Retrieval Dynamics of In-the-Moment and Long-Term Statistical Word Learning

Haley A. Vlach (haleyvlach@ucla.edu)
Catherine M. Sandhofer (sandhof@psych.ucla.edu)
Department of Psychology, University of California, Los Angeles
1285 Franz Hall, Los Angeles, CA 90095 USA

Abstract

Theories of word learning have proposed several tools that children and adults use to reduce the difficulty of the word learning problem. However, we propose that reducing difficulty may be detrimental—difficulty may promote long-term word learning. In this study, we tested predictions of desirable difficulties and the retrieval effort hypothesis in a cross-situational word learning paradigm. Learners were presented with objects and labels in three conditions of learning (easy, medium, and difficult) and tested either immediately or one week later. Results revealed a counterintuitive pattern of performance—initially, participants in the easy condition had the highest performance. However, after a one week delay, participants in the medium condition had the highest performance. Participants' self-report of retrieval difficulty during learning is used to account for differences in performance over time. This work is discussed in terms of the implications for several fields of cognitive science: statistical learning, human memory, and language and cognitive development.

Keywords: word and category learning; statistical learning; cross-situational learning; language and cognitive development;

Introduction

Word learning has been described as one of the greatest challenges of cognitive development. The world offers learners a seemingly infinite number of word-to-world mappings yet children and adults appear to learn words with great ease (Quine, 1960). How do learners accomplish such a difficult task?

Theories of word learning have focused on tools that learners use to make the word learning problem easier. These tools range from basic cognitive processes of attention (e.g., Smith, 2000), to social cues (e.g., Tomasello & Barton, 1994), to early constraints, such as mutual exclusivity (e.g., Markman, 1989). Although theories of word learning propose different mechanisms and tools, all theories suggest that reducing difficulty of the task is beneficial for word learning.

In this study, we take a radically different perspective on word learning. We propose that the difficulty children and adults encounter during word mapping may be beneficial to the learning process. That is, a desirable amount of difficulty may promote long-term word learning. We examine how difficult learning conditions affect learners' in-the-moment and long-term statistical word mapping. We take a mechanistic approach by examining a basic cognitive process that contributes to different learning outcomes: retrieval difficulty during learning. Taken together, this work demonstrates the central role of retrieval processes in statistical learning and word mapping.

Memory. Research on human memory has long sought to discover the conditions of learning that (a) create and store a representation of knowledge and (b) produce a representation that can be recalled and accessed over extended periods of time. This body of research has revealed several learning conditions that promote long-term memory. For example, distributing practice (e.g., Bjork & Allen, 1970), varying the conditions of practice (e.g., Smith & Rothkopf, 1984), and reducing feedback to the learner (e.g., Schmidt, 1991) have all been shown to promote long-term memory.

These learning conditions support long-term memory because they introduce difficulty for learners while knowledge is being acquired (see Bjork, 1994, for a review). Although introducing difficulty during learning often deters immediate performance, retention tests reveal higher long-term performance (compared to easier learning conditions). Because of the long-term benefit of such difficulties, these conditions of learning are often termed 'desirable difficulties' of learning (e.g., Bjork, 1994). Consequently, memory research suggests that creating difficulty during learning promotes long-term performance.

An example of a desirable difficulty of learning is distributed practice, often termed the spacing effect (e.g., Vlach, Sandhofer, & Kornell, 2008). The spacing effect is the robust phenomenon whereby memory is enhanced when learning events are distributed across time (i.e., spaced), instead of being presented in immediate succession (i.e., massed). Because of the time between learning events, spaced learning creates greater opportunities for forgetting (e.g., Bjork & Allen, 1970). Consequently, retrieving previous learning events is more difficult. On the other hand, massed presentations prevent forgetting because presentations are in immediate succession, making retrieval of previous learning events easier. In fact, upon immediate testing, massed presentations lead to a greater amount of learning than spaced presentations. However, if a test is administered following a delay, a spaced presentation schedule will yield more learning than the massed presentation schedule (e.g., Vlach et al., 2008).

Word Learning. Word learning tasks differ from memory tasks. In memory tasks, learners are asked to store a specific piece of information and then retrieve that piece of information later. However, in word learning tasks, children and adults are required to aggregate and abstract across learning events in order to infer and/or generalize to a new experience.

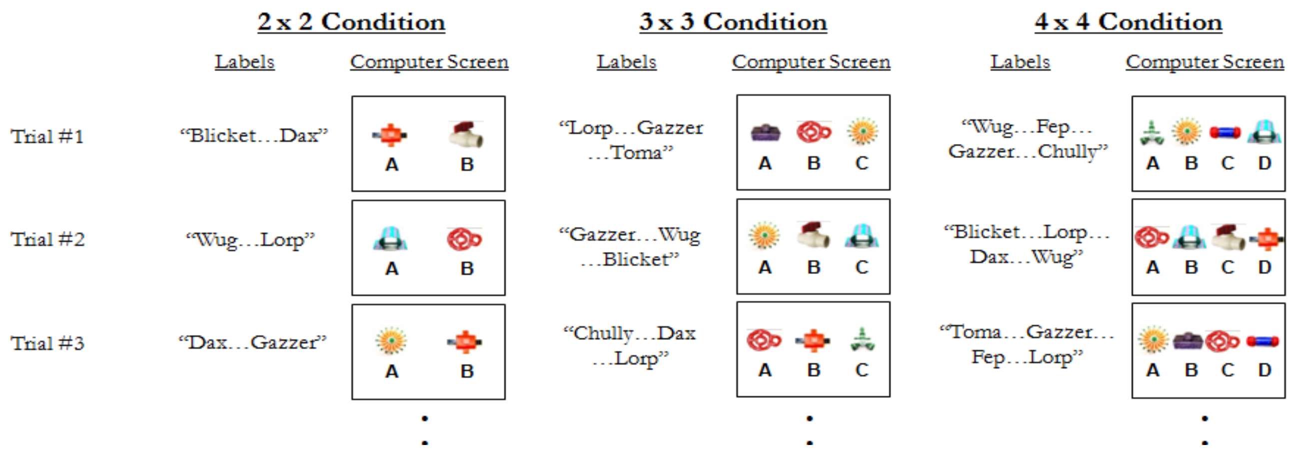

Figure 1. Example stimuli, for first three trials, from the different learning conditions: 2 x 2, 3 x 3, and 4 x 4.

Several researchers have long suggested that, although introducing difficulty during memory tasks is beneficial, these difficulties may be detrimental in more complex cognitive tasks (e.g., Gagne, 1950). For example, spaced learning was proposed to be particularly detrimental in generalization tasks. In fact, spaced learning was coined the “enemy of induction” (e.g., Gagne, 1950; see Kornell & Bjork, 2008, for a discussion). Despite speculations that desirable difficulties may be the “enemy of induction”, recent research suggests that imposing difficulty during learning promotes long-term word learning and generalization (e.g., Vlach et al., 2008).

***Why* is difficulty during learning beneficial?** One limitation of previous research on desirable difficulties in memory and word learning is that the mechanism(s) underlying performance differences are often poorly understood. That is, experiments are not commonly set up to determine and/or isolate the specific cognitive processes that cause higher long-term performance. Several theoretical frameworks, including desirable difficulties in learning, have suggested that more difficult (but eventually successful) retrieval is more beneficial for long-term learning than easier retrieval (e.g., Bjork, 1994). However, only within the last few years have researchers begun to test predictions of the retrieval effort hypothesis (see Pyc & Rawson, 2009, for a discussion).

Could the degree of difficulty in retrieving prior associations between word and objects promote long-term statistical word learning? From one perspective, difficulty in retrieving prior associations should deter statistical word learning because it would prevent learners from aggregating learning events together. If learners cannot aggregate learning events, they may not be able to determine which associations are more probable than others. On the other hand, memory mechanisms have been shown to promote word learning tasks (e.g., Vlach et al., 2008), and thus more effortful retrieval may support statistical word learning by promoting memory for associations.

The current study examined this question by presenting learners with a cross-situational word learning paradigm in three learning conditions: 2 x 2 (easy), 3 x 3 (medium), and 4 x 4 (hard). In the 2 x 2 condition, two objects and two words were presented in each learning trial, simultaneously. In the 3 x 3 condition, three objects and three labels were presented simultaneously. In the 4 x 4 condition, four objects and four labels were presented simultaneously. The conditions were classified as ‘easy’ (2 x 2), ‘medium’ (3 x 3), and ‘hard’ (4 x 4), based upon the amount of retrieval necessary to correctly map words to objects in one learning event. In the 2 x 2 condition, retrieving at least two prior associations is required to successfully map the words to objects in one learning event. In the 3 x 3 condition, retrieving at least three prior associations is required. In the 4 x 4 condition, retrieving at least four prior associations is required. During learning, participants were presented with a paper and pencil task. In this task, they were asked to report the objects for which they could successfully retrieve the corresponding label. These conditions and tasks allowed for a direct comparison of the effects of varying degrees of retrieval difficulty in statistical word learning.

Method

Participants

Participants were 64 undergraduates at University of California, Los Angeles. Participants received course credit for their participation.

Design

This study used a 3 x 2 design: Learning Condition (2 x 2, 3 x 3, and 4 x 4) and Testing Delay (Immediate or 1 Week Delay) were between-subjects factors. Participants were randomly assigned to one of the six conditions of the study.

Stimuli

Pictures of objects were presented on a 15-inch computer screen and the sound for the labels was presented through the computer’s speakers. As Figure 1 shows, the objects were pictures of novel objects. There were a total of 18 objects. The labels were novel words following the phonotactic probabilities of English (e.g., ‘blicket’). There were a total of 18 labels. Objects and labels were randomly paired together, for a total of 18 object-label pairs. In all conditions, there were a total of 6 presentations of each of the 18 object-label pairs. There were also an additional four objects and four labels presented during the training trial.

In the 2 x 2 condition, two objects and two words were presented in each learning trial, simultaneously. In the 3 x 3 condition, three objects and three labels were presented simultaneously. In the 4 x 4 condition, four objects and four labels were presented simultaneously. Figure 1 shows three learning trials of the experiment for each of the three conditions. It is important to note that the presentation order of objects and labels was randomized in each trial—the first word presented did not necessarily refer to the first object on the screen.

Because the same number of object-label pairs (18 pairs) were presented in each condition, the same number of times (6 presentations each), other presentation factors varied across conditions in order to ensure equivalent exposure to the object-label pairs. Table 1 outlines these variations, which were adapted from Yu and Smith (2007). Although the number of trials and time per trial varied, the total exposure time remained constant across the conditions.

Table 1. Three Learning Conditions

Condition	Number of Trials	Time per Trial (in secs)	Total Time (in secs)
2 x 2	54	6	324
3 x 3	36	9	324
4 x 4	27	12	324

Measures of Learning. In order to assess learners' in-the-moment and long-term performance, participants were given two pencil and paper tasks. The first task occurred during learning and was used to assess retrieval difficulty. Participants were given a worksheet in which they were asked to indicate the object(s) for which they had learned the corresponding label, on each trial. If they did not know the label for any of the objects, they were asked to circle 'None'. If they knew the label for the first object, they would circle 'A'. If they knew the labels for all of the objects, they would circle all of the letters. An example worksheet, for the 2 x 2 condition, is shown in Figure 2.

For each trial, circle A, B, A & B, or None.

Trial #			
1	A	B	None
2	A	B	None
3	A	B	None
4	A	B	None
⋮	⋮	⋮	⋮
54	A	B	None

Figure 2. An example of the worksheet used for the 2 x 2 condition. Participants self-reported the object(s) for which they could/could not retrieve the corresponding label.

For the 3 x 3 and 4 x 4 conditions, the number of letters corresponded to the number of objects and labels for each trial. For example, in the 4 x 4 condition, participants were given a worksheet that had 'A', 'B', 'C', 'D', and 'None' for each trial. In sum, this task assessed the quantity and times at which participants experienced success in retrieving labels for objects.

The second task was a final test designed to assess overall word learning performance. The final test consisted of four force-choice questions. Each question presented one label and asked participants to identify the corresponding object among four objects. The three foil objects were other objects used in the experiment. No one object was repeated in the tests. Thus, 16 of the 18 objects were used in the test.

Procedure

Participants were told that they would be shown children's toys and it was their job to figure out which word went with which toy. They were also instructed that it would be ambiguous as to which words went with which objects on each trial. Participants were then given a brief training exercise to demonstrate what the experiment would be like. The training consisted of three learning trials, each with two objects and two labels, immediately followed by a forced-choice test. Objects and labels used during training were not included during the rest of the experiment.

After the training trial, participants were informed that they would now be beginning the learning phase of the experiment. Participants were presented with learning trials according to the condition in which they were assigned (2 x 2, 3 x 3, or 4 x 4). The number and length of trials was also set according to the condition (see Table 1).

Participants were also instructed to complete a worksheet during the learning trials. Participants were asked to circle the object(s) for which they knew the corresponding label. If they did not know the corresponding labels for any of the objects, they were told to circle 'None'.

After viewing all of the learning trials and completing the worksheet, participants were given a final forced-choice test, depending upon the testing condition in which they were assigned. In the immediate condition, participants were given a test immediately following learning. In the one week delay condition, participants were asked to come back exactly 7 days after the learning session and complete the test.

Results

We asked whether difficulty would promote learners' long-term word learning in a cross-situational learning paradigm. If difficulty promoted word learning, we would expect to see lower performance immediately, but stronger performance long-term. However, if difficulty did not promote word learning, we would expect to see lower performance regardless of testing delay. Moreover, we wanted to examine if there were differences in learner's in-the-moment learning that could be contributing to differences in long-term performance. Specifically, we predicted that perhaps retrieval difficulty during learning might predict long-term performance.

Overall Word Learning Performance. We started our analysis by examining participant's overall word learning performance at the final test (see Figure 3). We conducted a 3 (Learning Condition) x 2 (Testing Delay) ANOVA, with the number of correct responses at the final test as the dependent measure. Results of this test revealed a significant main effect of learning condition, $F(2, 58) = 7.287$, $p = .002$, a significant main effect of testing delay, $F(1, 58) = 3.610$, $p = .052$, and a significant interaction of learning and testing delay, $F(2, 58) = 3.951$, $p = .025$.

In order to explore the interaction, we conducted two univariate ANOVAs, one in each testing condition. We then computed three planned comparisons using t-tests with Bonferroni corrections ($p < .05$) to determine the nature of the differences between learning conditions within each testing delay condition. If difficulty promoted word learning, we expected there to be differences in performance between learning conditions across the testing conditions.

In the immediate testing condition, there was a main effect of learning condition, $F(2, 31) = 6.270$, $p = .005$. Participants in the 2 x 2 condition had significantly higher performance than in the 4 x 4 condition, $p = .004$. Performance was also higher in the 2 x 2 condition than the 3 x 3 condition, $p = .047$. Finally, performance in the 3 x 3 condition was significantly higher than the 4 x 4 condition, p =.051. Thus, the greater the number of object-label pairings in each learning trial, the lower the performance.

However, there was a different pattern of results in the one week delay condition. There was a main effect of learning condition, $F(2, 27) = 4.925$, $p = .015$. Participants in the 3 x 3 condition had higher performance than both the 2 x 2 condition, $p = .046$, and 4 x 4 condition, $p = .016$. Participants in the 4 x 4 condition did not have significantly different performance than participants in the 2 x 2 condition, $p > .05$.

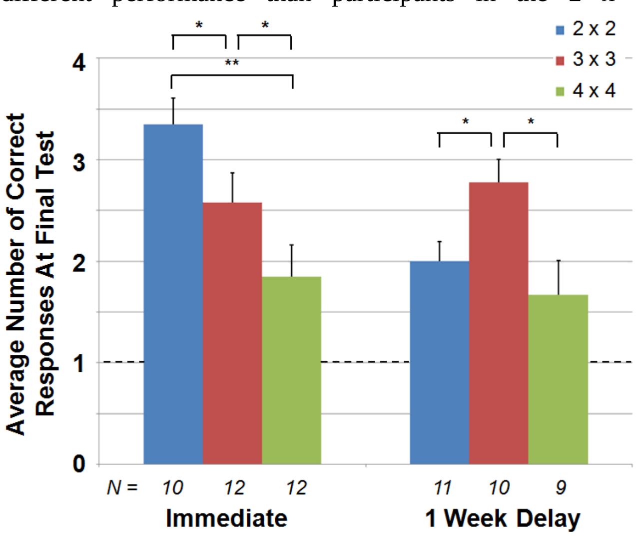

Figure 3. Average number of correct responses at final test, by condition. Dashed line represents chance performance. * = $p < .05$, ** = $p < .01$.

In sum, initially participants in the 3 x 3 condition had lower performance than the 2 x 2 condition. However, one week later, participants in the 3 x 3 condition had higher performance than participants in the 2 x 2 condition. This finding replicates that of previous research (e.g., Vlach & Sandhofer, 2010) and extends this work by demonstrating that retrieval difficult, but eventually successful retrieval can promote long-term learning.

In-the-Moment Word Learning Performance. We were particularly interested in determining if there were differences of in-the-moment learning that could be contributing to differences in long-term performance. Specifically, we thought that retrieval difficulty during learning could be a mechanism underlying performance. To explore this possibility, we analyzed participants' self-report of what they were retrieving. If there were differences in the number and timing of retrieval successes, this could be contributing to differences in performance.

We started by dividing learning into nine blocks of time, 36 seconds each. We chose this timescale because, over 36 seconds, participants in all of the conditions were exposed to the same number of object-label pairings. For example, in the 2 x 2 condition, there were 6 trials with 2 object-label pairings, for a total of 12 object-label pairings. In the 3 x 3 condition, there were 4 trials with 3 object-label pairings, for a total of 12 object-label pairings. Finally, in the 4 x 4 condition, there were 3 trials with 4 object-label pairings each, for a total of 12-object label pairings.

After dividing learning into nine timescales (36s), we then counted the number of objects that the participants reported as successfully retrieving the corresponding labels. To do this, we counted the number of objects (i.e., letters on the worksheet, see Figure 2) participants circled between the start of the experiment and the time point, resulting in a cumulative number of successful retrievals. For example, suppose a participant circled 2 objects between the time 0 and time 1, and 4 objects between time 1 and 2. The resulting values would be 0 for time point 0, 2 for time point 1, and 6 (i.e., 2+4) for time point 2.

We then computed a mixed 3 (Learning Condition) x 9 (Time Point) ANOVA, with learning condition as a between-subjects variable and time point as a within-subjects variable. Results of this test revealed a main effect of learning condition, $F(2, 61) = 16.876$, $p < .001$, a main effect of time point, $F(4, 488) = 277.917$, $p < .001$, and a significant interaction of learning condition and time point, $F(16, 488) = 14.004$, $p < .001$ (see Figure 4, next page).

In order to explore the interaction, we conducted nine univariate ANOVAs, one at each time point. We then computed planned comparisons using t-tests with Bonferroni corrections ($p < .05$) to determine the nature of the differences between learning conditions within each time point. If retrieval was difficult initially but then successful, we expected there to be differences in performance between learning conditions across the different time points.

At every time point, there was a main effect of learning condition (all *ps* < .001). There were also no floor or ceiling effects. However, across time points there were differences in which learning conditions were significantly different from one another. In time points 1-3, participants in the 2 x 2 condition reported successfully retrieving more pairings than participants in the 3 x 3 and 4 x 4 conditions (which did not significantly differ from each other, *ps* > .05). At time points 4-5, all learning conditions were significantly different from one another, *ps* < .05. Finally, in time points 7-9, participants in the 4 x 4 condition reported retrieving significantly less pairings than participants in the 2 x 2 and 3 x 3 conditions (which did not significantly differ from each other, *ps* > .05).

In sum, participants in the 3 x 3 condition reported significantly fewer retrieval successes initially (compared to participants in the 2 x 2 condition). This suggests that, at first, retrieval was difficult for participants in the 3 x 3 condition. However, by the end of the experiment, participants in the 3 x 3 conditions were more successfully retrieving pairings of objects and words (compared to participants in the 4 x 4 condition). This pattern of retrievals could explain both the lower immediate performance and the higher long-term performance of participants in the 3 x 3 condition.

Accuracy of Self-Report. As a final analysis of in-the-moment word learning, we examined the accuracy of participants' self-report. If participants were accurately reporting what they were retrieving, participants that reported successfully retrieving more should have higher final test performance. Thus, we analyzed the relationship between total number of retrieval successes during learning

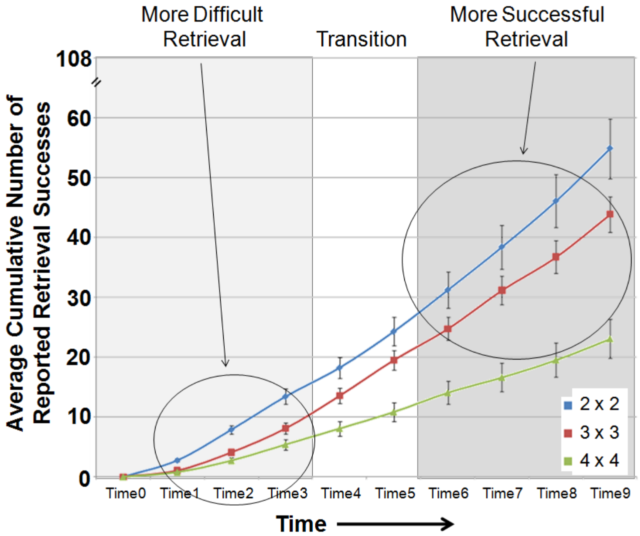

Figure 4. Average cumulative number of reported retrieval successes, by time point, for the three conditions of learning. Participants in the 3 x 3 condition started with a period of more difficult retrieval and ended with a period of more successful retrieval, relative to the other conditions.

and final test performance using Pearson's r. There was a significant correlation for both participants in the immediate condition, $r(32) = .441$, $p = .005$, and participants in the one week delay condition, $r(28) = .449$, $p = .041$.

Discussion

At the immediate test, participants in the 2 x 2 condition had higher performance than participants in the 3 x 3 and 4 x 4 conditions. However, at the one week delayed test, participants in the 3 x 3 condition had higher performance than participants in the 2 x 2 and 4 x 4 conditions. This finding replicates previous research demonstrating that difficult learning conditions often deter immediate word learning performance but promote long-term performance, relative to easier conditions of learning (e.g., Vlach & Sandhofer, 2010). This work extends this work by pinpointing a mechanism, retrieval difficulty, which may be contributing to differences in long-term performance.

Why did participants in the 3 x 3 condition have lower performance initially but higher performance at the one week delayed test? We predicted that perhaps retrieval difficulty during learning could account for differences in long-term performance. Specifically, we predicted that more difficult, but eventually successful, retrieval would promote long-term performance (often termed the retrieval effort hypothesis, see Pyc & Rawson, 2009).

Our analysis of participants' self-reported retrievals during learning revealed that participants in the 3 x 3 condition had a particularly unique pattern of retrievals, compared to participants in the 2 x 2 and 4 x 4 conditions. During the initial part of learning, participants in the 3 x 3 condition reported a lower number of retrieval successes, suggesting that retrieval was difficult. However, in the middle portion of learning, participants in the 3 x 3 condition reported significantly more retrieval successes than participants in the 4 x 4 condition, but significantly fewer than participants in the 2 x 2 condition. Finally, by the last portion of learning, the number of retrieval successes by participants in the 3 x 3 condition did not significantly differ from participants in the 2 x 2 condition. This suggests that, by the last portion of learning, participants in the 3 x 3 condition were experiencing more successful retrieval.

This work bridges several fields of research, including statistical learning, human memory, and language and cognitive development. As such, we have outlined the different contributions this work to each field of research.

Implications for Theories of Statistical Learning. This work highlights the importance of retrieval processes in statistical learning. In any given moment, a learner must retrieve potential associations from the past in order to aggregate them with the information currently being learned. Moreover, at a later point in time, learners need to retrieve prior associations in order to make statistical inferences. Whether it be during learning or at test, retrieval is central in statistical learning.

On a more global level, this work suggests a powerful role of memory in statistical learning. Previous research has attributed word learning to both basic cognitive processes of attention (e.g., Yu & Smith, 2007) and more complex algorithms and forms of inference making (e.g., Xu & Tenenbaum, 2007). However, this work suggests that memory processes, another basic cognitive process, should be added to the list of mechanisms driving word learning.

Implications for Theories of Human Memory. As noted in the introduction, one limitation of memory experiments is that they are not commonly set up to determine and/or isolate the specific cognitive processes that cause higher long-term performance in difficult learning conditions. However, the current study was designed to demonstrate the retrieval difficulty during learning could be one mechanism that promotes long-term performance. Consequently, this work contributes to recent research supporting the retrieval effort hypothesis (e.g., Pyc & Rawson, 2009).

This work also expands the retrieval effort hypothesis and the desirable difficulty framework by suggesting that not all retrieval attempts need to be successful in order for there to be a benefit for long-term performance. Indeed, participants in the 3 x 3 condition had a large number of retrieval failures before retrieval successes. Future work should continue to isolate the conditions under which more difficult retrieval is beneficial for long-term learning.

Implications for Models of Learning & Development. This work contributes to the idea that the *difficulty* of the word learning problem may promote language development. Although many tools may reduce difficulty and promote immediate performance, this may be detrimental long-term. Successful word learning is likely to be an optimal combination of the factors that create and reduce difficulty—future work should pursue this framework in order to determine the conditions that optimize learning.

This work also highlights that learning and development occur over time. Exploring in-the-moment learning is essential for our understanding of (a) how information is initially encoded and (b) how learners' prior experience is brought to bear on the moment at hand. The vast majority of research on cognitive development has focused on children and adults' learning at one moment in time.

However, in some theories of development (especially theories of language development), the common assumption is that performance at one moment will reflect performance at a later time. This study clearly demonstrates that this is not always the case. As such, generalizing results from one time point to another is dangerous.

In order to account for real-world learning, research should incorporate learning and testing over longer time-scales—over the course of weeks, months, and years. A complete theory of development not only accounts for learning in the moment and on each time scale, but also integrates them in order to understand how they influence each other over time.

Acknowledgments

We thank the undergraduate research assistants of the Language and Cognitive Development Lab at UCLA for their assistance in collecting data: Livon Ghermezi, Crystal Holloway, Nicole Lew, Chandni Parikh, and Justina Taban.

References

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalf and A. P. Shimura (Eds.), *Metacognition: Knowing about knowing*. Cambridge, MA: MIT Press.

Bjork, R. A., & Allen, T. W. (1970). The spacing effect: Consolidation or differential encoding? *Journal of Verbal Learning and Verbal Behavior, 9*, 567-572.

Gagne, R. M. (1950). The effect of sequence of presentation of similar items on the learning of paired-associates. *Journal of Experimental Psychology, 40*, 61-73.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science, 19*, 585-592.

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction*. Cambridge, MA: MIT Press.

Pyc, M. A., Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60*, 437-447.

Quine, W.V.O. (1960). *Word and object*. Cambridge, MA: MIT Press.

Schmidt, R. A. (1991). Frequent augmented feedback can degrade learning: Evidence and interpretations. In G. E. Stelmach & J. Requin (Eds.), *Tutorials in motor neuroscience*. Dordrecht: Kluwer.

Smith, L. B. (2000). Learning how to learn words: An associative crane. In *Becoming a Word Learner*. New York: Oxford University Press.

Smith, S. M., & Rothkopf, E. Z. (1984). Contextual enrichment and distribution of practice in the classroom. *Cognition and Instruction, 1*, 341-358.

Tomasello, M., & Barton, M. (1994). Learning words in non-ostensive contexts. *Developmental Psychology, 30*, 639-650.

Vlach, H. A., & Sandhofer, C. M. (2010). Desirable difficulties in cross-situational word learning. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.

Vlach, H. A., Sandhofer, C. M., & Kornell, N. (2008). The spacing effect in children's memory and category induction. *Cognition, 109*, 163-167.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review, 114*, 245-272.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science, 18*, 414-420.