was expected given the known encoding weaknesses of this clinical population. (Weaknesses that, in this case, were still evident even when pre-experiment speech sound accuracy was used as a covariate.) More important here is the fact that the RSR stability seen across time was identical in the two groups. This finding indicates to us that the accuracy-preservation function of spaced retrieval operates as effectively in children with DLD as it does in children with typical language development.

Another important finding from this study stems directly from our use of the more precise scoring system for phonetic accuracy. In the earlier studies using the original correct-incorrect scoring system (Haebig et al., 2019; Leonard et al., 2019a; Leonard et al., 2019b), there was considerable stability observed for both the DLD and TD groups from 5 min to 1 week. This was true not only for the RSR condition but also for the comparison conditions (IR in Haebig et al., repeated study in Leonard, Deevy et al., and Leonard, Karpicke et al.,). That is, although the comparison conditions did not show as many words successfully recalled at testing as in the RSR conditions, for those words that met the criterion for accuracy, there were no differences between the 5-min and 1-week recall scores. Our current findings indicate that when a more precise phonetic accuracy measure is used, stability is not as rock-solid as originally assumed. Stability was, in fact, still seen for both groups for words learned in the RSR condition. However, for words in the IR condition, only the TD group maintained the same level of accuracy from 5 min to 1 week. The children with DLD, in contrast, showed a significant reduction in accuracy.

In the earlier studies, the apparent stability over time was used to argue that the word learning weaknesses of children with DLD should probably not be characterized as "forgetting." These children may not have acquired as many words as their peers, but for those words they could remember in the short-term, they managed to hold onto for at least one week. We now have data that require a modification of this description. We are not sure whether to characterize as "forgetting" the finer-grained slippage that we observed

in the DLD group. Clearly, though, these children's phonetic representations of the words in the IR condition were not sufficiently robust to prevent further deterioration beyond the 5-min mark. Opportunities for effortful retrieval seem to be one way to change the course of this decline.

The findings of the present study join those of other studies in suggesting that RSR might prove to be an effective component in teaching new words to children. The comparison (IR) condition in this study represented a more tightly controlled version of an activity that is generally viewed as helpful to children's word learning – hearing new words frequently and having the opportunity to produce them. Yet, RSR proved more beneficial.

We are also encouraged by the fact that children with typical language development also benefitted from RSR. Although they did not show the declines over time in the IR condition seen in the children with DLD, the TD children showed consistently better phonetic accuracy in both the 5-min and 1-week time trials for words in the RSR condition than for words in the IR condition. For this reason, RSR seems to be applicable to children in general and probably should not be seen as a procedure used only in remediation.

Earlier studies of novel word learning have reported that children with DLD usually require more exposures to the words to meet the same criterion levels seen in their typically developing peers (e.g., Alt, 2011; Gray et al., 2014). This may continue to be true even if RSR is incorporated into the procedures. However, we can also note that the facilitative effect of RSR does not require additional exposures of the words to be taught. Indeed, in the Haebig et al. (2019) study serving as the source of data here, even the number of retrieval opportunities was the same as in the comparison condition. Thus, RSR seems to constitute a relatively efficient means of improving word learning.

Yet, both in the proportion of novel words learned (Haebig et al., 2019; Leonard et al., 2019a; Leonard et al., 2019b), and in the phonetic accuracy of these novel words (the present study), RSR has only produced success on a relative level. Absolute levels of recall and phonetic accuracy are still far from ideal. For example, although words in the RSR condition showed higher phonetic accuracy at the 1-week point than words in the IR condition (see Figure 2), these represented only the words that met the original standards of "correct" in the Haebig et al., study.

Further refinement is needed to increase the effectiveness of RSR. To boost children's recall, the number of words taught in a set and the specific spaced retrieval schedule to employ are two of the factors to consider. It is also important to point out that, as noted earlier, regardless of condition, each word in the Haebig et al., study (2019) was heard only 24 times and there were only six retrieval opportunities per word. An increase in both exposures and retrieval opportunities might serve to improve

children's performance level. Another factor – and most relevant to the present investigation – is whether a different mix of immediate retrieval trials and spaced retrieval trials might result in higher levels of recall and phonetic accuracy. For example, it would be a significant enhancement to RSR if phonetic accuracy on the final learning trial were increased while still showing the stability across time seen in the present study. Because immediate retrieval increases the success of subsequent spaced retrieval trials (Kueser et al., 2021), an increase in immediate retrieval trials might have this effect. For now, this expectation is only speculative. The data used by Kueser et al., to show the supportive effects of immediate retrieval on spaced retrieval trials included the RSR condition of Haebig et al. (2019) that we also examined in the present study. However, to be more confident that immediate retrieval was facilitative in the way we propose here, investigators in future studies might compare a condition with a mix of immediate and spaced retrieval trials (as in the present RSR condition) with a condition consisting of spaced retrieval trials only.

Summary

The present study has shown that RSR has advantages beyond facilitating the learning and recall of words whose phonetic production can be judged as "close enough." Even when recall is assessed at a more phonetically refined level, the boost provided by RSR is readily apparent. The gains from RSR seem to be enhanced through the insertion of immediate retrieval opportunities that promote short-term improvements in accuracy that spaced retrieval can then solidify. For children with DLD, RSR may avert a significant decline in phonetic accuracy after 5-min testing. For their peers with typical language development, RSR produces advantages that are present from the final learning trial through 1-week testing. Although the learning and recall profiles are not the same in the two groups, the RSR benefits to both are clear.

Declaration of conflicting interests

Funding

ORCID iDs

Laurence B. Leonard (ID) https://orcid.org/0000-0002-9189-4438
Eileen Haebig (ID) https://orcid.org/0000-0001-8216-7063

Supplemental material

Supplemental material for this article is available online.

References

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

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

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*(1), 3. https://doi.org/10.1186/s40359-015-0062-7

Butler, A., Karpicke, J., & Roediger, H. (2008). Correcting metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 918–928. https://doi.org/10.1037/0278-7393.34.4.918

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, second edition. DeKalb*. Janelle Publications.

Dunn, L., & Dunn, D. (2007). *Peabody picture vocabulary test, fourth edition*. AGS/Pearson.

Edwards, J., Beckham, M., & 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)

Gordon, K. (2020). The advantages of retrieval-based and spaced practice: implications for word learning in clinical and educational contexts. *Language, Speech, and Hearing Services in Schools*, *51*(4), 955–965. https://doi.org/10.1044/2020_LSHSS-19i-00001

Gordon, K., McGregor, K., & Arbisi-Kelm, T. (2020). Optimising word learning in post-secondary students with developmental language disorder: The roles of retrieval difficulty and retrieval success during training. *International*

Journal of Speech-Language Pathology, *23*(4), 405–418. https://doi.org/10.1080/17549507.2020.1812719

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., Pittman, A., & Weinhold, J. (2014). Effect of phonological probability and neighborhood density on word learning configuration by preschoolers with typical language development and specific language impairment. *Journal of Speech, Language, and Hearing Research*, *57*(3), 1011–1025. https://doi.org/10.1044/2014_JSLHR-L-12-0282

Haebig, E., Leonard, L., Deevy, P., Karpicke, J., Christ, S., Usler, E., Kueser, J., 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), 932–943. https://doi.org/10.1044/2018_JSLHR-L-18-0071

Greenslade, K., Plante, E., & Vance, R. (2009). The diagnostic accuracy and construct validity of the Structural Photographic Expressive Language Test–Preschool: Second Edition. *Language, Speech, and Hearing Services in Schools*, *40*(2), 150–160.

Haebig, E., Leonard, L., Deevy, P., Schumaker, J., Karpicke, J., & Weber, C. (2021). The neural underpinnings of processing newly taught semantic information: The role of retrieval practice. *Journal of Speech, Language, and Hearing Research*, *64*(8), 3195–3211. https://doi.org/10.1044/2021_JSLHR-20-00485

Jackson, E., Leitão, S., Claessen, M., & Boyes, M. (2021). Word learning and verbal working memory in children with developmental language disorder. *Autism and Developmental Language Impairments*, *6*, 1–20. https://doi.org/10.1177/23969415211004109

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., & Roediger, H. (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

Kaufman, A., & Kaufman, N. (2004). *Kaufman assessment battery for children, second edition*. American Guidance Service.

Kueser, J., Leonard, L., Deevy, P., Haebig, E., & Karpicke, J. (2021). Word-learning trajectories influence long-term recall in children with developmental language disorder and typical development. *Journal of Communication Disorders*, *94*, 106160. https://doi.org/10.1016/j.jcomdis.2021.106160

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

Lenth, R. (2019). *emmeans: Estimated marginal means, aka least-squares means*. https://CRAN.R-project.org/package=emmeans.

Leonard, L., Christ, S., Deevy, P., Karpicke, J., Weber, C., Haebig, E., Kueser, J., Souto, S., & Krok, W. (2021). A multi-study examination of the role of repeated spaced retrieval in the word learning of children with developmental language disorder. *Journal of Neurodevelopmental Disorders*, *13*, 20.

Leonard, L., Deevy, P., Karpicke, J., Christ, S., Weber, C., & Kueser, J. (2020). After initial retrieval practice, more retrieval produces better retention than more study in the word learning of children with developmental language disorder. *Journal of Speech, Language, and Hearing Research*, *63*(August), 2763–2776. https://doi.org/10.1044/2019_JSLHR-L-19-0221

Leonard, L., Deevy, P., Karpicke, J., Christ, S., Weber, C., Kueser, J., & Haebig, E. (2019a). 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*, 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., & Krok, W. (2019b). 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*, 944–964. https://doi.org/10.1044/2018_JSLHR-L-18-0071

Ma, X., Li, T., Duzi, K., Li, Z., Ma, X., Li, Y., & Zhou, A. (2020). Retrieval practice promotes pictorial learning in children aged six to seven years. *Psychological Reports*, *123*(6), 2085–2100. https://doi.org/10.1177/0033294119856553

McGregor, K., Arbisi, T., & Eden, N. (2017a). The encoding of word forms into memory may be challenging for college students with developmental language impairment. *Journal of Speech, Language, and Hearing Research*, *19*(1), 43–57. https://doi.org/10.3109/17549507.2016.1159337

McGregor, K., Arbisi-Kelm, T., Eden, N., & Oleson, J. (2020). The word learning profile of adults with developmental language disorder. *Autism and Developmental Language Impairments*, *5*, 1–19. https://doi.org/10.1177/2396941519899311

McGregor, K., Gordon, K., Eden, N., Arbisi-Kelm, T., & Oleson, J. (2017b). 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., Oleson, J., Bahnsen, A., & Duff, D. (2013). Children with developmental language impairment have vocabulary deficits characterized by limited breadth and depth. *International Journal of Language and Communication Disorders*, *48*(3), 307–319. https://doi.org/10.1111/1460-6984.12008

McGregor, K., Owen Van Horne, A., Curran, M., Wagner Cook, S., & Cole, R. (2021). The challenge of rich vocabulary instruction for children with developmental language disorder. *Language, Speech, and Hearing Services in Schools*, *52*(2), 467–484. https://doi.org/10.1044/2020_LSHSS-20-00110

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

Roediger, H., & 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., & DeLosh, E. (2015). Mnemonic benefits of retrieval practice at short retention intervals. *Memory (Hove, England)*, *23*(3), 403–419. https://doi.org/10.1080/09658211.2014.889710

Schopler, E., Van Bourgondien, M., Wellman, G., & Love, R. (2010). *Childhood autism rating scale* (2nd ed). Western Psychological Services.

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2019). *afex: Analysis of factorial experiments*. https://CRAN.R-project.org/package=afex.

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*(2), 497–506. https://doi.org/10.3758/BRM.42.2.497

Storkel, H., Komesidou, R., Pezold, M., Pill, A., Fleming, K., & Romine, R. (2019). The impact of dose and dose frequency on word learning by kindergarten children with developmental language disorder during interactive book reading. *Language, Speech, and Hearing Services in Schools*, *50*(4), 518–539. https://doi.org/10.1044/2019_LSHSS-VOIA-18-0131