



Published in final edited form as:

J Speech Lang Hear Res. 2014 October 1; 57(5): 1708–1721. doi:10.1044/2014_JSLHR-L-13-0150.

On-line learning from input versus off-line memory evolution in adult word learning: Effects of neighborhood density and phonologically-related practice

Holly L. Storkel, Daniel E. Bontempo, and Natalie S. Pak

University of Kansas

Abstract

Purpose—This study investigates adult word learning to determine how neighborhood density and practice across phonologically-related training sets influence on-line learning from input during training versus off-line memory evolution during no-training gaps.

Method—Sixty-one adults were randomly assigned to learn low or high density nonwords. Within each density condition, participants were trained on one set of words and then were trained on a second set of words, consisting of phonological neighbors of the first set. Learning was measured in a picture-naming test. Data were analyzed using multilevel modeling and spline regression.

Results—Steep learning during input was observed, with new words from dense neighborhoods and new words that were neighbors of recently learned words (i.e., second set words) being learned better than other words. In terms of memory evolution, large and significant forgetting was observed during 1-week gaps in training. Effects of density and practice during memory evolution were opposite of those during input. Specifically, forgetting was greater for high density and second set words than for low density and first set words.

Conclusion—High phonological similarity, regardless of source (i.e., known words or recent training), appears to facilitate on-line learning from input but seems to impede off-line memory evolution.

Keywords

vocabulary; word learning; adults; neighborhood density; input; memory evolution

People typically learn new words through exposure. The prevailing assumption is that words are learned from the input that a person hears. That is, learning a word is, at least partly, a function of experiences with the word. However, there is long-standing evidence that learning words can occur in the absence of additional input (Gaskell & Dumay, 2003; Leach & Samuel, 2007; Rice, Oetting, Marquis, Bode, & Pae, 1994; Storkel, 2001, 2003; Storkel & Lee, 2011). Specifically, past studies show that word learning performance may continue to improve after training is withdrawn. This pattern of results suggests that there are (at least)

two mechanisms that underlie adult word learning: on-line learning from input as well as “off-line” learning in the absence of input (Davis, Betta, Macdonald, & Gaskell, 2008; Davis & Gaskell, 2009; Dumay & Gaskell, 2007; Gaskell & Dumay, 2003; Leach & Samuel, 2007; Tamminen & Gaskell, 2008; Tamminen, Payne, Stickgold, Wamsley, & Gaskell, 2010). This two mechanism framework fits well with current theories of learning and memory. In particular, theories of learning and memory differentiate between change associated with swift episodic learning from input and change associated with slower memory consolidation during sleep, when it is thought that newly learned information is stabilized and integrated with existing information (McClelland, McNaughton, & O'Reilly, 1995; Norman & O'Reilly, 2003; O'Reilly & Rudy, 2000). Moreover, these theories suggest distinct neuroanatomical structures for each mechanism with the hippocampus being crucial for learning from input and the neocortex being critical for memory consolidation (McClelland et al., 1995; Norman & O'Reilly, 2003; O'Reilly & Rudy, 2000).

The discussion of memory consolidation has typically emphasized the benefits for word learning performance, creating a rather positive view of retention during gaps without training. Often overlooked in this discussion is a more negative aspect of withdrawing training: forgetting. In fact, in numerous studies, word learning performance declines when training is withdrawn (Storkel, Bontempo, Aschenbrenner, Maekawa, & Lee, 2013; Tamminen & Gaskell, 2013; Vlach & Sandhofer, 2012). Recently, Vlach and Sandhofer (2012) documented a curvilinear pattern of forgetting of a newly learned word for both adults and children. Specifically, for both groups, performance dropped rapidly between an immediate test and a 1-week delayed test, and then dropped less rapidly from the 1-week delayed test to a 1-month delayed test, as an asymptote was reached (i.e., performance at floor). While the evidence of memory consolidation as well as forgetting during gaps in word learning training may seem contradictory, emerging theories of memory suggest these counter processes may be critical in creating stable but flexible representations (Hardt, Nader, & Nadel, 2013; Stickgold & Walker, 2013). In particular, forgetting allows non-essential details of a word learning experience to be lost over time so that subsequent word learning experiences will reinforce the essential elements (Vlach & Sandhofer, 2012). Recent theories of memory have referred to this array of off-line memory changes as *memory evolution* to better reflect the idea that memories can be forgotten via decay or interference or retained via consolidation, integration, and generalization (Hardt et al., 2013; Stickgold & Walker, 2013). The purpose of the current study is to build on this framework that differentiates on-line learning during input from off-line memory evolution during no-input gaps by examining two variables that are known to influence adult word learning, specifically neighborhood density of the words to be learned and practice across different training sets, to determine how each variable influences each process.

Neighborhood Density

Neighborhood density is the number of words that are phonologically similar to a given word (Luce & Pisoni, 1998). During training, adults learn words with many neighbors (i.e., high density) more accurately than words with few neighbors (i.e., low density, Storkel, Armbruster, & Hogan, 2006). Thus, high density appears to facilitate learning from input, possibly due to the influence of density on working memory. That is, high density items are

held more accurately in working memory than low density items (Roodenrys & Hinton, 2002; Thomson, Richardson, & Goswami, 2005; Thorn & Frankish, 2005), potentially leading to the creation of a more accurate and/or detailed representation in long-term memory. Unfortunately, Storkel and colleagues (2006) did not examine retention across a no-training gap. Thus, it is unknown how neighborhood density influences memory evolution in word learning by adults (but see Storkel & Lee, 2011 for discussion of child word learning). Thus, one purpose of the current study was to examine the effect of neighborhood density on adult word learning during training and across no-training gaps to provide a clearer picture of how density influences on-line learning from input versus off-line memory evolution in adult word learning.

Practice across Training Sets

Gershkoff-Stowe and Han (2007, 2009) showed that learning of one set of words could facilitate learning of a second set of words for both children and adults. In Gershkoff-Stowe and Han (2009), adults received two training sessions for a 1st set of novel words, which were learned to a relatively high level of accuracy. Then, they were trained on a 2nd set of novel words in a third training session. When performance was compared at similar points in training (i.e., first training session for each set), adults were more accurate for the 2nd set of words than the 1st set of words. This improvement for the 2nd set of words is sometimes referred to as a *practice effect*. It was hypothesized that the 1st training set primed learning of the 2nd training set by activating semantic or conceptual networks or perhaps even the entire lexical network. Thus, it appears that learning from input can be enhanced by repeated experience learning novel words. However, the effect of this repeated practice on off-line memory evolution is not known because retention of novel words after a no-input gap was not compared for the 1st versus 2nd set. In addition, the 1st and 2nd sets were related in a general or contextual way only. That is, the sets were “matched approximately for object category” (p. 686, Gershkoff-Stowe & Hahn, 2007, presumably artifacts and natural kinds) and were trained in the same manner in the same environment (i.e., laboratory environment). It is unclear whether a more specific similarity relationship, namely phonological similarity, would alter the effect of practice. Based on the authors’ account, a specific similarity relationship should still result in a positive effect of practice on learning different sets of training words because practice on the 1st set of words would activate the existing network of phonologically similar words, namely the word’s neighborhood, potentially facilitating adding new words from the 2nd set to that same neighborhood. The current study tests this hypothesis by examining learning of a 2nd set of nonwords that are phonological neighbors of a 1st set of taught nonwords. In addition, performance is examined during training and following no-training gaps to elucidate the influence of practice on on-line learning from input versus off-line memory evolution in the absence of input.

Purpose

The purpose of this study was to determine how neighborhood density and practice across phonologically-related training sets influence on-line learning from input during training versus off-line memory evolution during no-training gaps. These two particular variables are of interest because they index different aspects of phonological similarity. In particular,

neighborhood density indexes the phonological similarity between a new word and already known words. Thus, this variable captures similarity between the new input and the existing contents of memory. In contrast, the relationship between the training sets indexes phonological similarity between new words, capturing similarity within the recent and current input. Examining how these two variables influence on-line learning from input during training versus off-line memory evolution during no-training gaps will shed light on whether the effect of phonological similarity is the same on these two processes regardless of the locus of the similarity: long-term memory in the case of density versus recent past input in the case of phonological practice. Based on studies of adults, it is predicted that high density nonwords will be learned better than low density nonwords during training and that 2nd set nonwords will be learned better than 1st set nonwords during training. Thus, both types of phonological similarity are expected to facilitate on-line learning from input during training. Prior studies of adults have not examined retention of nonwords varying by density or practice so it is unclear whether either or both types of similarity will enhance off-line memory evolution.

Method

Participants

Sixty-one adults between the ages of 18 – 31 years participated. The participants were recruited from the university community and received either payment or partial course credit. Based on self-report, all participants had normal hearing and were monolingual native English speakers. Likewise, none of the participants reported a history of speech, language, or cognitive delay. The participants scored within normal limits (i.e., received a standard score of 85 or higher) on a receptive vocabulary test (i.e., Peabody Picture Vocabulary Test - 4th edition, Dunn & Dunn, 2007). Participants were randomized to one of two stimulus conditions: (1) low density; (2) high density. The participant characteristics for the low versus high density groups are shown in Table 1. As shown in Table 1, the groups were comparable on all areas measured.

Stimuli

The stimuli used for this study were 48 nonwords, consisting of consonant-vowel-consonant (CVC) sequences that were not found in an online corpus of real English words (Storkel, 2013; Storkel & Hoover, 2010). Therefore, these nonwords should not be part of a native English speaker's lexicon. The 48 nonwords are shown in Table S1 of the online supplemental materials. Nonwords were selected from the corpus to manipulate neighborhood density, one of the independent variables. Neighborhood density refers to the total number of words in a corpus that are phonologically similar to a given word (Luce & Pisoni, 1998). Phonological similarity is defined as all the words that differ from the target word by exactly one sound segment (i.e., a substitution, addition, or deletion in any word position). Low density neighborhoods have few phonologically similar neighbors, whereas high density neighborhoods have many phonologically similar neighbors. One set of 24 nonwords from low density neighborhoods and one set of 24 nonwords from high density neighborhoods were selected. Low density was defined as a density 25th percentile of the CVC pool (cf., Storkel & Lee, 2011). High density was defined as a density 75th percentile

of the CVC pool (cf., Storkel & Lee, 2011). The low density nonwords had 2–7 real-word neighbors ($M = 5.42$, $SD = 1.35$), whereas high density nonwords had 15–22 neighbors ($M = 17.04$, $SD = 2.12$).

Each set of 24 nonwords (i.e., low vs. high density) consisted of 12 pairs of phonological neighbors, as shown in Table S1 of the supplement. This pairing was needed to examine the effect of phonologically related training sets. That is, one member of the pair was trained first (e.g., Group A /ver/) during initial training sessions, followed by training on the second member of the pair (e.g., Group B /ner/) during later training sessions. This creates the right conditions for testing the effect of practice learning phonologically related nonwords (i.e., does learning /ver/ subsequently influence learning of /ner/). The set of items that were trained first are referred to as the 1st set. In complement, the set of items that were trained second are referred to as the 2nd set. Note that half of the participants in each density condition were trained in the Group A then Group B order, whereas the other half of the participants in each condition were trained in the reverse order: Group B then Group A. These groupings are shown in Table S1 of the supplement, and the neighborhood density associated with each group is shown in Table 2. As shown in Table 2, Group A was similar in density to Group B within a given density condition (low or high). Thus, density is matched across the 1st and 2nd set. Of the 12 pairs of phonological neighbors in each density condition, four pairs were taken from each of three types of neighbors: (1) neighbors that shared the initial consonant and vowel (i.e., CV_ neighbors), such as /fub/-/fuk/ in the low density condition; (2) neighbors that shared the vowel and final consonant (i.e., _VC neighbors), such as /ver/ - /ner/ in the low density condition; (3) neighbors that shared the consonants (i.e., C_C neighbors), such as /zis/ - /zos/ in the low density condition. Each neighbor pair came from a different phonological neighborhood. That is, aside from these neighbor pairs, the nonwords in a condition were not neighbors of one another.

Stimuli also were selected to control phonotactic probability, namely the likelihood of occurrence of a sound or sound sequence in words of the language. This is based on two measures: positional segment sum and biphone sum (Storkel, 2004). The *positional segment sum* is the sum of the positional segment frequency of each sound in the word. *Positional segment frequency*, in turn, is the sum of the log frequency of every word in the corpus with the specific sound in the target word position, divided by the sum of the log frequency of every word in the corpus with any sound in the target position. *Biphone sum* is the sum of the biphone frequency of each adjacent pair of sounds in the word. Similar to positional segment frequency, *biphone frequency* is the sum of the log frequency of every word in the corpus with the target pair of sounds in the target word position, divided by the sum of the log frequency of every word in the corpus containing any sound in the target word position.

The nonwords used as stimuli in this study had “mid” phonotactic probability, defined as adult segment sum and biphone sums between the 25th and 75th percentiles. Table 2 provides the phonotactic probability values for each nonword group for low and high density conditions. The phonotactic probability for low density and high density conditions were compared to ensure minimal variability. There was no significant difference for the segment sum, $F(1, 44) = 0.33$, $p > 0.50$, $\eta^2 < .01$. Segment sum for low density nonwords ($M = .10$, $SD = .02$, range .08–.13) was similar to that of high density nonwords ($M = .11$, $SD = .02$,

range .08 –.13). In contrast, there was a significant effect for biphone sum, $F(1, 44) = 11.57$, $p = 0.001$, $\eta p^2 = .21$. Here, the low density nonwords ($M = .0020$, $SD = .0008$, range .0012 – .0035) tended to be lower in biphone sum than the high density nonwords ($M = .0027$, $SD = .0007$, range .0013 – .0036). However, the effect size for biphone sum was relatively small, especially when compared to the effect size for density, $F(1, 44) = 503.38$, $p < 0.001$, $\eta p^2 = .92$. Although biphone sum is not perfectly matched across the low and high density conditions, the two conditions vary more in density than biphone sum and are well matched on segment sum.

Each nonword was paired with a nonobject. The same group of 24 nonobjects was used for the low and high density conditions. Thus, there is no difference in nonobject characteristics across the two density conditions. The nonobjects were selected from those published by Kroll & Potter (1984). The reference number of the selected nonobjects are shown in Table S1 of the online supplement. Within each density condition, one group of 12 nonwords (e.g., Group A) was paired with a group of 12 nonobjects (e.g., Group 1), and the second group of 12 nonwords (e.g., Group B) was paired with the other 12 nonobjects (e.g., Group 2). Recall that these groupings correspond to the order of training, namely 1st set versus 2nd set. Consequently, two nonobject characteristics were matched across groups/sets. Adult semantic set size for each nonobject refers to the number of different semantic neighbors reported by two or more adult participants (Storkel & Adlof, 2009a). Group 1 nonobjects ($M = 10.5$, $SD = 0.5$, range 10–11) had similar set sizes to Group 2 nonobjects ($M = 10.5$, $SD = 0.5$, range 10–11). Objectlikeness is the degree to which the nonobject resembles a real object on a 7-point scale, according to ratings published by Kroll & Potter (1984). Group 1 nonobjects ($M = 4.4$, $SD = 0.9$, range 3.2–5.6) were similar to Group 2 nonobjects on objectlikeness ratings ($M = 4.5$, $SD = 0.9$, range 3.2–5.9). Note that half of the participants received Group A nonwords paired with Group 1 nonobjects and Group B nonwords paired with Group 2 nonobjects. The other half of the participants received the reverse pairing (i.e., Group A with Group 2; Group B with Group 1).

Procedures

Each participant attended four sessions, at approximately seven-day intervals ($M = 7.6$, $SD = 2.8$, range 1–26). Recall that participants were randomized to receive either low density or high density stimuli. Each session consisted of multiple training-testing cycles administered via a computer with DirectRT software (Jarvis, 2002). During the *training phase*, participants passively listened to recordings of the nonword stimuli and viewed the nonobject on the computer. The nonword stimuli were presented in sentences in the word-final position. For each nonword in the set, the participant heard “This is a _____. Listen closely, it’s called a _____. Remember, it’s a _____. Look at the _____. Don’t forget the _____.” Thus, each participant received five exposures to each nonword per training cycle. The *testing phase* consisted of a production task, in which the participant was presented with a picture of each nonobject and asked to name the corresponding nonword. The accuracy of the participants’ responses on the production task served as the primary outcome measure of the study.

Session 1 focused exclusively on training and testing of the 1st set of stimuli (i.e., Group A or B nonwords paired with Group 1 or 2 nonobjects) for the participant's assigned density condition (low vs. high density). Session 1 began with a baseline naming test. In this test, the nonobject pictures were presented and participants were asked to guess the name. As expected, no participant produced the correct nonword for a nonobject at baseline. Instead, participants tended to use real words to describe some aspect of the nonobject (e.g., "wheels" for a nonobject that appeared to have wheels) or to name a real object that they perceived as similar to the nonobject (e.g., "pretzel" for a nonobject with a twisty shape like a pretzel). Baseline testing was followed by three cycles, each including one training phase and one testing phase. Tests during this session measure learning from input. Session 2 began with a test for the 1st set stimuli, which measured the participant's retention of 1st set stimuli after the one week break. This test taps memory evolution over the gap in training. This post-test was followed by an additional three cycles of training and testing for 1st set stimuli. Again, these tests during training measure learning from input. Session 3 also began with a post-test to measure retention of 1st set stimuli following the one week break. As with the first retention test, this second retention test taps memory evolution in the absence of input. 2nd set stimuli were then trained following the same procedures (i.e., baseline testing followed by three cycles of training and testing). These tests tap learning from input for the 2nd set stimuli. In Session 4, participants performed a post-test for the 2nd set nonwords. This retention test taps memory evolution for the 2nd set stimuli. The receptive vocabulary test (Dunn & Dunn, 2007) also was administered during Session 4.

In total, each participant received 30 exposures to each nonword in the 1st set and 15 exposures to each nonword in 2nd set during training. The 1st set received greater training than the 2nd set to ensure that the 1st set would be learned to a high level, establishing the necessary conditions for the practice effect. Excluding baseline testing, each participant performed eight production tests for 1st set stimuli (i.e., 6 tests during input; 2 tests of memory evolution in the absence of input) and four production tests for 2nd set stimuli (i.e., 3 tests during input; 1 test of memory evolution in the absence of input).

Procedural reliability was calculated for 13% of the sample using a checklist of items related to correct examiner behavior (e.g., appropriate instructions provided, appropriate nonword-nonobject set administered) and correct equipment function (e.g., all items presented appropriate number of times). Procedural reliability was computed as the percent of items on the checklist that were scored as appropriately administered divided by the total number of items on the checklist. Procedural reliability was high ($M = 96\%$, $SD = 6\%$, range 82% – 100%). The one instance of low reliability (i.e., 82%) was due to equipment issues (i.e., issues with the quality of the recording for the production test).

Scoring

All sessions were audio- and video-recorded. Participant responses on the production task were phonetically transcribed and scored. Interjudge transcription reliability was computed for 13% of the sample. Mean percent agreement was 96% ($SD = 2\%$, range 93% – 98%). Based on these transcriptions, responses were scored as correct or incorrect, and this served as the dependent variable for the primary analysis. Correct responses were those with all

three phonemes produced in the correct positions. Responses were scored incorrect if they contained substitutions or deletions of any of the target phonemes. Interjudge scoring reliability was computed for 13% of the sample, with mean agreement of 100% ($SD = 1\%$, range 98% – 100%).

Incorrect responses were further scored for a secondary descriptive analysis. First, incorrect responses were coded for the number of phonemes correct (i.e., 0, 1, or 2) to quantify phonological similarity between the target and the incorrect response. Second, incorrect responses that shared 2 of 3 phonemes with the target (e.g., response /ves/ for target /ver/) were compared to the list of real word neighbors of the target. Incorrect responses were scored as a real word neighbor if they shared 3 of 3 phonemes with one of the items on the real word neighbor list (e.g., response /ves/ for target /ver/ scored as real word neighbor “vase”). Likewise, incorrect responses in the 2nd set that shared 2 of 3 phonemes with the target (e.g., response /ner/ for target /ver/) were compared to the 1st set neighbor and scored as a 1st set neighbor if 3 of 3 phonemes were shared (e.g., /ner/ is a set neighbor for /ver/). These two additional measures for phonologically similar errors quantify the influence of real word neighbors and 1st set trained items. Third, all incorrect responses were compared to the list of trained nonwords and scored as a trained nonword error if 3 of 3 phonemes were shared. This measure quantifies the extent to which participants are confusing nonwords that are being trained together. Lastly, for incorrect responses that shared 0 of 3 phonemes, it was noted whether the participant gave no response (e.g., said nothing, “I can’t remember”) to quantify the extent to which participants had not formed any representation or had experienced a retrieval failure. Error coding was completed by a first judge and then by a second judge with discrepancies resolved through discussion.

Results

Primary Analysis Approach

The accuracy data were analyzed using multilevel modeling. Multilevel modeling (MLM), also called mixed effects modeling, hierarchical linear modeling, or random coefficient modeling, is becoming a preferred method for analyzing repeated measures data because it allows for a variety of variance/covariance structures, thus being more flexible regarding dependencies arising from repeated measures as well as accommodating missing and/or unbalanced data (Cnaan, Laird, & Slasor, 1997; Gueorguieva & Krystal, 2004; Hoffman & Rovine, 2007; Misangyi, LePine, Algina, & Goeddeke, 2006; Nezlek, Schroder-Abe, & Schutz, 2006; Quene & van den Bergh, 2004). Moreover, random effects of participants and items can be accommodated in the same analysis by incorporating crossed random intercepts, and this is becoming the favored analysis approach for psycholinguistic data (cf., Baayen, Davidson, & Bates, 2008; Locker, Hoffman, & Bovaird, 2007; Quene & van den Bergh, 2008). Note that the dependent variable for this study was accuracy (i.e., correct or incorrect), which is a binary variable. Thus, a logistic MLM was used. The analysis proceeded in several model building steps; however, only the final model is reported here because certain variables are less interpretable in preliminary models that exclude other variables (e.g., main effects are less interpretable/meaningful when a significant interaction is present). As shown in Table 3 (i.e., rows 16 and 17), McKelvey and Zavoina’s r -squared

(Hox, 2010; Windmeijer, 1995) and Somers' D (Group, 2014; Newson, 2002) were used as indicators of model fit. The reported final model in Table 3 was among the models with the highest r-squared and Somers' D, with very minimal change in these values when other non-significant effects were added to the model.

Participant and Item Effects

Table 3 shows the final model. First, the crossed random effects of participants and items (i.e., nonwords) were examined to better understand variability. To facilitate insight into the magnitude of individual differences, participant and item level variance was expressed as a median odds ratio (MOR, Merlo et al., 2006). Conceptually, the MOR conveys the median increase in the odds of a correct response between a pair of participants or items that are alike on all other covariates. The MOR has the further advantage of being on the same scale as the odds ratio (OR), which was used as the effect size for the fixed effects (e.g., neighborhood density). In terms of interpreting the magnitude of an effect stated in an odds ratio metric, there are not straightforward guidelines because the interpretation is somewhat dependent on the rate of the event of interest (Chen, Cohen, & Chen, 2010). However, generally values near 1.5 are interpreted as small effects, those near 2–3 are interpreted as moderate effects, and those near 5–6 and higher are considered large effects (Chen et al., 2010). As shown in Table 3 (i.e., rows 14 and 15), the MOR for participants was 1.92 (95% CI = 1.70 – 2.22) and the MOR for items was 3.00 (95% CI = 2.44 – 3.88). Thus, the variability between participants is associated with a median difference of 1.92 in the odds of a correct response when comparing two randomly drawn participants who are alike on other covariates. This is a moderate effect size. Likewise, the variability between nonwords is associated with a median difference of 3.00 in the odds of a correct response between two randomly drawn nonwords that are alike on other covariates. Again, this is a moderate effect size. Taken together, random effects of participants and items were moderate in size with the variance of items being slightly larger than that of participants.

Significant predictors of participant and item variability also were included in the model, as shown in Table 3 (rows 1–3). In terms of participant characteristics, the odds of a correct response were 1.60 (95% CI = 1.00 – 2.56) times higher for males relative to females. In terms of mean accuracy, males had a mean accuracy of 66% (95% CI = 60–73%) compared to 58% (95% CI = 54–62%) for females. In the final model reported in Table 3, the effect of gender was no longer significant. However, it was significant in earlier models and just missed traditional criteria for significance in the final model. For this reason, gender was retained in the model as a reference point for future research. Turning to vocabulary scores, the odds of a correct response were 1.02 (95% CI = 1.01 – 1.04) times higher for each one point increase in standard score on the *Peabody Picture Vocabulary Test*. Splitting the sample at the median standard score of 113 showed that mean accuracy of participants with scores at or above the median was 64% (95% CI = 59–69%), whereas mean accuracy of participants with scores below the median was 55% (95% CI = 51–59%). In terms of item characteristics, the odds of a correct response were 1.22 (95% CI = 1.06 – 1.39) times higher for a one unit decrease in semantic set size, replicating Storkel and Adlof (2009b). However, mean accuracy values were quite similar. Specifically, mean accuracy was 59% (95% CI = 53–65%) for higher set size items compared to 60% (95% CI = 54–67%) for lower set size

items. This small mean difference suggests that this term might be tapping other characteristics of the nonobjects that were not entered in the model (e.g., strength of the first semantic associate, Storkel & Adlof, 2009a). This effect should be interpreted with caution. Note that the magnitude of these participant and item level predictors was relatively small.

Input versus Memory Evolution Tests

Turning to variables that address the research questions, Figure 1 shows the data for test (each individual panel), neighborhood density as a continuous variable (x-axis), and set (top versus bottom row). Starting with the effect of test, recall that the absolute timing of different tests varied. That is, the first three tests (excluding baseline) occurred during input (i.e., first training phase). Thus, these tests are separated by minutes. In contrast, the fourth test occurred after a gap without any training. Thus, test 3 and test 4 are separated by approximately 1 week. The next set of tests (i.e., tests 5, 6, and 7) again occurred during input (i.e., second training phase), returning to a separation in tests on the order of minutes. Finally, the last test (i.e., test 8) again occurred after a no training gap, creating a separation of approximately 1 week from the previous test. To address this variability in absolute timing of tests and in the mental activities that occurred between tests (i.e., on-line learning from input during training vs. off-line memory evolution during no-training gaps), spline regression and multiple intercepts were used to model the effect of test. With linear splines, the effect of an explanatory variable (i.e. test) is assumed to be piecewise linear on a specified number of segments separated by knots where the slope changes (Gould, 1993; Panis, 1994). Coding can be for the slope or the change in slope (cf., Storkel et al., 2013). To capture abrupt changes in level (i.e., after one week of memory evolution), an additional intercept parameter can be paired with each knot. Coding for this model is shown in Table S2 of the online supplemental materials.

The fixed effects of test are shown in rows 4–7 in Table 3. The first coefficient (row 4 of Table 3) captures the slope across all tests that were administered, excluding baseline. This term generally captures learning from input because additional coefficients capture memory evolution, as described later. This all-test slope was significant, indicating that the odds of a correct response were 3.75 (95% CI = 3.36 – 4.18) times higher for each subsequent test relative to the immediately preceding test. Note that this odds ratio indicates a moderate effect of training. This also can be seen in Figure 1, where accuracy generally increases across each panel. The second coefficient (row 5 of Table 3) captures the abrupt change across the first 1-week gap in training. This is the first retention test, capturing off-line memory evolution. As shown in Table 3, this effect was significant with a large effect size. Specifically, the odds of a correct response were 25.00 (95% CI = 18.18 – 34.48) times lower for the test point following the no training gap relative to the test point preceding the no training gap. Thus, participants forgot the newly learned words when training was withdrawn for 1-week. This is apparent in Figure 1 by comparing the Test 3 and Test 4 panels, which clearly show a precipitous drop in performance from Test 3 to Test 4. Note that performance at Test 4 looks quite similar to performance at Test 1, indicating that the forgetting that occurred during the 1-week gap negated much of the learning that occurred during input. The third coefficient (row 6 of Table 3) was intended to capture any change in slope that might occur in during later training after this first gap. As shown in Table 3, there

was a non-significant change in slope in these later tests, with the trend being for a slightly shallower slope across this training-testing period. As shown in the Test 4–8 panels of Figure 1, accuracy began to approach ceiling in later tests, accounting for this non-significantly shallower slope during later training and testing. Although this effect was non-significant, it was retained in the model because the full set of predictors is needed to describe the complexity of time in this design. Note that models dropping these non-significant effects did not lead to appreciable changes in odds ratios or p-values for the other effects in the model. The fourth coefficient (row 7 of Table 3) captured abrupt change across the second 1-week gap in training. This is the second retention test, providing an additional window into off-line memory evolution. As shown in Table 3, this effect was significant with a large effect size and again indicated forgetting when training was withdrawn. Specifically, the odds of a correct response were 29.41 (95% CI = 18.52 – 45.45) times lower for the test point following the no training gap relative to the test point preceding the no training gap. Comparison of the Test 7 and Test 8 panels in Figure 1 also shows a clear drop in accuracy consistent with forgetting. Once again the forgetting that occurred during the 1-week gap in training negated much of the learning that occurred during training, as can be seen by comparing Test 8 to Tests 4 and 5. Note that the odds ratio for the two retention coefficients (i.e., 25.00 versus 29.41) is quite similar, indicating similarly large effects of forgetting across each 1-week gap.

Neighborhood Density

Turning to neighborhood density, the effect of density was modeled using both main effects and interactions (see rows 8–11 of Table 3). For all of these terms, density was modeled as a continuous predictor (i.e., raw density), rather than as dichotomously coded variable (i.e., low vs. high). As shown in row 8 of Table 3, the main effect of density was not significant. The direction of the main effect is for high density nonwords to be responded to more accurately than low density nonwords. In particular, the odds of a correct response were 1.03 (95% CI = 0.97 – 1.10) times higher for each one neighbor increase in density. This non-significant main effect of density was retained in the model because interaction terms involving density were significant. First, the interaction between density and tests (row 9 of Table 3) was significant with a small effect size, indicating that the difference between low and high density words widens as training progresses. The fit line in each panel of Figure 1 depicts the effect of density. Across panels, which correspond to test, the steepness of this line generally increases, indicating a greater advantage for high over low density words as training progresses. Second, the interaction between density and the first retention test (row 10 of Table 3) is significant with a small effect size. Here, there is a greater drop in word learning performance across the no-training gap for higher density nonwords than for lower density nonwords. As shown in Figure 1, the fit line for density has a shallower slope at Test 4 than at Test 3, indicating a more precipitous drop in accuracy for high density nonwords across the no-training gap, yielding accuracy that is more similar across low and high density nonwords after memory evolution has taken place. Third, the interaction between density and the second retention test (row 11 of Table 3) is significant with a small effect size. The interpretation is similar to the first retention test. As shown in Figure 1, the fit line for density has a shallower slope at Test 8 than at Test 7, indicating greater forgetting for high density nonwords across the second no-training gap. Lastly, an interaction between

density and later tests (i.e., Tests 4–8 that occurred after the first gap in training) was tested but found to be non-significant (OR = 0.99, 95% CI = 0.97 – 1.01). Thus, this interaction was deleted from the final model and is not shown in Table 3. This non-significant interaction suggests that the relationship between density and test is consistent across all tests, without any significant change occurring in the later tests as ceiling is approached. To summarize, high density nonwords were learned better than low density nonwords during training, but high density nonwords experienced greater forgetting during the no-training gaps than low density nonwords.

Turning to the error data, Table 4 shows the error pattern for input versus retention tests for low and high density items. As shown in Table 4, errors during input testing (i.e., Tests 1, 2, 3, 5, 6, and 7) were primarily close phonological approximations of the target, sharing 2 of 3 phonemes with the target. Notably, the percentage of close phonological approximations was relatively similar across low and high density items. However, these phonological approximations were frequently real word neighbors for high density items but rarely real word neighbors for low density items. The percentages of other error types during input testing were similar across low and high density items. Turning to retention testing (i.e., Tests 4, 8), errors tended to be dissimilar to the target with the predominant error pattern being 0 of 3 phonemes shared. This fits well with the accuracy analysis by further showing the impact of forgetting. Specifically, participants have now either completely lost even partial representations for an item or have lost access to partial (or full) representations of an item. Like the accuracy analysis, there is evidence in the error analysis that the impact of forgetting is greater for high density than for low density. During training, close phonological approximations (i.e., 2 of 3 phonemes shared) were relatively similar across low and high density items. However, at retention, fewer close phonological approximations are observed for high than low density. Likewise, the no response rate during retention testing is higher for high density items than for low density items, indicating that forgetting has a larger impact on high than low density items. Overall, phonological approximation errors were observed for both low and high density items during training, and these were most often real words for high density items. In contrast, after memory evolution, errors were dissimilar from the targets for both low and high density items, with fewer phonological approximations and more no response errors for high than low density items.

Practice across Phonologically-Related Training Sets

Turning to the influence of practice on learning phonologically similar nonwords, the effect of Set was modeled using both main effects and interactions (see rows 12–13 of Table 3). The main effect of Set was significant with a small effect size (see row 12 of Table 3). Specifically, the odds of a correct response were 1.23 (95% CI = 1.06 – 1.43) times higher for the 2nd set relative to the 1st set. Thus, learning the 1st set appeared to facilitate learning of the 2nd set. Although this can be seen by comparing the upper (Set 1) and lower (Set 2) rows of plots in Figure 1, it is a bit difficult. For clearer illustration with mean values, at Test 1, mean accuracy is 36% (95% CI = 30–42%) for 2nd set nonwords compared to 31% (95% CI = 25–36%) for 1st set nonwords. Likewise, at Test 2, mean accuracy is 63% (95% CI = 55–70%) for 2nd set nonwords compared to 60% (95% CI = 53–67%) for 1st set nonwords. At Test 3, mean accuracy is 73% (95% CI = 65–81%) for 2nd set nonwords compared to

71% (95% CI = 64–79%) for 1st set nonwords. Although it appears that the effect of set may be diminishing across tests, this interaction was not significant (OR = 0.93, 95% CI = 0.77 – 1.12) and is not shown in the Table 3 final model. However, there was a significant interaction of set with the first retention test (see row 13 of Table 3), and this effect had a moderate effect size. Here, the effect of set reverses, with the 1st set showing a smaller drop in accuracy compared to the 2nd set across the no-training gap. This again may be difficult to ascertain from Figure 1. In terms of mean values for clearer comparison, mean accuracy for 1st set at Test 3 is 71% (95% CI = 64–79%) versus 39% (95% CI = 33–44%) at Test 4. In comparison, mean accuracy for the 2nd set at Test 3 is 73% (95% CI = 65–81%), but mean accuracy drops more dramatically to 29% (95% CI = 24–34%) at Test 4. Overall, the 2nd set appears to be learned more accurately than the 1st set during training but the 1st set appears to be retained more accurately than the 2nd set across the no training gap.

Turning to the error data, Table 5 shows the error pattern for input versus retention tests for 1st and 2nd set items. As shown in Table 5, errors during input testing (i.e., Tests 1, 2, 3, 5, 6, and 7) were primarily close phonological approximations of the target, sharing 2 of 3 phonemes. In general, the error types were similar across 1st and 2nd set items during training. Turning to retention testing (i.e., Tests 4, 8), the predominant error is 0 of 3 phonemes shared. As previously discussed for density, this reflects forgetting across the gap in training. More important for understanding the effect of set is that differences between 1st set and 2nd set errors begin to emerge at retention. Specifically, 0 of 3 phoneme errors are more prevalent for the 2nd set than the 1st set. In complement, 2 of 3 phoneme errors are more prevalent for the 1st set than the 2nd set. These data fit well with the hypothesis that forgetting had a greater impact on the 2nd set than the 1st set. Here, partial representations (or access to partial representation) appears to be retained better for the 1st set than the 2nd set. In addition, the number of 1st set neighbors that are produced as errors to 2nd set targets doubled from the input tests to the retention tests. This suggests that 1st set neighbors may be interfering with retention of 2nd set targets. A related observation is that the 2nd set shows more other trained nonword errors than the 1st set at retention. This could indicate confusion between 1st set and 2nd set items that are not neighbors of one another and/or confusion among 2nd set items. Either scenario fits the hypothesis that forgetting has a greater impact on 2nd set than 1st set items, potentially due to confusion among trained items during the gap.

Additional Models

An interaction between density and set was explored but this term was not significant (OR = 1.01, 95% CI = 0.99 – 1.03). Thus, this interaction was not retained in the final model. Likewise, an interaction between density and vocabulary scores was explored to address the possibility of individual differences in the effect of density. However, this interaction term was not significant (OR = 0.999, 95% CI = 0.997 – 1.000) and, therefore, was not retained. In addition, the influence of the type of neighbor on the effect of set was explored. These data are shown in Table S3 of the on-line supplement. Recall that there were three ways that 1st set and 2nd set nonwords could be related. Specifically, they could share the initial consonant and vowel (i.e., CV_ neighbors), the vowel and final consonant (i.e., _VC neighbors), or the consonants (i.e., C_C neighbors). It is possible that the effect of set could

be stronger or weaker across these different types of relationships so models incorporating interactions between set and type of neighbor were explored (see Table S3 of the supplement). These models showed that the effect of set was consistent across these three types of neighbors.

Discussion

The purpose of this study was to determine how neighborhood density and practice across phonologically-related training sets influence on-line learning from input during training versus off-line memory evolution during no-training gaps. In terms of learning from input, steep learning during training was observed with improvements diminishing slightly (albeit non-significantly) at later test points as ceiling performance was approached. During learning from input, new words from dense neighborhoods were responded to more accurately than new words from sparse neighborhoods, replicating prior work (Storkel et al., 2006). In addition, words that were neighbors of previously learned words (i.e., 2nd set words) were learned better than words that were not neighbors of previously learned words (i.e., 1st set words), replicating prior work (Gershkoff-Stowe & Hahn, 2007, 2009). In terms of the overall pattern of memory evolution, large and significant forgetting was observed across both 1-week gaps in training. Effects of density and practice during memory evolution were opposite of those observed during input. Specifically, high density words experienced greater forgetting than low density words, and words that were neighbors of previously learned words (i.e., 2nd set words) experienced greater forgetting than words that were not neighbors of previously learned words (i.e., 1st set words). This asymmetry in the effect of the variables across on-line learning and off-line memory evolution supports the need to differentiate these two mechanisms to understand word learning.

On-line learning during training versus off-line memory evolution

As in most prior studies, adult participants demonstrated strong and robust on-line learning during training. In contrast, off-line memory evolution appears to vary across studies with some studies showing improvements in performance, namely memory consolidation, across gaps in training (Gaskell & Dumay, 2003; Leach & Samuel, 2007; Rice et al., 1994; Storkel, 2001, 2003; Storkel & Lee, 2011), and others showing declines in performance, namely forgetting, across gaps in training (Storkel et al., 2013; Tamminen & Gaskell, 2013; Vlach & Sandhofer, 2012). There are numerous differences in methodology that could account for why some studies show evidence of memory consolidation while others show evidence of forgetting. For the current study, which showed clear evidence of forgetting during gaps in training, one contributing factor likely was the size of the gap in training, which was 1-week. As shown by Vlach and Sandhofer (2012) for word learning, forgetting is a curvilinear function such that the rate of forgetting is initially rapid but then slows over time. Vlach and Sandhofer tested word learning immediately following training, 1-week after training, and 1-month after training. Visual inspection of their figure suggests that 1-week after training approximates the turn in the curvilinear function that marks the end of the rapid phase of forgetting. Thus, forgetting may have been minimized if the gap in training had been smaller. An additional factor that may have contributed to forgetting in the current study is that the training exposure was not particularly rich. That is, the training solely relied

on passive auditory exposure, rather than more active participation (e.g., repeating the word) or more detailed description of the nonobjects (e.g., highlighting perceptual or semantic characteristics) or nonwords (e.g., highlighting phonological characteristics). In addition, although participants in this study had an opportunity to generate the words during repeated naming testing, they did not receive any feedback on the accuracy of their responses. Thus, a richer and more interactive training procedure may have minimized forgetting or may have even resulted in memory consolidation (Vlach & Sandhofer, 2012).

Regardless of the reason that forgetting was observed in this study rather than memory consolidation, the impact of forgetting on performance was striking. That is, visual inspection of Figure 1 showed that forgetting after a gap in training lead to performance that was similar to performance observed at the start of training. Moreover, comparison of odds ratios further shows that the impact of forgetting was 6 to 7 times larger than the impact of one additional training (i.e., the inverse-retention odds ratios of 25–29 are approximately 6–7 times larger than the testing odds ratio of 3.75). In essence, forgetting negated the effects of several training cycles. This underscores the importance of understanding what can be done during training to either minimize forgetting or enhance consolidation during gaps in training, especially in educational and clinical contexts. In these types of naturalistic contexts, the magnitude of forgetting observed in the current study would severely limit overall gains.

Neighborhood density

As in prior studies, words with many neighbors (i.e., high density) were learned more accurately than words with few neighbors (i.e., low density, Storkel et al., 2006). The prior account of this effect is that high density sound sequences are held more accurately in working memory than low density sound sequences (Roodenrys & Hinton, 2002; Thomson et al., 2005; Thorn & Frankish, 2005). The more accurate representation in working memory may then support the creation of a more accurate and/or detailed representation in long-term memory, leading to more accurate responding for high density words during training than for low density words. One might be inclined to find support for this hypothesis in the error analysis, where more real words were reported for high density than for low density words, which could suggest that the real words are activated and providing support for learning. However, it is unclear how to interpret these real word errors. By definition, high density words have more real word neighbors than low density words. Thus, a one phoneme change to a high density word has a greater probability of yielding a real word neighbor than that same change to a low density word. For example, the substitution of /b/ for /z/ results in a real word for the high density nonword /zin/ (i.e., “bean”) but not for the low density nonword /zis/ (i.e., “beas” is not a real word). For this reason, it is unclear whether real word errors reflect an interaction between the target and the neighbors or are actually phonological substitutions that result in real words by chance.

Turning to memory evolution, this facilitation of learning high density words during training is then countered by greater forgetting of high density words during the 1-week gap in training. This greater forgetting of high density words than low density words suggests that interference may occur between newly learned words and existing known words in long-

term memory (see Hardt et al., 2013 for a discussion of interference in neocortical memory systems). Because a high density word is similar to many more existing known words than a low density word is, interference would be greater for high density than low density words, accounting for the greater forgetting of high than low density words. This hypothesis is supported by the error analysis which showed fewer phonological approximations and more no responses for high density than for low density items, indicating that partial representations or access to partial representations had been lost for high density items after memory evolution. This pattern indicates greater interference for high density than for low density items, although it fails to reveal the source of that interference. Thus, whether real word neighbors are actually the cause of greater interference for high density words requires additional verification.

Practice across Training Sets

This study extends prior work on learning similar sets of words (Gershkoff-Stowe & Hahn, 2007, 2009) to the learning of phonologically similar sets of words. As with conceptually similar set of words, learning a second set of phonologically similar words facilitated word learning relative to learning a first set of words. In keeping with the hypothesized mechanism for conceptually similar word sets, learning the first set of words presumably primes the existing phonological neighborhoods of these words. Thus, when a second word from the same phonological neighborhood is taught (i.e., 2nd set), it is easier to learn the second word than when the neighborhood has not been previously primed (i.e., 1st set). A more novel observation from the current study is that this facilitative effect of practice is reversed when retention is examined. Here, 2nd set words experienced greater forgetting than 1st set words across a 1-week gap in training. This finding suggests that retention of the words learned in the 1st set interfered with retention of the words learned in the 2nd set. That is, the phonological overlap between two recently learned words leads to confusion between the two words, negatively impacting the later learned (i.e., 2nd set) word. This hypothesis is supported by the finding that 1st set nonwords were reported in place of 2nd set nonwords more frequently after the gap.

It is important to note that it cannot be unequivocally determined whether the phonological overlap between the sets was crucial in obtaining the effects observed. A control condition where phonologically unrelated words were taught in each set was not included. Thus, it is possible that the effect might be attributable to a broader mechanism. That is, activating the lexicon to learn any kind of word may facilitate learning any type of subsequent word, regardless of whether these words are related in some way or not. Likewise, the 1st set of words might induce greater forgetting in the 2nd set of words, even if the words are unrelated. There is some suggestion of this in the error analysis where other trained nonwords were reported in place of 2nd set nonwords more frequently after the gap. The other trained nonwords are unrelated to the target (due to the way the stimuli were selected). Taken together, there is evidence of interference from 1st set related nonwords as well as unrelated nonwords on the 2nd set nonwords after memory evolution. This suggests the need to further tease apart the role of similarity among trained items on learning from input and memory evolution.

Conclusion

Two types of phonological similarity were examined: (1) similarity between a new word and existing words in long-term memory (i.e., neighborhood density); (2) similarity between recent and current input (i.e., 1st set vs. 2nd set neighbors). The results across these two types of phonological similarity were the same. Specifically, phonological similarity, regardless of source, appears to facilitate on-line learning from input, replicating prior studies. In contrast, phonological similarity, regardless of source, appears to impede off-line memory evolution. That is, greater forgetting across a gap in training was observed when similarity was high than when it was low. This suggests that high similarity may increase interference compared to low similarity. Overall, the degree of forgetting was striking. When combined with the other findings, there is a clear need to explore memory evolution in word learning to determine how to minimize forgetting and enhance memory consolidation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The project described was supported by Grants DC 08095, DC 05803, and HD02528 from NIH. The contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. The second author was supported by the Analytic Techniques and Technology (ATT) Core of the Center for Biobehavioral Neurosciences of Communication Disorders (BNCD, DC05803). We would like to thank staff of the Participant Recruitment and Management Core (PARC) of the Center for Biobehavioral Neurosciences of Communication Disorders (BNCD, DC05803) for assistance with recruitment of preschools and children; staff of the Word and Sound Learning Lab (supported by DC 08095) for their contributions to stimulus creation, data collection, data processing, and reliability calculations.

References

- Baayen RH, Davidson DJ, Bates DM. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*. 2008; 59(4):390–412.
- Chen H, Cohen P, Chen S. How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communication in Statistics - Simulation and Computation*. 2010; 39:860–864.
- Cnaan A, Laird NM, Slasor P. Using the general linear mixed model to analyze unbalanced repeated measures and longitudinal data. *Statistics in Medicine*. 1997; 16:2349–2380. [PubMed: 9351170]
- Davis MH, Betta AMD, Macdonald MJE, Gaskell MG. Learning and consolidation of novel spoken words. *Journal of Cognitive Neuroscience*. 2008; 21:803–820. [PubMed: 18578598]
- Davis MH, Gaskell MG. A complementary systems account of word learning: neural and behavioural evidence. *Philosophical Transactions of the Royal Society of London, Series B, Biological Science*. 2009; 364(1536):3773–3800. doi:364/1536/3773 [pii] 10.1098/rstb.2009.0111 [doi].
- Dumay N, Gaskell MG. Sleep-Associated Changes in the Mental Representation of Spoken Words. *Psychological Science*. 2007; 18(1):35–39. [PubMed: 17362375]
- Dunn, LM.; Dunn, LM. *Peabody picture vocabulary test–4th edition*. San Antonio, TX: Pearson; 2007.
- Gaskell MG, Dumay N. Lexical competition and the acquisition of novel words. *Cognition*. 2003; 89(2):105–132. [PubMed: 12915296]
- Gershkoff-Stowe L, Hahn ER. Fast mapping skills in the developing lexicon. *Journal of Speech, Language, and Hearing Research*. 2007; 50:682–697.

- Gershkoff-Stowe, L.; Hahn, ER. Differential effects of lexical neighborhoods on receptive and productive word learning; Paper presented at the Symposium for Research on Child Development; Denver, CO. 2009.
- Gould W. Linear splines and piecewise linear functions. *Stata Technical Bulletin*. 1993; 15:13–17.
- Group USC. How can I get a Somers' D after logistic regression in Stata? 2014 from <http://www.ats.ucla.edu/stat/stata/faq/somersdlogit.htm>.
- Gueorguieva R, Krystal JH. Move over ANOVA: Progress in analyzing repeated measures data and its reflection in papers published in the Archives of General Psychiatry. *Archives of General Psychiatry*. 2004; 61:310–317. [PubMed: 14993119]
- Hardt O, Nader K, Nadel L. Decay happens: The role of active forgetting in memory. *Trends in Cognitive Sciences*. 2013; 23(3):111–120. [PubMed: 23369831]
- Hoffman L, Rovine M. Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods*. 2007; 39:101–117. [PubMed: 17552476]
- Hox, J. *Multilevel analysis: Techniques and applications*. 2nd edition ed.. New York: Routledge; 2010.
- Jarvis, BG. *DirectRT research software (Version 2002)*. New York, NY: Empirisoft; 2002.
- Kroll JF, Potter MC. Recognizing words, pictures, and concepts: A comparison of lexical, object, and reality decisions. *Journal of Verbal Learning & Verbal Behavior*. 1984; 23(1):39–66.
- Leach L, Samuel AG. Lexical configuration and lexical engagement: When adults learn new words. *Cognitive Psychology*. 2007; 55:306–353. [PubMed: 17367775]
- Locker L Jr, Hoffman L, Bovaird JA. On the use of multilevel modeling as an alternative to items analysis in psycholinguistic research. *Behavior Research Methods*. 2007; 39(4):723–730. [PubMed: 18183884]
- Luce PA, Pisoni DB. Recognizing spoken words: The neighborhood activation model. *Ear & Hearing*. 1998; 19:1–36. [PubMed: 9504270]
- McClelland JL, McNaughton BL, O'Reilly RC. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*. 1995; 102:419–457. [PubMed: 7624455]
- Merlo J, Chaix B, Ohlsson H, Beckman A, Johnell K, Hjerpe P, Larsen K. A brief conceptual tutorial of multilevel analysis in social epidemiology: using measures of clustering in multilevel logistic regression to investigate contextual phenomena. *Journal of epidemiology and community health*. 2006; 60:290–297. [PubMed: 16537344]
- Misangyi VF, LePine JA, Algina J, Goeddeke F. The adequacy of repeated measures regression for multilevel research: Comparisons with repeated measures ANOVA, multivariate repeated measures ANOVA, and multilevel modeling across various multilevel research designs. *Organizational Research Methods*. 2006; 9:5–28.
- Newson R. Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences. *Stata Journal*. 2002; 2:45–64.
- Nezlek JB, Schroder-Abe M, Schutz A. Multilevel analyses in psychological research. Advantages and potential of multi-level random coefficient modeling. *Psychologische Rundschau*. 2006; 57:213–223.
- Norman KA, O'Reilly RC. Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review*. 2003; 110:611–646. [PubMed: 14599236]
- O'Reilly RC, Rudy JW. Computational principles of learning in the neocortex and hippocampus. *Hippocampus*. 2000; 10:389–397. [PubMed: 10985278]
- Panis C. Linear splines and piecewise linear functions. *Stata Technical Bulletin*. 1994; 18:27–29.
- Quene H, van den Bergh H. On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication*. 2004; 43:103–121.
- Quene H, van den Bergh H. Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of memory & language*. 2008; 59(4):413–425.
- Rice ML, Oetting JB, Marquis J, Bode J, Pae S. Frequency of input effects on word comprehension of children with specific language impairment. *Journal of Speech & Hearing Research*. 1994; 37(1): 106–121. [PubMed: 8170118]

- Roodenrys S, Hinton M. Sublexical or lexical effects on serial recall of nonwords? *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 2002; 28(1):29–33.
- Stickgold R, Walker MP. Sleep-dependent memory triage: Evolving generalization through selective processing. *Nature Neuroscience*. 2013; 16:139–145.
- Storkel HL. Learning new words: Phonotactic probability in language development. *Journal of Speech, Language, and Hearing Research*. 2001; 44(6):1321–1337.
- Storkel HL. Learning new words II: Phonotactic probability in verb learning. *Journal of Speech, Language, and Hearing Research*. 2003; 46(6):1312–1323.
- Storkel HL. Methods for minimizing the confounding effects of word length in the analysis of phonotactic probability and neighborhood density. *Journal of Speech, Language, and Hearing Research*. 2004; 47(6):1454–1468.
- Storkel HL. A corpus of consonant-vowel-consonant (CVC) real words and nonwords: Comparison of phonotactic probability, neighborhood density, and consonant age-of-acquisition. *Behavior Research Methods*. 2013; 45:1159–1167. [PubMed: 23307574]
- Storkel HL, Adlof SM. Adult and child semantic neighbors of the Kroll and Potter (1984) nonobjects. *Journal of Speech, Language and Hearing Research*. 2009a; 52(2):289–305. doi: 1092-4388_2009_07-0174 [pii] 10.1044/1092-4388(2009/07-0174) [doi].
- Storkel HL, Adlof SM. The effect of semantic set size on word learning by preschool children. *Journal of Speech, Language and Hearing Research*. 2009b; 52(2):306–320. doi: 1092-4388_2009_07-0175 [pii] 10.1044/1092-4388(2009/07-0175) [doi].
- Storkel HL, Armbruster J, Hogan TP. Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research*. 2006; 49(6): 1175–1192.
- Storkel HL, Bontempo DE, Aschenbrenner AJ, Maekawa J, Lee SY. The effect of incremental changes in phonotactic probability and neighborhood density on word learning by preschool children. *Journal of Speech, Language, and Hearing Research*. 2013; 56:1689–1700.
- Storkel HL, Hoover JR. An on-line calculator to compute phonotactic probability and neighborhood density based on child corpora of spoken American English. *Behavior Research Methods*. 2010; 42:497–506. [PubMed: 20479181]
- Storkel HL, Lee SY. The independent effects of phonotactic probability and neighborhood density on lexical acquisition by preschool children. *Language and Cognitive Processes*. 2011; 26:191–211. [PubMed: 21643455]
- Tamminen J, Gaskell MG. Newly learned spoken words show long-term lexical competition effects. *Q J Exp Psychol (Colchester)*. 2008; 61(3):361–371. doi:783032454 [pii] 10.1080/17470210701634545 [doi].
- Tamminen J, Gaskell MG. Novel word integration in the mental lexicon: Evidence from unmasked and masked semantic priming. *Quarterly Journal of Experimental Psychology*. 2013; 66:1001–1025.
- Tamminen J, Payne JD, Stickgold R, Wamsley EJ, Gaskell MG. Sleep spindle activity is associated with the integration of new memories and existing knowledge. *The Journal of Neuroscience*. 2010; 30(43):14356–14360. [PubMed: 20980591]
- Thomson JM, Richardson U, Goswami U. Phonological similarity neighborhoods and children's short-term memory: typical development and dyslexia. *Memory & Cognition*. 2005; 33(7):1210–1219. [PubMed: 16532854]
- Thorn AS, Frankish CR. Long-term knowledge effects on serial recall of nonwords are not exclusively lexical. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 2005; 31(4):729–735. doi:2005-08130-010 [pii] 10.1037/0278-7393.31.4.729 [doi].
- Vlach HA, Sandhofer CM. Fast mapping across time: Memory processes support children's retention of learned words. *Frontiers in Psychology*. 2012; 3:1–8. [PubMed: 22279440]
- Windmeijer FAG. Goodness-of-fit measures in binary choice models. *Econometric Reviews*. 1995; 14:101–116.

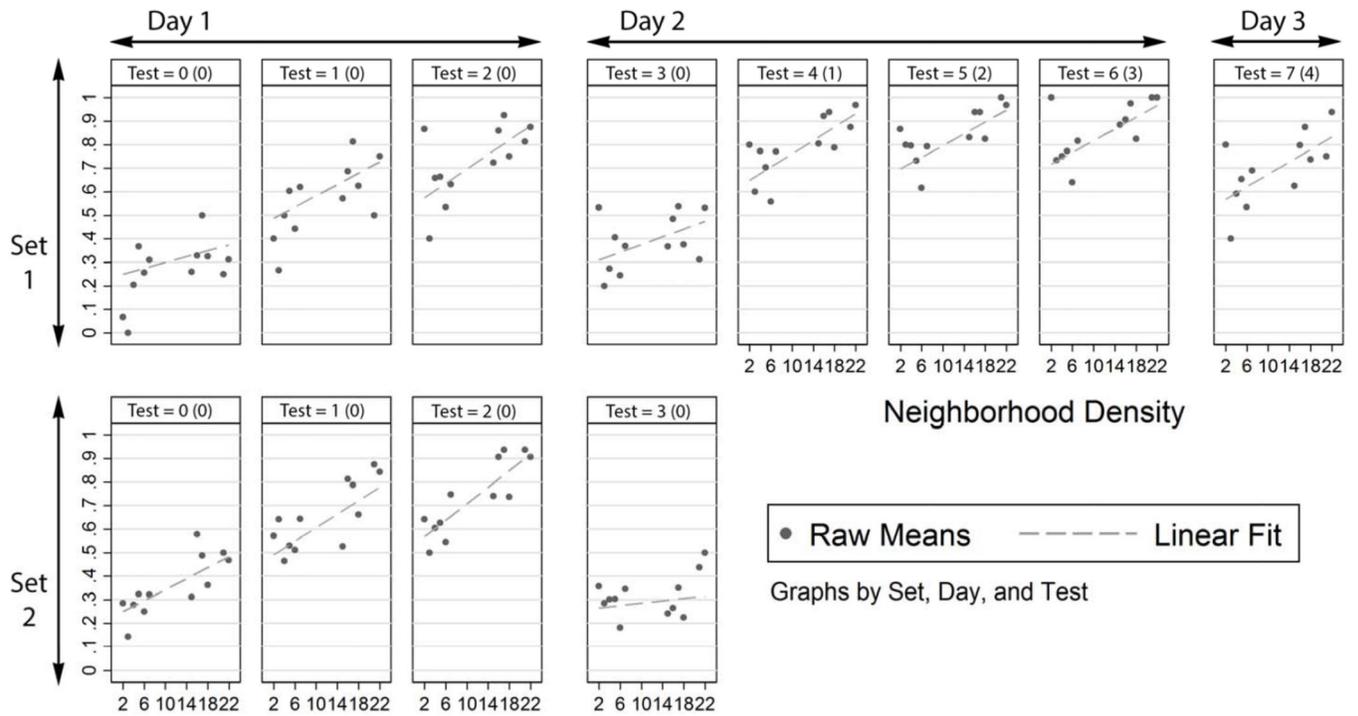


Figure 1. Probability of a correct response by the number of neighbors for the 1st set nonwords (top) and 2nd set nonwords (bottom). Each panel from left to right corresponds to a given test point: test 1, 2, 3, 4 (1-week gap), 5, 6, 7, 8 (1-week gap). The dashed linear fit line in each panel illustrates the effect of neighborhood density.

Table 1

Participant characteristics for each stimulus condition.

	Low Density Stimulus Condition	High Density Stimulus Condition
Number	29	32
Gender	Females = 24 Males = 5	Females = 26 Males = 6
Age ²	$M = 22$ years $SD = 3$ years Range = 18–29 years	$M = 22$ years $SD = 3$ years Range = 18–31 years
PPVT-4 ^{1, 2}	$M = 109$ $SD = 13$ Range = 89–136	$M = 115$ $SD = 12$ Range = 88–140
Race	White = 28 Asian = 1 American Indian = 0 Unknown = 0	White = 28 Asian = 2 American Indian = 1 Unknown = 1
Ethnicity	Non-Hispanic = 29 Hispanic = 0 Unknown = 0	Non-Hispanic = 30 Hispanic = 1 Unknown = 1

¹ Standard score on the *Peabody Picture Vocabulary Test – 4th edition* (Dunn & Dunn, 2007)

² Differences between the groups were not statistically significant, all $t(59) < 1.7$, all $p > 0.10$, $\eta_p^2 < 0.05$.

Table 2

Stimuli characteristics.

	Low Density <i>M (SD)</i>	High Density <i>M (SD)</i>
Group A		
Phonotactic Probability: Segment Sum	0.10 (0.02)	0.11 (0.02)
Phonotactic Probability: Biphone Sum	0.002 (0.001)	0.003 (0.001)
Neighborhood Density	6 (1)	17 (2)
Group B		
Phonotactic Probability: Segment Sum	0.11 (0.02)	0.11 (0.01)
Phonotactic Probability: Biphone Sum	0.002 (0.001)	0.003 (0.001)
Neighborhood Density	5 (2)	17 (2)

Note: Low Density and High Density do not differ significantly for segment sum, $F(1, 44) = 0.33, p > 0.50, \eta_p^2 < .01$; but do differ significantly for biphone sum, $F(1, 44) = 11.57, p = 0.001, \eta_p^2 = .21$, and neighborhood density (as intended), $F(1, 44) = 503.38, p < 0.001, \eta_p^2 = .92$.

Table 3

Final Multi-level Model.

Row #	Variable	Odds Ratio (95% confidence interval)	Reciprocal for OR < 1 ^I (95% confidence interval)
1	Gender	1.60 (1.00 – 2.56)	
2	PPVT-4 Standard Score	1.02 (1.01 – 1.04)**	
3	Semantic Set Size	0.82 (0.72 – 0.94)** I	1.22 (1.06 – 1.39)
4	All Tests (Test 1–8) Slope	3.75 (3.36 – 4.18)***	
5	Retention 1 Intercept (1-week gap 1)	0.04 (0.03 – 0.06)*** I	25.00 (18.18 – 34.48)
6	Later Tests (Tests 4–8) Slope	0.89 (0.77 – 1.01) I	1.13 (0.99 – 1.30)
7	Retention 2 Intercept (1-week gap 2)	0.03 (0.02 – 0.05)*** I	29.41 (18.52 – 45.45)
8	Neighborhood Density	1.03 (0.97 – 1.10)	
9	Density × All Tests Slope	1.04 (1.03 – 1.05)***	
10	Density × Retention 1 Intercept	0.89 (0.85 – 0.92)*** I	1.13 (1.09 – 1.17)
11	Density × Retention 2 Intercept	0.89 (0.85 – 0.94)*** I	1.12 (1.07 – 1.18)
12	Set	1.23 (1.06 – 1.43)**	
13	Set × Retention 1 Intercept	0.29 (0.22 – 0.38)*** I	3.45 (2.61 – 4.55)
14	MOR for Participants	1.92 (1.70 – 2.22)	
15	MOR for Items	3.00 (2.44 – 3.88)	
16	McKelvey & Zavoina's r^2	0.27	
17	Somers' D	0.50 (0.46 – 0.54)	

*
 $p < .05$ **
 $p < .01$ ***
 $p < .001$

All models included an overall intercept term that serves as the traditional constant.

^IFor OR < 1, a correct response is less likely for a higher than a lower value of the variable. The third column shows the reciprocal of the OR for ease of interpretation.

Table 4

Descriptive error analysis for input versus retention tests by density.

Phonemes Correct	Tests During Input (i.e., Tests 1, 2, 3, 5, 6, 7)		Retention Tests (i.e., Tests 4, 8)	
	Low Density Stimulus Condition	High Density Stimulus Condition	Low Density Stimulus Condition	High Density Stimulus Condition
0 of 3	<i>M</i> = 26% <i>SD</i> = 18%	<i>M</i> = 34% <i>SD</i> = 16%	<i>M</i> = 46% <i>SD</i> = 28%	<i>M</i> = 65% <i>SD</i> = 22%
1 of 3	<i>M</i> = 18% <i>SD</i> = 8%	<i>M</i> = 15% <i>SD</i> = 9%	<i>M</i> = 14% <i>SD</i> = 8%	<i>M</i> = 11% <i>SD</i> = 11%
2 of 3	<i>M</i> = 56% <i>SD</i> = 18%	<i>M</i> = 51% <i>SD</i> = 17%	<i>M</i> = 40% <i>SD</i> = 24%	<i>M</i> = 25% <i>SD</i> = 20%
2 of 3: Real Word Neighbors	<i>M</i> = 6% <i>SD</i> = 6%	<i>M</i> = 53% <i>SD</i> = 19%	<i>M</i> = 6% <i>SD</i> = 9%	<i>M</i> = 49% <i>SD</i> = 33%
0–2 of 3: Other Trained Nonword	<i>M</i> = 6% <i>SD</i> = 8%	<i>M</i> = 12% <i>SD</i> = 11%	<i>M</i> = 9% <i>SD</i> = 14%	<i>M</i> = 20% <i>SD</i> = 26%
0 of 3: No Response	<i>M</i> = 45% <i>SD</i> = 33%	<i>M</i> = 42% <i>SD</i> = 30%	<i>M</i> = 48% <i>SD</i> = 38%	<i>M</i> = 61% <i>SD</i> = 35%

Table 5

Descriptive error analysis for input versus retention tests by set.

Phonemes Correct	Tests During Input (i.e., Tests 1, 2, 3, 5, 6, 7)		Retention Tests (i.e., Tests 4, 8)	
	1 st Set	2 nd Set	1 st Set	2 nd Set
0 of 3	<i>M</i> = 28%	<i>M</i> = 35%	<i>M</i> = 51%	<i>M</i> = 62%
	<i>SD</i> = 20%	<i>SD</i> = 22%	<i>SD</i> = 30%	<i>SD</i> = 30%
1 of 3	<i>M</i> = 16%	<i>M</i> = 17%	<i>M</i> = 11%	<i>M</i> = 13%
	<i>SD</i> = 11%	<i>SD</i> = 14%	<i>SD</i> = 12%	<i>SD</i> = 13%
2 of 3	<i>M</i> = 56%	<i>M</i> = 48%	<i>M</i> = 37%	<i>M</i> = 25%
	<i>SD</i> = 21%	<i>SD</i> = 25%	<i>SD</i> = 29%	<i>SD</i> = 27%
2 of 3: 1 st Set Neighbor	N/A ¹	<i>M</i> = 7%	N/A ¹	<i>M</i> = 18%
		<i>SD</i> = 13%		<i>SD</i> = 32%
0–2 of 3: Other Trained Nonword	<i>M</i> = 11%	<i>M</i> = 9%	<i>M</i> = 11%	<i>M</i> = 23%
	<i>SD</i> = 13%	<i>SD</i> = 12%	<i>SD</i> = 18%	<i>SD</i> = 33%
0 of 3: No Response	<i>M</i> = 40%	<i>M</i> = 45%	<i>M</i> = 57%	<i>M</i> = 55%
	<i>SD</i> = 32%	<i>SD</i> = 39%	<i>SD</i> = 41%	<i>SD</i> = 40%

¹ 1st set Neighbor was only coded for the 2nd set items because the purpose was to quantify the extent to which the first trained items influenced responding during training of second set items. Note that items within a set were not neighbors of one another. Any confusion within a set is captured by the Other Trained Nonword category.