Effortful Retrieval Practice Effects in Lexical Access: A Role for Semantic Competition

Abhijeet Patra*, Hilary J. Traut, Mackenzie Stabile, Erica L. Middleton

Moss Rehabilitation Research Institute, 50 Township Line Rd., Elkins Park, PA, 190127, USA

*Correspondence concerning this article should be addressed to Abhijeet Patra, Research Department, Moss Rehabilitation Research Institute, 50 Township Line Rd., Elkins Park, PA, 19027, USA. E-mail: abhijeet.patra14@gmail.com.

Footnote: Hilary J. Traut is currently a graduate student at Department of Psychology & Neuroscience, University of Colorado Boulder, USA and Mackenzie Stabile is currently a graduate student at Department of Psychological Sciences, University of Connecticut, USA.

Effortful Retrieval Practice Effects in Lexical Access: A Role for Semantic Competition

Word retrieval difficulty (*lexical access deficit*) is prevalent in aphasia. Studies have shown that practice retrieving names from long-term memory (retrieval practice) improves future name retrieval for production in people with aphasia (PWA), particularly when retrieval is effortful. To explicate such effects, this study examined a potential role for semantic competition in the learning mechanism(s) underlying effortful retrieval practice effects in lexical access in 6 PWA. Items were trained in a blocked-cyclic naming task, in which repeating sets of pictures drawn from semantically-related versus unrelated categories underwent retrieval practice with feedback. Naming accuracy was lower for the related items at training, but next-day accuracy did not differ between the conditions. However, greater semantic-relatedness of an item to its set in the related condition was associated with lower accuracy at training but higher accuracy at test. Relevance to theories of lexical access and implications for naming treatment in aphasia are discussed.

Keywords: lexical access; retrieval practice; semantic blocking; naming impairment; aphasia

1. Introduction

Naming impairment is a common problem and impediment to functional communication in people with aphasia (PWA). Naming impairment manifests as frequent word substitutions (e.g., semantic error as in *giraffe* for *zebra*), distortions in the form of the word (e.g., *bobot* for *robot*), or outright response failures (i.e., omission) when attempting to name familiar, everyday objects, people, places, etc. *Lexical access deficit*, or difficulty retrieving words and/or their forms during naming, is a major contributor to naming impairment in aphasia (e.g., Schwartz et al., 2006). Evidence is amassing that practice retrieving names (e.g., for depicted objects) from long-term memory (hereafter, *naming practice*) is more beneficial to later naming accuracy in PWA compared to practice that does not involve retrieving names from long-term memory (e.g., Friedman et al., 2017; Middleton et al., 2015, 2016, 2019; Schuchard & Middleton, 2018a, 2018b). Furthermore, Middleton et al. (2016) found that naming practice that was more effortful conferred more durable learning (defined in Section 1.2) in PWA.

This new and growing evidence base regarding the effects of naming practice and retrieval effort on lexical access has far outpaced theoretical explication of such effects. The current study takes a step towards addressing this theory gap by evaluating a potential role for semantically-driven lexical competition (hereafter, *semantic competition*) in the learning mechanism(s) underlying effortful retrieval effects in lexical access.

1.1 Semantic Context Effects in Naming

An extensive literature on *semantic context effects* indicates that picture naming (e.g., zebra) makes subsequent naming of items from the same semantic category (e.g., giraffe) more effortful, manifesting as increased naming error rates and/or latencies. Semantic context effects in naming have been extensively studied in neurotypical speakers, and less so in PWA, using the semantic variant of the blocked-cyclic naming task (e.g., Belke, 2008, 2013; Belke et al., 2005; Belke & Stielow, 2013; Biegler et al., 2008; Damian et al., 2001; Damian & Als, 2005; Harvey & Schnur, 2015; McCarthy & Kartsounis, 2000; Patra et al., 2021; Schnur et al., 2006; Wilshire & McCarthy, 2002). In this task, participants name repeating sets of pictures drawn from either the same category (*homogeneous condition*) or multiple categories (*mixed condition*). Typically, a set is presented in a 'block' comprised of multiple (usually 6) successive 'cycles', and the items in the set are presented in random order in each cycle. A *semantic blocking effect* manifests as slower naming latencies and/or more naming errors in the homogeneous compared to the mixed context (e.g., Belke et al., 2005; Damian et al., 2001; Harvey & Schnur, 2015; Schnur et al., 2006). In addition, several studies have observed cumulative semantic interference – growing decrement in naming accuracy or increment in naming latency – across cycles (Harvey & Schnur, 2015; Schnur et al., 2006; but see Belke, 2008; Belke & Stielow, 2013). Also, the semantic blocking effect does not diminish with additional time between trials (Biegler et al., 2008; Schnur et al., 2006) or intervening trials in the blocks (Damian & Als, 2005; Navarrete et al., 2012). Semantic context effects have also been extensively studied in a related paradigm termed the continuous naming task (Howard et al., 2006), in which multiple exemplars from each of several categories are presented serially and in intermixed fashion for naming in a large list. In continuous naming, the effect of semantic context manifests as an incremental, cumulative increase in naming difficulty (e.g., in latencies or errors) with the presentation of each additional exemplar in a category (termed *ordinal position effect*; e.g., Belke, 2013; Harvey et al., 2019; Howard et al., 2006; Navarrete et al., 2010). The overall consensus in the literature is that semantic context effects – the semantic blocking effect in blocked-cyclic naming and the ordinal position effect in the continuous naming task – arise from a learning process that persistently decreases the accessibility of related items, at least within the timeframe of the task (Howard et al., 2006; Oppenheim et al., 2010).

Naming is a complex process that begins with visual recognition/categorization of the object, followed by mapping from the encoded meaning (i.e., semantics) to a word, retrieval and encoding of the word's phonology, and finally, articulation. The stages dedicated to mapping from semantics to a word, and from the word to phonology, are typically regarded as the two main stages of lexical access (e.g., Dell, 1990; Dell et al., 1997; Dell & O'Seaghdha, 1992; Dell & Reich, 1981; Fay & Cutler, 1977; Fromkin, 1971; Garrett, 1975, 1976; Levelt et al., 1999; Rapp & Goldrick, 2000; Schwartz, 2014; Stemberger, 1985; cf., Caramazza, 1997; Caramazza & Miozzo, 1997). Studies collectively have identified that semantic blocking effects in naming localize to the mapping from semantics to a word (e.g., Damian et al., 2001; Goldrick & Rapp, 2007; Kroll & Stewart, 1994; Vigliocco et al., 2002). For greatest experimental sensitivity in our study, we recruited people with aphasia whose naming impairment can be attributed, at least in part, to problems retrieving words from semantics, as well as in retrieving phonology, as opposed to solely arising from processes that are peripheral to lexical access (see Section 2.1). We revisit the issue relating to the characterization of our participants and learning effects in the current study in the Discussion (Section 4).

Several studies have observed that the deleterious effect of semantic context on naming is enhanced with increasing semantic similarity with previously named items. In blocked-cyclic naming, Navarrete et al. (2012) reported an enhanced semantic blocking effect for sets composed of more semantically similar (e.g., cat and dog) versus less semantically similar (e.g., cat and zebra) category members. Likewise, Vigliocco et al. (2002) reported an enhanced semantic blocking effect when sets were composed of items drawn from more similar (e.g., clothing and body parts) versus less similar (e.g., clothing and vehicles) categories. Studies using the continuous naming task have likewise found that the degree of semantic similarity between items affects the magnitude of interference from prior naming (Harvey et al., 2019; Rose & Abdel Rahman, 2017; see Alario & del Prado Martín, 2010 for a discussion). For example, in a recent study involving PWA, Harvey et al. (2019) found higher similarity between the first exemplar (ordinal position 1) and second exemplar (ordinal position 2) in a category that were presented in a session resulted into heightened semantic error rates at ordinal position 2.

To summarize, prior semantic context can enhance naming difficulty in a persistent manner (at least within the timeframe of the task); such effects localize to the mapping from semantics to words; and, semantic context effects are enhanced with greater semantic similarity of prior trials to the current naming trial. A possibility that has yet to be examined, however, is whether training naming amidst enhanced semantic competition, via blocked-cyclic naming, may ultimately promote more persistent gains from naming training, and/or enhanced accuracy measured at a later session. The following sections consider an empirical basis (Section 1.2), followed by the theoretical motivation (Section 1.3), for this possibility.

1.2 Retrieval-based Learning Effects in Lexical Access

The field of aphasiology has demonstrated growing interest in how basic research on fundamental learning mechanisms can help elucidate the treatment process and improve efficacy. Inspired by the neuroscientific principle of Hebbian learning (Hebb, 1949), pioneering studies by Fillingham and colleagues (Fillingham et al., 2005a, 2005b, 2006) examined an 'errorless learning' naming treatment for aphasia whereby on each training trial, the object for naming was presented along with its name, and the name was repeated by the patient. This approach was designed to capitalize on the Hebbian notion that cell arrays that fire together, wire together by assuring the correct response (object's name) given the stimulus (depicted object) on every trial. Fillingham et al. (2005a, 2005b, 2006) compared errorless learning to 'errorful' naming treatment, in which the participant was encouraged to attempt to retrieve the name for the object with cueing support (e.g., presentation of word onset), often leading to naming error. Whether the correct name was provided as feedback after errorful trials was variable across the studies. Single-case analyses in each study revealed that most PWA benefitted from both types of training approaches despite substantially higher error rates during errorful treatment (Fillingham et al., 2005a, 2005b, 2006; see also, Conroy et al., 2009). A later study (McKissock & Ward, 2007) revealed that errorful learning provided the same benefit as errorless learning across a group of PWA, but only when correct-answer feedback was provided following the naming attempt.

Middleton et al. (2015) revisited errorless learning naming treatment for aphasia, but compared it to a retrieval-based naming treatment. In contrast to the typical errorful treatment from prior studies (e.g., Fillingham et al., 2005a, 2005b, 2006), the retrieval-based naming treatment was informed by best practices derived from the retrieval practice (a.k.a. test-enhanced learning) literature (for recent reviews, see Kornell & Vaughn, 2016; Rowland, 2014), including a focus on correct retrieval during treatment and consistent provision of feedback. The results from Middleton et al.'s study showed that though the rate of production of the name was highest in the errorless learning condition during training, both naming practice conditions outperformed the errorless learning condition on a next-day test of naming (hereafter, *delayed test of naming*), with the advantage persisting for the cued naming practice condition after one week. This constituted the first empirical demonstration that retrieval practice—a learning factor examined primarily in the context of knowledge acquisition—can persistently enhance the retrievability of existing lexical representations for production (i.e., lexical access). In Middleton et al., indications that retrieval practice impacted lexical access included that neuropsychological characterization of the PWA was consistent with lexical access deficit underlying their naming impairment, and the study materials were pictures of familiar, common objects (e.g., scissors; caterpillar; pizza) with high name agreement.

Two features of retrieval practice are important to consider in research seeking to design interventions or training regimens that maximize the benefits from retrieval practice. First, the potency of retrieval practice training is driven mainly by correct retrievals; failed retrievals, even followed by correct-answer feedback, confer detectable but weak learning (Dunlosky & Rawson, 2012; Kornell et al., 2011; Middleton et al., 2015; Pashler et al., 2005; Wissman & Rawson, 2018). Second, information retrieved under more effortful conditions receives greater strengthening, i.e., learning is more *durable* (e.g., Karpicke & Bauernschmidt, 2011; Karpicke & Roediger III, 2007; Pashler et al., 2003; Pyc & Rawson, 2009). The signature pattern of more durable learning from enhanced retrieval effort is typically demonstrated with an interaction between training condition and time (training versus test) with opposing patterns of performance—higher error rate at training but better performance at test for the effortful condition (Karpicke & Roediger III, 2007; Middleton et al., 2016; Pashler et al., 2003; for discussion, see Schmidt & Bjork, 1992). That is, although making training more difficult means fewer items, or fewer trials per item, benefit from the strengthening that successful retrieval practice confers, the information that is successfully retrieved under more effortful conditions receives greater strengthening. This greater strengthening can ultimately confer enhanced test performance in the more effortful condition.

To evaluate whether more effortful retrieval confers more durable learning in aphasia treatment, Middleton et al. (2016) compared naming practice versus errorless learning in a group of PWA with lexical access deficit but additionally examined how the spacing of trials impacted later performance. For present purposes, the most illustrative aspect of that study involved manipulating the number (i.e., *lag*) of other-item trials between repeated naming attempts for an item. Presenting items at different degrees of spacing (lag 5, 15, or 30) permitted examination of how increased retrieval effort with increased spacing in the naming condition affected training and test performance. First, Middleton et al. observed an interaction indicating that though naming practice success rate at training dropped precipitously as the spaced schedule lags increased (reflective of more difficult retrieval with increasing lag), performance on the delayed tests was similar across the spaced lags, consistent with greater strengthening from retrievals at higher lags. The strongest evidence for an effect of effortful retrieval on later performance was reported in an analysis that statistically controlled for differences in training performance across lags. In that analysis, increasing lag was associated with *increasing* delayed test performance. In other words, naming practice that is more effortful for people with aphasia can come at the cost of heightened errors during training, but successful retrieval trials confer greater strengthening under more effortful training conditions.

1.3 Learning from Inhibition

To advance a mechanistic understanding of effortful retrieval effects in lexical access, we consider a well-researched phenomenon in the memory and learning literature, specifically *retrieval-induced forgetting* (RIF). RIF studies have shown that retrieving a target from long-term memory (FRUIT-O____, answer: *orange*) decreases subsequent retrievability of related items, i.e., competitors (FRUIT-B_____, answer: *banana*; for reviews, see Murayama et al., 2014; Storm & Levy, 2012; Verde, 2012). Controversy surrounds whether RIF manifests because competitors are inhibited when the target is retrieved, or because the target is strengthened from retrieval, which interferes with subsequent retrieval of its competitors (for debate, see Anderson, 2003; Raaijmakers & Jakab, 2013; Storm & Levy, 2012; Verde, 2012). However, features of RIF point to a role for inhibition. For example, counter to the interference account, not just any strengthening event creates RIF; rather, RIF is specific to retrieval practice (e.g., studying FRUIT-ORANGE does not decrease retrievability of FRUIT-B_____, answer: banana; Anderson et al., 2000; Bäuml, 2002). For our purposes, most important are observations that greater inhibition is conferred as a competitor is more (versus less) related to the category (Anderson et al., 1994; Storm et al., 2005, 2007), and inhibition from RIF can *potentiate* learning (Storm et al., 2008). Specifically, Storm et al. found items that are first inhibited via RIF and then strengthened (i.e., via restudy) are *more* retrievable later relative to items that do not undergo inhibition before strengthening. Furthermore, the superior retrievability of previously inhibited (versus non-inhibited) items was found to accumulate with each inhibition-study cycle, a phenomenon Storm et al. dubbed *accelerated relearning*.

Now turning to the lexical access literature, as we reviewed in Section 1.1, semantic context effects in naming implicate learning. A prominent, computationally explicit framework for understanding such learning is the *dark-side model* of incremental learning in lexical access (Oppenheim et al., 2010). In the dark-side model, following each naming trial, a learning algorithm strengthens the retrieval connections between semantics and the target word (the *light* side of retrieval), and weakens connections to competitors, i.e., words concurrently active via overlapping semantics with the target (the *dark* side of retrieval). Importantly, learning is *error-based* in that the degree of weight change is driven by how over- (i.e., competitor) or under- (i.e., target) activated each word node was relative to a desired ("correct") activation value. Coupling the notion of error-based learning with accelerated relearning, we consider the possibility that (a) the greater cyclic inhibition and strengthening of items in homogeneous (versus mixed) sets in blocked-cyclic naming should ultimately confer more durable learning in naming, and (b) the degree of relatedness of items in a homogeneous set should also relate to the durability of learning.

1.4 Present Study

As this is a training study, we first identified training items for each PWA that elicited naming error from a large picture corpus of common, everyday objects. Different sets of items were trained in a homogeneous versus a mixed context in blocked-cyclic naming in each of seven rounds. Each round comprised a training session and a next-day delayed test of naming of the items trained in that round. Correct-answer feedback (target name was auditorily presented) was provided after each naming attempt during training.

In the present design, if greater semantic competition enhances retrieval effort, the homogeneous condition should be associated with enhanced naming error rates during training compared to the mixed condition (i.e., semantic blocking effect). Furthermore, if the enhanced effort from training in a semantic context confers more durable learning, we expect to see the signature interaction of training condition (homogeneous versus mixed) and time (training versus delayed test of naming) on accuracy with enhanced delayed test accuracy for the homogeneous condition compared to the mixed. This pattern, which was observed in a prior study of effortful retrieval effects in lexical access (Middleton et al., 2016), would constitute strong evidence for a role for semantic competition in conferring more durable learning. However, depending on how our effortful retrieval manipulation is situated with regards to the trade-off between greater strengthening versus greater rates of failed retrieval during training (for discussion see Bjork, 1994; Pashler et al., 2003), we may observe similar levels of accuracy at the next-day test in the two conditions.

Next, because increasing semantic similarity between items in a set increases naming difficulty (Navarrete et al., 2012; Vigliocco et al., 2002), at training, we would expect poorer accuracy for items in homogeneous sets as the semantic similarity of an item to its set-mates increases. However, according to the effortful retrieval hypothesis, this greater difficulty should confer more durable learning, resulting in an interaction between an item's similarity to its set members and time (training versus delayed test), with enhanced test accuracy with increasing similarity of an item to its set-mates. Lastly, we report the standard indices of semantic blocking in accuracy and latencies, specifically the effect of context and possible accumulation of semantic interference across cycles at training, to contribute to the relatively small literature on semantic context effects in PWA (Biegler et al., 2008; Harvey et al., 2019; Harvey & Schnur, 2015; McCarthy & Kartsounis, 2000; Schnur et al., 2006; Scott & Wilshire, 2010).

2. Method

In comparison to neurotypical adults, studies involving individuals with neurological damage (e.g., people with stroke aphasia) require a strategy of achieving experimental sensitivity in the face of greater between-participant and within-participant variability. For example, PWA of even the same aphasia subtype (e.g., Broca's aphasia) can show great variability in their residual cognitive and linguistic skills, which can interact in unpredictable ways with experimental manipulations. In addition, within an individual, increased variability in performance from one trial to another within a task is a hallmark feature of neurological damage (MacDonald et al., 2006). In the present study, we addressed these challenges by (1) including participants with a relatively homogeneous profile in terms of their cognitive-linguistic deficits, and (2) designing the study to confer a large number of observations per condition per participant. We have adopted similar strategies in our prior work to provide stable results within and across participants (Middleton et al., 2015, 2016, 2019, 2020). For example, Middleton et al. (2016; 2019) showed statistically robust learning effects in a participant sample of four PWA with approximately 50 observations per condition per participant. With these studies as benchmarks, we set our recruitment goal for the current experiment at six PWA, with a more ambitious target of 84 observations per condition per participant.

2.1 Participants

Six participants were recruited from the Moss Rehabilitation Research Institute Participant Registry. All participants gave informed consent under a protocol approved by the Institutional Review Board of Einstein Healthcare Network, and were reimbursed $15 per hour of participation.

The inclusion criteria for the study were age range between 21-80 years, have English as their native or primary language, and give evidence of having the linguistic and cognitive capacity to understand the consent form and give informed consent. Participants were included without respect to gender, race, or ethnic background. Table 1 provides demographic and neuropsychological characteristics of the participant sample, which comprised 3 males and 3 females. Mean age was 51.7 years (*SD* = 13.5), all participants were pre-morbidly right-handed with one exception (participant 3), and mean education level was 14.7 years (*SD* = 2.6). All participants were diagnosed with post-stroke aphasia in the chronic phase as determined by the Western Aphasia Battery (WAB) Aphasia Quotient (AQ) (Kertesz, 2007).

The study participants were selected from a large (>100) pool of previously characterized and potentially available people with chronic post-stroke aphasia. These participants were prioritized for recruitment because they were able to commit to the months-long protocol, and their neuropsychological profile was consistent with detectable naming impairment attributable, at least in part, to lexical access deficit. The six participants presented with mild to moderate naming impairment on the Philadelphia Naming Test (Roach et al., 1996). The sample showed no worse than mild impairment on tests of nonverbal semantic comprehension (Pyramids & Palm Trees test; Howard & Patterson, 1992) and word comprehension (spoken word-to-picture verification task; Roach et al., 1996), suggesting deficits in semantics or lexical-semantics was not a major contributor to their naming impairment. The sample also demonstrated good or very good word repetition, suggestive of minor contribution of post-lexical encoding or articulation problems to their naming impairment (see Table 1). No participant exhibited worse than moderate apraxia of speech. Appendix A provides a breakdown, per participant, of naming error types on the large set of items administered in the item selection task (described in Section 2.2.1). Some incidence of phonological error in naming was present across the sample, but naming errors consistent with an impairment in word retrieval (i.e., semantic errors, descriptions, and no response errors; Chen et al., 2019; Schwartz et al., 2009) were most prominent.

<<Insert Table 1>>

2.2 Materials and Procedure

To enhance experimental sensitivity, a large picture corpus was used to select training items for each participant that elicited naming error prior to the main study. The corpus comprises 660 unique common objects (hereafter, *660-item set*) collected from published picture corpora (Brodeur et al., 2010; Szekely et al., 2004) and various internet sources. Items in the corpus are characterized by several variables that can affect naming including visual complexity, name agreement, log word frequency, number of phonemes, and number of syllables. Visual complexity and name agreement values were collected from published corpora when available; otherwise, these values were obtained in normative studies with a minimum of 40 responses per item. Mean name agreement for the 660-item set is 93% (*SD* = 6%; range = 80-100%). Log frequency values for the picture names were collected from SubtlexUS (Brysbaert & New, 2009). Picture names that did not appear in SubtlexUS were assigned a log frequency value of zero. Audio recordings of the picture names were created by a female native English speaker.

The picture corpus was divided into 19 categories of related items informed by category production norms (Van Overschelde et al., 2004) and experimenter intuition. The goal was to divide the 660-item corpus into a large number of categories, each comprised of a large number of items, to increase the chances of obtaining a sufficient number of errorful items for a sufficient number of categories to populate the design for each participant (see Section 2.2.1). The categories were organized around items forming natural kinds or taxonomies (e.g., fruits and vegetables, body parts), or thematically and/or functionally related groups (e.g., accessories, toys and games, office supplies). The range of exemplars across categories was 18-43 items. There were 78 items that did not belong to any category (i.e., *uncategorized items*), some of which were used as fillers (see Section 2.2.2). Table 2 lists the 19 related categories with sample category members.

<<Insert Table 2>>

2.2.1 Item Selection Task

In the item selection task, the 660-item set was administered in its entirety for naming twice, one administration per week on different weeks preceding the main experiment. On each naming trial, the participant was shown the picture and instructed to name the picture to the best of their ability. They were provided 20 seconds to do so, after which the software automatically advanced; or, if the participant indicated they were finished attempting to name the picture, the experimenter advanced the trial prior to the end of 20 seconds. This procedure developed in our prior work (e.g., Middleton et al., 2016) was instituted to eliminate experimenter feedback of any kind regarding the potential correctness of the naming response.

To identify items for training, we selected the 14 categories with the highest proportion of items that were errorful across both administrations of the item selection task for a participant. Within each of the 14 selected categories, the 12 most consistently errorful items were selected for training. This resulted in a number of items selected for training that were accurately named once or twice during item selection. For a participant's selected categories, the 12 selected items per category were randomly assigned into the homogeneous and mixed conditions while controlling for item selection naming accuracy, log frequency, visual complexity, number of phonemes, name agreement, and number of syllables (see Table 3). Mixed sets were comprised of 6 exemplars from different categories. When necessary, the sets were manually altered so that no items within a single set shared a phonological onset.

<<Insert Table 3 here>>

For all homogenous items selected for training for each participant, an item's semantic similarity to each of its set mates was estimated in a pairwise fashion using latent semantic analysis (LSA; Landauer et al., 1998), and an item's mean semantic similarity across its set mates was derived (hereafter, *item-to-set semantic similarity*). Table 4 provides an example of LSA-based item-to-set semantic similarity estimates for a hypothetical homogeneous set.

<<Insert Table 4 here>>

2.2.2 Training and Delayed Test of Naming Sessions

In the main experiment, participants underwent seven 'rounds', with each round comprising a training session and a next-day delayed test of naming of items trained in the prior session. For each participant, each round occurred in a different week. The training session in each round was devoted to training two homogeneous sets, which were unrelated to each other, and two mixed sets.[1] Individual sets were trained in one round only for a participant. In a training session, all items across the two mixed sets were from different categories, and those categories were unrelated to the two homogeneous categories also trained in that session. Order of the conditions within a session were counterbalanced across the seven training sessions and across

[1] Due to experimenter error, one participant received training on three homogeneous sets and one mixed set in Round 1, and three mixed sets and one homogeneous set in Round 2.

participants. Within a session, each set underwent five sequential cycles of naming in which items in a set were presented in pseudo-random order with the constraint that the same item was not presented contiguously across two cycles. At the onset of each training trial, the depicted object was displayed, and the participant was provided 8 seconds to attempt to name the object. This was immediately followed by feedback, where the target name was auditorily presented and the participant repeated the name, after which the next trial was initiated.

In each delayed test of naming, the 24 critical items from the preceding training session were tested but they were distributed among 25 filler items in a pseudo-random order with the constraint of a minimum of 6 trials for other items between any two category members. Filler items were selected from the remaining items in the 660-item corpus that were not selected for training for a participant. Different fillers were used in each of the rounds for a participant. The addition of these filler items was intended to mitigate the potential for testing itself to instantiate a semantic context effect such as that observed in continuous naming (Howard et al., 2006).

Delayed test of naming trial structure followed the procedure used during item selection (see Section 2.2.1). To permit off-line measurement of naming response latencies on correct trials, simultaneous with picture presentation on each trial during training and test, the experimental software played a beep to mark the start of the trial. All sessions were recorded and transcribed into IPA for analysis by a trained expert. Including the item selection phase, mean time of participation was M = 15.2 (SD = 2.3) weeks and M = 17.3 (SD = .74) total sessions per participant.

2.3 Analyses

All participants completed seven rounds except participants 4 and 6. Participant 4 missed the round 6 delayed test and participant 6 missed the round 2 delayed test, both due to inclement weather. As a consequence, the data for the corresponding training sessions for these two participants were dropped from the analyses. The procedure produced 4800 training trials (i.e., 7 rounds x 4 sets x 6 items x 5 cycles x 4 participants + 6 rounds x 4 sets x 6 items x 5 cycles x 2 participants) and 960 delayed test trials (i.e., 7 rounds x 4 sets x 6 items x 4 participants + 6 rounds x 4 sets x 6 items x 2 participants) after excluding trials for filler items. With the exception of participants 4 and 6, the design produced 84 observations per condition per participant.

Naming accuracy and naming onset latency (correct trials only) were calculated based on the participant's first complete, non-fragmented naming attempt per trial. To determine naming accuracy, phonological overlap (Lecours & Lhermitte, 1969; see formula below) between the naming attempt and target name was first calculated. Phonological overlap provides a continuous measure of phonological similarity to a target that is standardized across different word lengths. Shared phonemes were identified independent of position, and credit was assigned only once if a response had two instances of a single target phoneme (e.g., /kakt/ for cat is not considered correct). Semantic errors and descriptions (including all non-noun responses) received an overlap score of zero so as to avoid rewarding coincidental phonological similarity to a target. A response was coded as correct if phonological overlap was equal to or greater than 0.75; responses with phonological overlap less than 0.75 were considered incorrect. For accuracy, including the item selection phase, training and test phase, the protocol produced a total of 14,680 hand-coded responses across all six participants.

$$\text{Phonological overlap} = \frac{\text{number of shared phonemes in target and response } \times 2}{\text{total number of phonemes in target and response}}$$

To measure onset latency on correct naming trials, trained research staff used Praat software (Boersma & Weenink, 2016) to view the formants and glottal pulses of the responses. Onset latency was calculated from the trial-onset beep to the first glottal pulse that extended through at least two formants for voiced segments, and to the first visible increase in energy due to sound for unvoiced segments. For latency, including the training and test phase, the protocol produced a total of 4,715 hand-coded responses across all six participants. In preparation for the latency analyses, we removed outliers using the mean absolute deviation (MAD) method (for the upper range: +6SD from median, for the lower range: -3SD from median) and log-transformed the latencies (Leys et al., 2013; Wiley & Rapp, 2019).

Naming accuracy was modelled with mixed logistic regression using the glmer function in R version 3.6.0 (R Core Team, 2019) with alpha = .05 for tests of significance. To evaluate whether greater effort at training (homogeneous versus mixed) leads to more durable learning, we assessed a potential interaction of a two-level factor of condition (homogeneous versus mixed) and a two-level factor of time (training versus test) on naming accuracy (correct/incorrect response). Sum contrasts were applied to the condition (+1 for mixed and -1 for homogenous) and time (+1 for training and -1 for test) factors. A significant interaction was followed by simple-effects models to inspect potential effects of the condition factor at each timepoint. To evaluate whether greater effort due to higher item-to-set semantic similarity (defined in Section 2.2.1) within the homogenous set confers more durable learning, we assessed a potential interaction of the time factor and item-to-set semantic similarity entered as a numerical fixed effect, with the significant interaction followed with simple-effects models to inspect potential effects of the item-to-set semantic similarity variable at each timepoint. Though not of a priori interest, the same sequence of analyses was applied to naming onset latencies for correct trials using mixed linear regression. For completeness, the model results applied to naming onset latencies are reported in Appendix E1-E2. Finally, an analysis of forgetting to assess retention of accuracy performance from training to test in the homogeneous and mixed conditions was conducted (details in Section 3.1).

Item-specific variables (i.e., covariates) that can affect naming but are not of theoretical interest (log frequency, syllable length, number of phonemes, visual complexity, and name agreement; see Section 2.2.1) were entered as fixed effects in all models but were dropped if not significant. Random intercepts for participants and items were included in all models to capture the correlation among observations that can arise from multiple participants giving responses to overlapping sets of items. By-participant random slopes for the experimental factors were also included if they improved model fit by a chi-square test of deviance in model log likelihood (alpha = .05). Naming accuracy model results are reported in Tables 5 and 6. Models examining classic indices of semantic context effects in blocked-cyclic naming (i.e., semantic blocking effect; cumulative semantic interference) are described in Section 3.3. Lastly, for readers interested in more classic indices of treatment effects (i.e., pre to post-treatment change), models reporting change in naming accuracy from item-selection to the delayed test across the group (mixed logistic) and per participant (simple logistic) in the homogeneous and mixed conditions are reported in Appendix B. All participants showed significant improvement in both conditions.

3. Results

3.1 Interaction of Time and Condition

The results revealed a significant interaction of the time factor and the condition factor (estimate = 0.10, SE = 0.05, *Z* = 2.12, *p* = .03; Table 5). Figure 1 presents mean naming accuracy across the participants in the mixed and homogeneous conditions at the training and test timepoints. The simple-effects model applied to training performance revealed a significant decrement in naming accuracy in the homogeneous condition compared to the mixed condition (estimate = 0.11, SE = 0.04, *Z* = 2.82, *p* = .005; Table 5). This finding is in line with the existing semantic blocking literature--naming items in a homogeneous versus mixed context is associated with heightened naming error. However, the simple-effects model applied to test showed no decrement in naming accuracy in the homogeneous condition compared to mixed (estimate = -0.09, SE = 0.09, *Z* = -1.03, *p* = .30; Table 5). In fact, numerically, naming accuracy at test was higher for the homogeneous condition compared to the mixed condition.

One way to examine differential strengthening of retrieved information via effort manipulations is to examine 'forgetting' (e.g., Roediger III & Karpicke, 2006). In the present study, this involved examining the rate of change in naming accuracy going from training to test for each condition separately. The results revealed a marginal decrement in naming accuracy going from training to test for the mixed condition (estimate = 0.12, SE = 0.07, *Z* = 1.85, *p* = .06; see Appendix C for full model) but not for the homogenous condition (estimate = -0.09, SE = 0.07, *Z* = -1.26, *p* = .20; see Appendix C for full model). In fact, the homogenous condition showed a numerical improvement (i.e., a gain of 2.5 %) in naming accuracy going from training (naming accuracy = 0.823) to test (naming accuracy = 0.844). We provide interpretation of the forgetting findings and the time by condition interaction in the Discussion.

<<Insert Table 5 here>>

<<Insert Figure 1 here>>

3.2 Interaction of Time and Item-to-Set Semantic Similarity

The results revealed a significant interaction between item-to-set semantic similarity and time for naming accuracy (estimate = -2.10, SE = 0.76, *Z* = -2.76, *p* =.005; Table 6, Figure 2). The simple-effects model applied to training performance revealed a significant decrement in naming accuracy as item-to-set semantic similarity increased (estimate = -1.28, SE = 0.57, *Z* = -2.21, *p* =.02; Table 6), an effect similar to that observed in other semantic blocking studies (Navarrete et al., 2012; Vigliocco et al., 2002). However, a finding that heretofore has not been examined or reported, at the delayed test, as an item's semantic similarity to its set increased, naming accuracy increased (estimate = 3.11, SE = 1.53, *Z* = 2.03, *p* =.04; Table 6). This suggests that a homogenous item with greater semantic similarity with its set – despite having less opportunity to be retrieved successfully during training – receives greater strengthening from the enhanced effort that is required when retrieved amongst greater versus lesser semantic competition.

<<Insert Table 6 here>>

<<Insert Figure 2 here>>

3.3 Classic Indices of Semantic Blocking Effects

As described in Section 3.1, we observed the standard semantic blocking effect in the form of a significant decrement in naming accuracy for the homogeneous condition compared to the mixed condition during training. To examine whether the difference between conditions grew across cycles, we modelled a cycle-by-condition interaction, using linear contrasts for cycle (see Schad et al., 2020) and sum coding for the condition factor. The cycle-by-condition interaction was significant (estimate = 0.49, SE = 0.13, *Z* = 3.93, *p* <.001; see Appendix D for model output, and Appendix F for naming accuracy means as a function of condition and cycle). However, it is problematic to interpret this interaction as evidence for cumulative semantic interference because the homogeneous and mixed sets did not differ in accuracy at Cycle 1 (estimate = -0.02, SE = 0.07, *Z* = -0.35, *p* =.72; Appendix D). This likely reflects the within-block semantic priming that can offset naming difficulty at Cycle 1 for homogeneous sets compared to mixed sets (for discussion, see Belke & Stielow, 2013). When Cycle 1 was dropped from the analysis, the interaction of cycle-by-condition was no longer significant (estimate = -0.04, SE = 0.12, *Z* = -0.30, *p* =.76; Appendix D). In other words, the semantic blocking effect in accuracy did not grow across cycles 2-5. Following the same analysis trajectory for latencies, including examination of simple-effects only in the presence of a significant interaction, there was no condition by time interaction (*p* = .27; Appendix E1) and no cycle-by-condition interaction (*p* = .88; see Appendix E3 for model output, and Appendix G for naming latency means as a function of condition and cycle).

4. Discussion

The goal of the present study was to examine a potential role for semantic competition in the theoretical explication of effortful retrieval practice effects in lexical access. To do this, the current study probed the durability of learning from training of errorful naming items for people with aphasia amidst more versus less semantic competition using the blocked-cyclic naming paradigm as a training intervention.

With regards to naming accuracy, we observed a significant interaction of time and condition, with lower accuracy during training in the homogeneous condition compared to mixed but no difference in accuracy for the two conditions at test. Also, an analysis of forgetting revealed a trend for better retention of performance from training to test in the homogeneous versus the mixed condition. However, the marginal nature of the forgetting effect and lack of a difference between the conditions at test constitutes a failure to provide strong evidence for the effortful retrieval hypothesis, and does not align with reports of greater test performance in the more effortful condition in other studies of effortful retrieval learning effects (e.g., Karpicke & Roediger III, 2007; Middleton et al., 2016; Pashler et al., 2003). We next consider two explanations of these results.

One possibility is that though presenting items for naming training in a homogeneous condition induces greater retrieval effort and naming error, this enhanced effort is unrelated to learning. That is, similar performance at test in the homogeneous and mixed conditions may have resulted from the fact that during the training, participants engaged in multiple trials of retrieval practice followed by correct-answer feedback, which strengthened items to a comparable degree in the two conditions. Likewise, all participants benefitted strongly from both the homogeneous and mixed training contexts (see Appendix B), suggesting retrieval practice with feedback confers potent benefits regardless of the semantic context at training.

A second possibility is that successful retrievals that are more effortful at training due to semantic blocking confer more durable learning, but that our effortful retrieval manipulation was suboptimal as regards the tradeoff between greater training error rate and greater benefit from enhanced training effort. As discussed in the memory literature, unsuccessful retrievals during retrieval practice confer weak learning compared to successful retrievals (Dunlosky & Rawson, 2012; Kornell et al., 2011; Middleton et al., 2015; Pashler et al., 2005; Wissman & Rawson, 2018). Thus, more effortful training conditions can surpass a point of 'desirable difficulty' if retrieval failures during training are too frequent, which can partially or completely eliminate the advantage to later performance from increasing the effort during training (Bjork, 1994; Pashler et al., 2003). In the current study, the additional retrieval effort required by enhanced semantic competition may have surpassed the point of desirable difficulty, leading to similar test performance for the homogeneous and mixed conditions. Future studies may revisit these issues by parametrically varying the effort required for retrieval via a manipulation of different degrees of semantic relatedness and examining more and longer retention intervals to increase experimental power for measuring forgetting in the different conditions. Another strategy could involve controlling for the number of correct retrievals during training between the homogeneous and mixed conditions by dropping items from further training when they reach a pre-assigned criterion of performance (e.g., Schuchard et al., 2020).

In the present study, findings from the semantic similarity analysis provided the strongest evidence of greater strengthening of items trained amidst enhanced semantic competition. First, we observed a cross-over interaction between item-to-set similarity and time. Specifically, increasing similarity of a homogeneous item to its set mates (item-to-set similarity) was associated with decreasing naming accuracy during training, reflective of enhanced retrieval difficulty. On the other hand, we observed that increasing item-to-set semantic similarity was associated with increasing naming accuracy at test. This indicates that greater retrieval effort due to greater interference from more highly related set mates at training conferred greater strengthening of items.

The results from the semantic similarity analysis are compatible with theories in the learning and memory literature that postulate a role for retrieval effort in the potency of learning from retrieval practice (e.g., Karpicke & Bauernschmidt, 2011; Karpicke & Roediger III, 2007; Pashler et al., 2003; Pyc & Rawson, 2009). In the case of lexical access, this study provides original evidence that increasing the effort required for retrieval of a target word by manipulating preceding semantic context affects a target word's retrievability at a future session. To more fully characterize the underlying learning mechanism, headway may be made by relating the present results to current theories of effortful retrieval effects. For example, according to the inhibitory account of retrieval induced forgetting, inhibition of related items when retrieving a target decreases the future accessibility of those competitors in a persistent fashion (Anderson et al., 2000; Storm et al., 2007). However, when a competitor becomes a target, its lower accessibility from prior inhibition potentiates the benefit it receives from a strengthening event (Storm et al., 2008). Though no models of lexical access yet exist that account for the present results, those that include mechanisms for retrieval-based weakening and strengthening (e.g., Oppenheim et al., 2010) may provide a better foundation for understanding the present results than those that only propose strengthening of targets following retrieval (e.g., Howard et al., 2006). Explicit, computational investigations are required to examine whether the fundamental assumptions of such models are ultimately compatible with the present results.

In addition to the training effects examined in the present study, we probed classic indices of semantic context effects in blocked-cyclic naming including a semantic blocking effect at training as well as cumulative semantic interference across cycles during training. The semantic blocking effect was apparent in the observation of decreased naming accuracy in the homogeneous condition compared to the mixed condition at training. However, we did not find evidence for cumulative semantic interference, which is not entirely unexpected. In an extensive review, Belke and Stielow (2013) found evidence of cumulative semantic interference only for participants with moderate to severe aphasia where neurological damage involved left frontal cortical sites, specifically left-inferior frontal gyrus. In the present study, PWA were not selected based on lesion profile; rather they were selected because of their cognitive-linguistic profile consistent with lexical access deficit as a contributor to their naming impairment and willingness to commit to the months-long protocol. In addition, the present study differed in important ways from the standard blocked-cyclic naming paradigm in that the items selected for training were largely errorful, and feedback was given on each trial. It is unclear which aspect of our design may have precluded observing cumulative semantic interference.

The present work bears on theories of lexical access by demonstrating an effortful retrieval effect in people with aphasia whose naming deficit is consistent with lexical access deficit. The effect of effortful retrieval in this current study is likely to localize, at least in part, to the first stage of lexical access in our participants. The majority of naming errors produced during item-selection testing were semantic substitutions, omissions, and descriptions (Appendix A), and such errors localize to neuroanatomical areas implicated in semantically-driven word retrieval (Chen et al., 2019; Schwartz et al., 2009, 2011). Though semantic naming errors in particular have also been attributed to dysregulated or degraded semantic representations (Gainotti et al., 1981; Hillis et al., 1990; Jefferies & Lambon Ralph, 2006), the participants in our sample had notably mild nonverbal semantic and word comprehension deficits (Section 2.1). Second, effortful retrieval in the present study was induced by a semantic context manipulation. Through careful experimentation, studies have localized semantic context effects in blocked-cyclic naming to the first (semantics-to-word) stage of lexical access (Damian et al., 2001; Kroll

& Stewart, 1994; Vigliocco et al., 2002). However, we consider the possibility that enhanced effort from semantic competition may have impacted phonological retrieval in our participants. A rationale could be that, because of cascading activation, greater semantic competition provokes enhanced activation of competitor phonemes, translating into greater learning when the correct phonemes are ultimately retrieved. Such a possibility could be evaluated in a future study examining semantic-competition induced effortful retrieval effects in individuals with aphasia with relatively pure stage-1 versus stage-2 lexical access deficits. This is one of the many potentially exciting future directions for research seeking to manipulate semantic competition to enhance the efficacy of treatments for aphasia.

Acknowledgements

We would like to thank Erica Johns for her contribution to data collection and data coding. We also thank Myrna Schwartz, Gary Dell, and members of the Moss Rehabilitation Research Institute (MRRI) Language and Learning Lab for their feedback on this project.

Funding

This work was supported by the National Institutes of Health [grant number R01 DC015516-01A1] and the Albert Einstein Society, Philadelphia, PA awarded to Erica L. Middleton.

Disclosure statements

We declare no potential conflicts of interest.

References

Alario, F.-X., & del Prado Martín, F. M. (2010). On the origin of the "cumulative semantic inhibition" effect. *Memory & Cognition*, *38*(1), 57–66. https://doi.org/10.3758/MC.38.1.57

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language*, *49*(4), 415–445. https://doi.org/10.1016/j.jml.2003.08.006

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *20*(5), 1063–1087. https://doi.org/10.1037//0278-7393.20.5.1063

Anderson, M. C., Bjork, E. L., & Bjork, R. A. (2000). Retrieval-induced forgetting: Evidence for a recall-specific mechanism. *Psychonomic Bulletin & Review*, *7*(3), 522–530. https://doi.org/10.3758/BF03214366

Bäuml, K.H. (2002). Semantic generation can cause episodic forgetting. *Psychological Science*, *13*(4), 356–360. https://doi.org/10.1111/1467-9280.00464

Belke, E. (2008). Effects of working memory load on lexical-semantic encoding in language production. *Psychonomic Bulletin & Review*, *15*(2), 357–363. https://doi.org/10.3758/PBR.15.2.357

Belke, E. (2013). Long-lasting inhibitory semantic context effects on object naming are necessarily conceptually mediated: Implications for models of lexical-semantic encoding. *Journal of Memory and Language*, *69*(3), 228–256. https://doi.org/10.1016/j.jml.2013.05.008

Belke, E., Meyer, A. S., & Damian, M. F. (2005). Refractory effects in picture naming as assessed in a semantic blocking paradigm. *The Quarterly Journal of Experimental Psychology Section A*, *58*(4), 667–692. https://doi.org/10.1080/02724980443000142

Belke, E., & Stielow, A. (2013). Cumulative and non-cumulative semantic interference in object naming: Evidence from blocked and continuous manipulations of semantic context. *Quarterly Journal of Experimental Psychology*, *66*(11), 2135–2160. https://doi.org/10.1080/17470218.2013.775318

Biegler, K. A., Crowther, J. E., & Martin, R. C. (2008). Consequences of an inhibition deficit for word production and comprehension: Evidence from the semantic blocking paradigm. *Cognitive Neuropsychology*, *25*(4), 493–527. https://doi.org/10.1080/02643290701862316

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe, & A. Shimamura (Eds.), *Metacognition: Knowing about knowing*. 185–205. Cambridge, MA: MIT Press.

Boersma, P., & Weenink, D. (2016). *Praat: Doing phonetics by computer [computer program].* Version 6.0.14, retrieved 10th January 2016 from http://www.praat.org/

Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The Bank of Standardized Stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PloS One*, *5*(5), e10773. https://10.1371/journal.pone.0010773

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology*, *14*(1), 177–208. https://doi.org/10.1080/026432997381664

Caramazza, A., & Miozzo, M. (1997). The relation between syntactic and phonological knowledge in lexical access: Evidence from the 'tip-of-the-tongue' phenomenon. *Cognition*, *64*(3), 309–343. https://doi.org/10.1016/S0010-0277(97)00031-0

Chen, Q., Middleton, E., & Mirman, D. (2019). Words fail: Lesion-symptom mapping of errors of omission in post-stroke aphasia. *Journal of Neuropsychology*, *13*(2), 183–197. https://doi.org/10.1111/jnp.12148

Conroy, P., Sage, K., & Ralph, M. A. L. (2009). Errorless and errorful therapy for verb and noun naming in aphasia. *Aphasiology*, *23*(11), 1311–1337. https://doi.org/10.1080/02687030902756439

Damian, M. F., & Als, L. C. (2005). Long-lasting semantic context effects in the spoken production of object names. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1372. https://doi.org/10.1037/0278-7393.31.6.1372

Damian, M. F., Vigliocco, G., & Levelt, W. J. (2001). Effects of semantic context in the naming of pictures and words. *Cognition*, *81*(3), B77–B86. https://doi.org/10.1016/S0010-0277(01)00135-4

Dell, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes*, *5*(4), 313–349. https://doi.org/10.1080/01690969008407066

Dell, G. S., & O'Seaghdha, P. G. (1992). Stages of lexical access in language production. *Cognition*, *42*(1–3), 287–314. https://doi.org/10.1016/0010-0277(92)90046-K

Dell, G. S., & Reich, P. A. (1981). Stages in sentence production: An analysis of speech error data. *Journal of Verbal Learning and Verbal Behavior*, *20*(6), 611–629. https://doi.org/10.1016/S0022-5371(81)90202-4

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, *104*(4), 801. https://doi.org/10.1037/0033-295X.104.4.801

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, *22*, 271-280. https://doi.org/10.1016/j.learninstruc.2011.08.003

Fay, D., & Cutler, A. (1977). Malapropisms and the Structure of the Mental Lexicon. *Linguistic Inquiry*, *8*(3), 505–520.

Fillingham, J. K., Sage, K., & Lambon Ralph, M. A. (2005a). Further explorations and an overview of errorless and errorful therapy for aphasic word-finding difficulties: The number of naming attempts during therapy affects outcome. *Aphasiology*, *19*(7), 597–614. https://doi.org/10.1080/02687030544000272

Fillingham, J. K., Sage, K., & Lambon Ralph, M. A. (2005b). Treatment of anomia using errorless versus errorful learning: Are frontal executive skills and feedback important? *International Journal of Language & Communication Disorders*, *40*(4), 505–523. https://doi.org/10.1080/13682820500138572

Fillingham, J. K., Sage, K., & Lambon Ralph, M. A. (2006). The treatment of anomia using errorless learning. *Neuropsychological Rehabilitation*, *16*(2), 129–154. https://doi.org/10.1080/09602010443000254

Friedman, R. B., Sullivan, K. L., Snider, S. F., Luta, G., & Jones, K. T. (2017). Leveraging the test effect to improve maintenance of the gains achieved through cognitive rehabilitation. *Neuropsychology*, *31*(2), 220–228. https://doi.org/10.1037/neu0000318

Fromkin, V. A. (1971). The Non-Anomalous Nature of Anomalous Utterances. *Language*, *47*(1), 27–52. https://doi.org/10.2307/412187

Gainotti, G., Miceli, G., Caltagirone, C., Silveri, M. C., & Masullo, C. (1981). The Relationship Between Type of Naming Error and Semantic-Lexical Discrimination in aphasic Patients. *Cortex*, *17*(3), 401–409. https://doi.org/10.1016/S0010-9452(81)80028-7

Garrett, M. F. (1976). Syntactic processes in sentence production. In R. J. Wales & E. Walker (Eds.), New approaches to language mechanisms. Amsterdam: North-Holland.

Garrett, M.F. (1975). The analysis of sentence production. In G. Bower (Ed.), *Psychology of Learning and Motivation*. Vol IX. Academic Press.

Goldrick, M., & Rapp, B. (2007). Lexical and post-lexical phonological representations in spoken production. *Cognition*, *102*(2), 219–260. https://doi.org/10.1016/j.cognition.2005.12.010

Harvey, D. Y., & Schnur, T. T. (2015). Distinct loci of lexical and semantic access deficits in aphasia: Evidence from voxel-based lesion-symptom mapping and diffusion tensor imaging. *Cortex*, *67*, 37–58. https://doi.org/10.1016/j.cortex.2015.03.004

Harvey, D. Y., Traut, H. J., & Middleton, E. L. (2019). Semantic interference in speech error production in a randomised continuous naming task: Evidence from aphasia. *Language,* *Cognition and Neuroscience*, *34*(1), 69–86. https://doi.org/10.1080/23273798.2018.1501500

Hebb, D. O. (1949). *The organization of behavior; a neuropsychological theory* (pp. xix, 335). Wiley.

Hillis, A. E., Rapp, B., Romani, C., & Caramazza, A. (1990). Selective impairment of semantics in lexical processing. *Cognitive Neuropsychology*, *7*(3), 191–243. https://doi.org/10.1080/02643299008253442

Howard, D., Nickels, L., Coltheart, M., & Cole-Virtue, J. (2006). Cumulative semantic inhibition in picture naming: Experimental and computational studies. *Cognition*, *100*(3), 464–482. https://doi.org/10.1016/j.cognition.2005.02.006

Howard, D., & Patterson, K. (1992). Pyramids and palm trees: A test of semantic access from pictures and words. Bury St Edmunds: Thames Valley Test.

Jefferies, E., & Lambon Ralph, M. A. (2006). Semantic impairment in stroke aphasia versus semantic dementia: A case-series comparison. *Brain*, *129*(8), 2132–2147. https://doi.org/10.1093/brain/awl153

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1250–1257. https://doi.org/10.1037/a0023436

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

Kertesz, A. (2007). Western aphasia battery-revised. San Antonio, TX: PsychCorp.

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*, 85-97. https://doi.org/10.1016/j.jml.2011.04.002

Kornell, N., & Vaughn, K. E. (2016). Chapter Five - How Retrieval Attempts Affect Learning: A Review and Synthesis. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 65, pp. 183–215). Academic Press. https://doi.org/10.1016/bs.plm.2016.03.003

Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connection between bilingual memory representations. *Journal of Memory and Language*, *33*(2), 149–174. https://doi.org/10.1006/jmla.1994.1008

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Lecours, A. R., & Lhermitte, F. (1969). Phonemic Paraphasias: Linguistic Structures and Tentative Hypotheses. *Cortex*, *5*(3), 193–228. https://doi.org/10.1016/S0010-9452(69)80031-6

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–38. https://doi.org/10.1017/S0140525X99001776

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766. https://doi.org/10.1016/j.jesp.2013.03.013

MacDonald, S. W. S., Nyberg, L., & Bäckman, L. (2006). Intra-individual variability in behavior: Links to brain structure, neurotransmission and neuronal activity. *Trends in Neurosciences*, *29*(8), 474–480. https://doi.org/10.1016/j.tins.2006.06.011

McCarthy, R. A., & Kartsounis, L. D. (2000). Wobbly words: Refractory anomia with preserved semantics. *Neurocase*, *6*(6), 487–497. https://doi.org/10.1080/13554790008402719

McKissock, S., & Ward, J. (2007). Do errors matter? Errorless and errorful learning in anomic picture naming. *Neuropsychological Rehabilitation*, *17*(3), 355–373. https://doi.org/10.1080/09602010600892113

Middleton, E. L., Rawson, K. A., & Verkuilen, J. (2019). Retrieval practice and spacing effects in multi-session treatment of naming impairment in aphasia. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *119*, 386–400. https://doi.org/10.1016/j.cortex.2019.07.003

Middleton, E. L., Schuchard, J., & Rawson, K. A. (2020). A Review of the Application of Distributed Practice Principles to Naming Treatment in Aphasia. *Topics in Language Disorders*, *40*(1), 36–53. https://doi.org/10.1097/TLD.0000000000000202

Middleton, E. L., Schwartz, M. F., Rawson, K. A., & Garvey, K. (2015). Test-enhanced learning versus errorless learning in aphasia rehabilitation: Testing competing psychological principles. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *41*(4), 1253–1261. https://doi.org/10.1037/xlm0000091

Middleton, E. L., Schwartz, M. F., Rawson, K. A., Traut, H., & Verkuilen, J. (2016). Towards a Theory of Learning for Naming Rehabilitation: Retrieval Practice and Spacing Effects. *Journal of Speech, Language, and Hearing Research : JSLHR*, *59*(5), 1111–1122. https://doi.org/10.1044/2016_JSLHR-L-15-0303

Murayama, K., Miyatsu, T., Buchli, D., & Storm, B. C. (2014). Forgetting as a consequence of retrieval: A meta-analytic review of retrieval-induced forgetting. *Psychological Bulletin*, *140*(5), 1383–1409. https://doi.org/10.1037/a0037505

Navarrete, E., Del Prato, P., & Mahon, B. Z. (2012). Factors Determining Semantic Facilitation and Interference in the Cyclic Naming Paradigm. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00038

Navarrete, E., Mahon, B. Z., & Caramazza, A. (2010). The cumulative semantic cost does not reflect lexical selection by competition. *Acta Psychologica*, *134*(3), 279–289. https://doi.org/10.1016/j.actpsy.2010.02.009

Oppenheim, G. M., Dell, G. S., & Schwartz, M. F. (2010). The dark side of incremental learning: A model of cumulative semantic interference during lexical access in speech production. *Cognition*, *114*(2), 227–252. https://doi.org/10.1016/j.cognition.2009.09.007

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 3-8. https://doi.org/10.1037/0278-7393.31.1.3

Pashler, H., Zarow, G., & Triplett, B. (2003). Is Temporal Spacing of Tests Helpful Even When It Inflates Error Rates? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1051–1057. https://doi.org/10.1037/0278-7393.29.6.1051

Patra, A., Bose, A., & Marinis, T. (2021). Semantic context effects in monolingual and bilingual speakers. *Journal of Neurolinguistics*, *57*, 100942. https://doi.org/10.1016/j.jneuroling.2020.100942

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*(4), 437–447. https://doi.org/10.1016/j.jml.2009.01.004

Raaijmakers, J. G. W., & Jakab, E. (2013). Is Forgetting Caused by Inhibition? *Current Directions in Psychological Science*, *22*(3), 205–209. https://doi.org/10.1177/0963721412473472

Rapp, B., & Goldrick, M. (2000). Discreteness and interactivity in spoken word production. *Psychological Review*, *107*(3), 460–499. https://doi.org/10.1037/0033-295X.107.3.460

Roach, A., Schwartz, M. F., Martin, N., Grewal, R. S., & Brecher, A. (1996). The Philadelphia Naming Test: Scoring and rationale. *Clinical Aphasiology*, *24*, 121–133.

Roediger III, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Rose, S. B., & Abdel Rahman, R. (2017). Semantic similarity promotes interference in the continuous naming paradigm: Behavioural and electrophysiological evidence. *Language, Cognition and Neuroscience*, *32*(1), 55–68. https://doi.org/10.1080/23273798.2016.1212081

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

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*, *110*, 104038. https://doi.org/10.1016/j.jml.2019.104038

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*(4), 207–218. https://doi.org/10.1111/j.1467-9280.1992.tb00029.x

Schnur, T. T., Schwartz, M. F., Brecher, A., & Hodgson, C. (2006). Semantic interference during blocked-cyclic naming: Evidence from aphasia. *Journal of Memory and Language*, *54*(2), 199–227. https://doi.org/10.1016/j.jml.2005.10.002

Schuchard, J., & Middleton, E. L. (2018a). The Roles of Retrieval Practice Versus Errorless Learning in Strengthening Lexical Access in Aphasia. *Journal of Speech, Language, and Hearing Research : JSLHR*, *61*(7), 1700–1717. https://doi.org/10.1044/2018_JSLHR-L-17-0352

Schuchard, J., & Middleton, E. L. (2018b). Word repetition and retrieval practice effects in aphasia: Evidence for use-dependent learning in lexical access. *Cognitive Neuropsychology*, *35*(5–6), 271–287. https://doi.org/10.1080/02643294.2018.1461615

Schuchard, J., Rawson, K. A., & Middleton, E. L. (2020). Effects of distributed practice and criterion level on word retrieval in aphasia. *Cognition*, *198*, 104216. https://doi.org/10.1016/j.cognition.2020.104216

Schwartz, M. F. (2014). Theoretical analysis of word production deficits in adult aphasia. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634). https://doi.org/10.1098/rstb.2012.0390

Schwartz, M. F., Dell, G. S., Martin, N., Gahl, S., & Sobel, P. (2006). A case-series test of the interactive two-step model of lexical access: Evidence from picture naming. *Journal of Memory and Language*, *54*(2), 228–264. https://doi.org/10.1016/j.jml.2005.10.001

Schwartz, M. F., Kimberg, D. Y., Walker, G. M., Brecher, A., Faseyitan, O. K., Dell, G. S., Mirman, D., & Coslett, H. B. (2011). Neuroanatomical dissociation for taxonomic and thematic knowledge in the human brain. *Proceedings of the National Academy of Sciences*, *108*(20), 8520–8524. https://doi.org/10.1073/pnas.1014935108

Schwartz, M. F., Kimberg, D. Y., Walker, G. M., Faseyitan, O., Brecher, A., Dell, G. S., & Coslett, H. B. (2009). Anterior temporal involvement in semantic word retrieval: Voxel-based lesion-symptom mapping evidence from aphasia. *Brain: A Journal of Neurology*, *132*(Pt 12), 3411–3427. https://doi.org/10.1093/brain/awp284

Scott, R. M., & Wilshire, C. E. (2010). Lexical competition for production in a case of nonfluent aphasia: Converging evidence from four different tasks. *Cognitive Neuropsychology*, *27*(6), 505–538. https://doi.org/10.1080/02643294.2011.598853

Stemberger, J. P. (1985). An interactive activation model of language production. In A. W. Ellis (Ed.), *Progress in the Psychology of Language* (Vol. 1, pp. 143-186). NJ: Erlbaum.

Storm, B. C., Bjork, E. L., & Bjork, R. A. (2005). Social metacognitive judgments: The role of retrieval-induced forgetting in person memory and impressions. *Journal of Memory and Language*, *52*(4), 535–550. https://doi.org/10.1016/j.jml.2005.01.008

Storm, B. C., Bjork, E. L., & Bjork, R. A. (2007). When intended remembering leads to unintended forgetting. *The Quarterly Journal of Experimental Psychology*, *60*(7), 909–915. https://doi.org/10.1080/17470210701288706

Storm, B. C., Bjork, E. L., & Bjork, R. A. (2008). Accelerated relearning after retrieval-induced forgetting: The benefit of being forgotten. *Journal Of Experimental Psychology-Learning Memory And Cognition*, *34*(1), 230–236. https://doi.org/10.1037/0278-7393.34.1.230

Storm, B. C., & Levy, B. J. (2012). A progress report on the inhibitory account of retrieval-induced forgetting. *Memory & Cognition*, *40*(6), 827–843. https://doi.org/10.3758/s13421-012-0211-7

Szekely, A., Jacobsen, T., D'Amico, S., Devescovi, A., Andonova, E., Herron, D., Lu, C. C., Pechmann, T., Pléh, C., Wicha, N., Federmeier, K., Gerdjikova, I., Gutierrez, G., Hung, D., Hsu, J., Iyer, G., Kohnert, K., Mehotcheva, T., Orozco-Figueroa, A., … Bates, E. (2004). A new on-line resource for psycholinguistic studies. *Journal of Memory and Language*, *51*(2), 247–250.

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, *50*(3), 289–335. https://doi.org/10.1016/j.jml.2003.10.003

Verde, M. F. (2012). Chapter Two - Retrieval-Induced Forgetting and Inhibition: A Critical Review. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 56, pp. 47–80). Academic Press. https://doi.org/10.1016/B978-0-12-394393-4.00002-9

Vigliocco, G., Vinson, D. P., Damian, M. F., & Levelt, W. (2002). Semantic distance effects on object and action naming. *Cognition*, *85*(3), B61-69. https://doi.org/10.1016/s0010-0277(02)00107-5

Wiley, R. W., & Rapp, B. (2019). Statistical analysis in Small-N Designs: Using linear mixed-effects modeling for evaluating intervention effectiveness. *Aphasiology*, *33*(1), 1–30. https://doi.org/10.1080/02687038.2018.1454884

Wilshire, C. E., & McCarthy, R. A. (2002). Evidence for a context-sensitive word retrieval disorder in a case of nonfluent aphasia. *Cognitive Neuropsychology*, *19*(2), 165–186. https://doi.org/10.1080/02643290143000169

Wissman, K. T., & Rawson, K. A. (2018). Test potentiated learning: Three independent replications, a disconfirmed hypothesis, and an unexpected boundary condition. *Memory*, *26*(4), 385-405. https://doi.org/ 10.1080/09658211.2017.1350717

Table 1. Participant demographic and neuropsychological characteristics

Participant	Age	Years of Education	Gender	MPO	WAB AQ	Aphasia Subtype	Speech Apraxia	PNT Acc	Nonverbal Comp	Word Comp	Word Rep
1	47	13	F	82	81.5	Anomic	Mild	78	83	94	97
2	49	13	F	24	81.1	Anomic	Mild	74	92	96	93
3	29	15	M	3	73.2	Conduction	None	59	94	100	97
4	55	16	M	10	76.7	Anomic	None	64	96	92	87
5	67	19	M	7	58.8	Broca's	Moderate	57	94	96	89
6	63	12	F	20	92.9	Anomic	None	78	88	98	94
Mean	51.7	14.7		24.3	77.4			68.3	91.2	96	92.8
Min, Max	29,67	12,19		3,82	58.8,92.9			57,78	83,96	92,100	89,97
Controls[a]								97		99	100
Cutoff[b]					93.8				90		

Notes. MPO = months post-stroke onset; WAB AQ = Western Aphasia Battery Aphasia Quotient, a measure of aphasia severity (Kertesz, 2007); PNT = Philadelphia Naming Test (Roach et al., 1996) performance, where ACC = accuracy in percentages; Nonverbal Comp = an associative picture-picture matching task of nonverbal comprehension, in percentages (Howard & Patterson, 1992); Word Comp = a spoken word-picture verification task of word comprehension, in percentages (Roach et al., 1996); Word Rep = a test of immediate word repetition, in percentages (Philadelphia Repetition Test; Dell et al., 1997).

[a] Average performance for neurotypical control sample
[b] Scores below cutoff indicate clinically significant impairment

Table 2. List of 19 related categories with sample category members

Category	Examples	Category	Examples
1. Accessories	Belt, bracelet, scarf	11. Non-Mammals	Alligator, fly, shark
2. Body Parts	Arm, ear, tongue	12. Office Supplies	Calculator, envelope, pen
3. Clothing	Dress, jeans, sweater	13. Parts of Buildings	Chimney, fireplace, roof
4. Food	Bacon, cheese, steak	14. Structures	Airport, bridge, hut
5. Fruits & Vegetables	Apple, broccoli, orange	15. Toiletries	Comb, razor, towel
6. Furnishings	Bed, chair, dresser	16. Tools & Hardware	Ax, ladder, rake
7. Kitchen Items	Blender, fork, plate	17. Toys & Games	Ball, crayon, baseball
8. Mammals	Bear, goat, rabbit	18. Types of People	Baby, cowboy, nurse
9. Musical Instruments	Accordion, banjo, drum	19. Vehicles	Bus, rocket, train
10. Nature	Acorn, cactus, river		

Table 3. Mean (SD) per variable across participants' personalized item sets as a function of condition

Training	Naming Accuracy[1]	#Syllables	#Phonemes	Log frequency	Name agreement	Visual Complexity
	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)
Homogeneous	.34 (.34)	2.14 (.86)	6 (2.02)	.74 (.53)	.92 (.06)	2.67 (.81)
Mixed	.35 (.34)	2.18 (.91)	6.14 (2.07)	.73 (.54)	.91 (.06)	2.71 (.78)

Notes. [1]Mean naming accuracy from the item selection phase

Table 4. Example of LSA-based item-to-set semantic similarity estimates for a hypothetical homogeneous set

Set	Attic	Chimney	Mailbox	Ceiling	Window	Tile	Item-to-set semantic similarity
Attic	-----	.35	.10	.53	.56	.25	.36
Chimney	.35	----	.05	.47	.36	.31	.31
Mailbox	.01	.05	----	.07	.20	.06	.10
Ceiling	.53	.47	.07	----	.50	.59	.43
Window	.56	.36	.20	.50	----	.28	.38
Tile	.25	.31	.06	.59	.28	----	.30

Note. LSA = Latent Semantic Analysis (Landauer et al., 1998)

Table 5. Mixed logistic regression model results on naming accuracy: Time by Condition interaction

Interaction Model: Time by Condition				
Fixed Effect	Coef.	SE	*Z*	*p*
Intercept	-0.25	0.69	-0.36	.72
Training[a]	-0.03	0.11	-0.30	.76
Mixed[b]	0.01	0.05	0.30	.77
Interaction of Time and Condition				
Training[a] x Mixed[b]	0.10	0.05	2.12	.03*
Log Frequency	0.38	0.09	4.08	<.001***
Syllable Length	-0.17	0.05	-3.42	<.001***
Name Agreement	2.34	0.72	3.25	<.001***
Random Effect	s^2			
Participant	0.51			
Item	0.12			
Participant: Time[a]	0.05			
Simple-Effects Model: Effect of Condition at training				
Fixed Effect	Coef.	SE	*Z*	*p*
Intercept	-0.15	0.68	-0.22	.82
Effect of Condition				
Mixed[b]	0.11	0.04	2.82	.005**
Log Frequency	0.33	0.09	3.47	<.001***
Syllable Length	-0.17	0.05	-3.39	<.001***
Name Agreement	2.21	0.72	3.03	.002**
Random Effect	s^2			
Participant	0.10			
Item	0.06			
Simple-Effects Model: Effect of Condition at test				
Fixed Effect	Coef.	SE	*Z*	*p*
Intercept	1.26	0.34	3.71	<.001***
Effect of Condition				
Mixed[b]	-0.09	0.09	-1.03	.30
Log Frequency	0.84	0.21	4.10	<.001***
Random Effect	s^2			
Participant	0.52			
Item	0.28			

Note. Sum coding was used for Time (Training = +1, Test = -1) and Condition (Mixed = +1, Homogenous = -1). Excluding the intercepts, Coef. = model estimation of the change in naming accuracy (in log odds) from the reference category for each fixed effect; SE = standard error of the estimate; *Z* = Wald Z test statistic, two-tailed; s^2 = Variance for by-participant random intercepts, by-items random intercepts, and by-participants random slopes.

[a] Reference is Test timepoint.

[b] Reference is Homogeneous condition.

Table 6. Mixed logistic regression model results on naming accuracy: Time by Item-to-Set Semantic Similarity interaction for homogeneous items only

Interaction Model: Time by Item-to-Set Semantic Similarity

Fixed Effect	Coef.	SE	*Z*	*p*
Intercept	-0.32	0.94	-0.34	.73
Training[a]	0.18	0.11	1.58	.11
Item-to-Set Semantic Similarity	0.72	0.81	0.89	.38
Interaction of Time and Item-to-Set Semantic Similarity				
Training[a] x Item-to-Set Semantic Similarity	-2.10	0.76	-2.76	.005**
Log Frequency	0.41	0.13	3.04	.002**
Syllable Length	-0.20	0.07	-2.75	.006**
Name Agreement	2.35	1.00	2.35	.02*
Random Effect	s^2			
Participant	0.14			
Item	0.21			

Simple-Effects Model: Effect of Item-to-Set Semantic Similarity at training

Fixed Effect	Coef.	SE	*Z*	*p*
Intercept	0.13	0.95	0.14	.88
Item-to-Set Semantic Similarity	-1.28	0.57	-2.21	.02*
Log Frequency	0.40	0.13	2.96	.003**
Syllable Length	-0.20	0.07	-2.71	.006**
Name Agreement	2.00	1.02	1.96	.04*
Random Effect	s^2			
Participant	0.11			
Item	0.13			

Simple-Effects Model: Effect of Item-to-Set Semantic Similarity at test

Fixed Effect	Coef.	SE	*Z*	*p*
Intercept	1.58	0.39	4.06	<.001***
Item-to-Set Semantic Similarity	3.11	1.53	2.03	.04*
Random Effect	s^2			
Participant	0.46			
Item	0.42			

Note. Sum coding was used for Time (Training = +1, Test = -1). Excluding the intercepts, Coef. = model estimation of the change in naming accuracy (in log odds) from the reference category for each fixed effect; SE = standard error of the estimate; *Z* = Wald Z test statistic, two-tailed; s^2 = Variance for by-participant random intercepts and by-items random intercepts.

[a] Reference is Test timepoint.

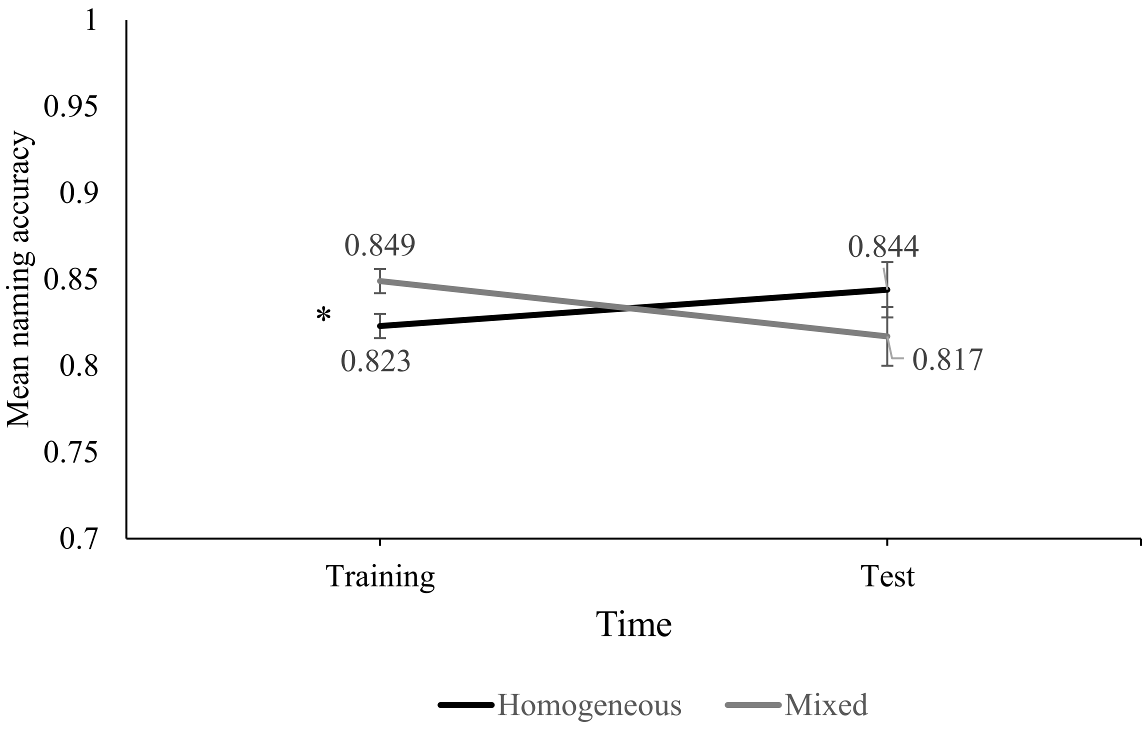


Figure 1. Mean naming accuracy for the two conditions (homogeneous and mixed) across the two time-points (training and test). Error bar represents the standard error of the mean.

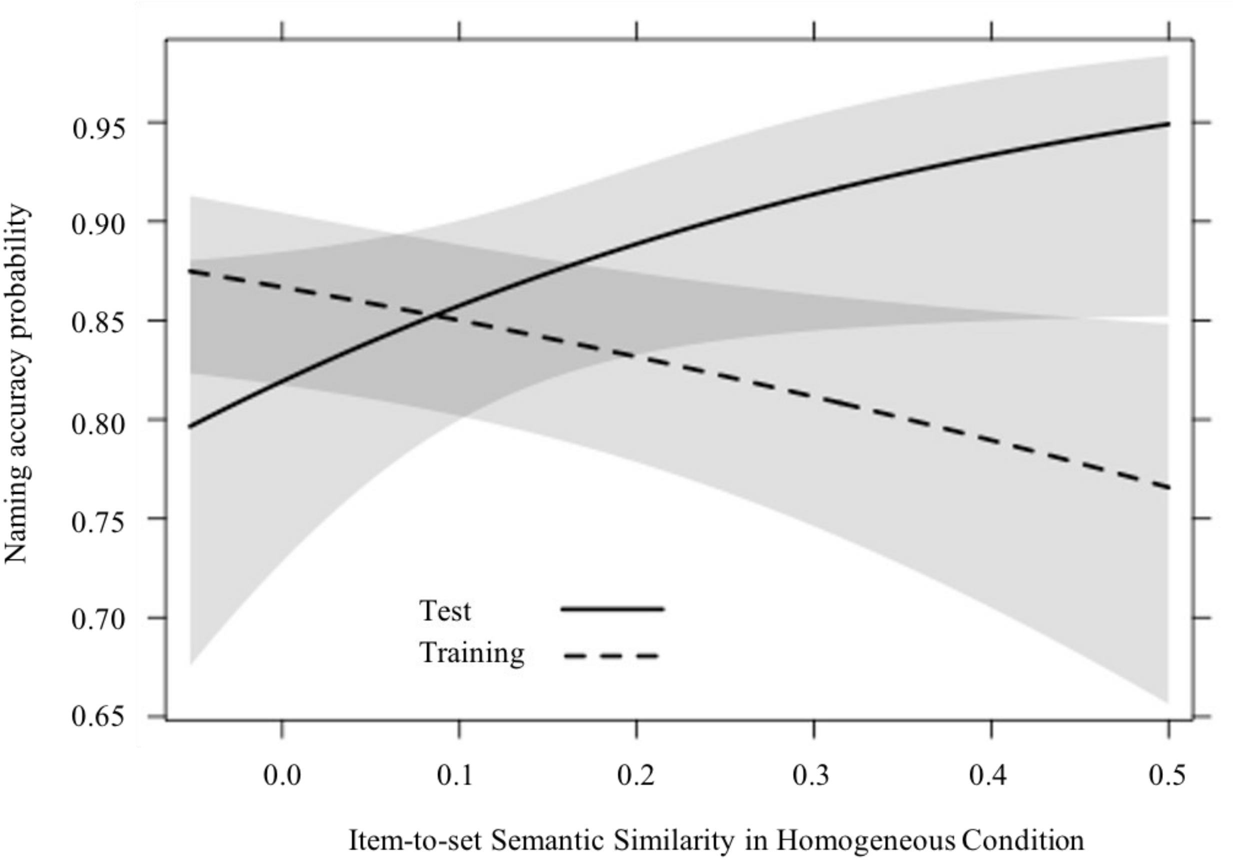


Figure 2. A predictor effect plot visualizing the interaction between time (training and test) and item-to-set semantic similarity in the homogeneous condition. The shaded regions represent pointwise confidence bands for the fitted values.

Appendix A

Breakdown of correct and incorrect responses on the item selection task

	Correct		Incorrect			
	Fully Correct	Minor Deviations	Phonol error	Sem err	NR/D	Other
Phonological overlap	1	.75-.99	0-.74	na	na	na
Participant						
1	.60	.07	.02	.04	.27	.01
2	.68	.07	.04	.08	.11	.02
3	.62	.06	.11	.07	.13	.01
4	.53	.05	.06	.19	.13	.03
5	.52	.11	.07	.07	.22	.02
6	.65	.07	.07	.12	.08	.01
Average	.60	.07	.06	.09	.16	.02

Note. Fully correct, phonological overlap score = 1.0; minor deviations, overlap score between .75-.99; phonol error = phonologically related word or nonword response, with overlap score between 0-.74; sem err: semantically related response; NR/D = no response or description; other = unrelated response, named picture part; na = not applicable.

Appendix B

Model results for effect of Phase (item-selection versus delayed test) on naming accuracy across the group (mixed logistic regression) and per participant (simple logistic regression) in the homogeneous and mixed conditions

Group analysis: Effect of Phase for homogenous items				
Fixed Effect	Coef.	SE	*Z*	*p*
Intercept	-5.24	1.25	-4.20	<.001***
Test[a]	1.50	0.09	16.30	<.001***
Log Frequency	0.31	0.15	2.12	.03*
Name Agreement	6.02	1.34	4.49	<.001***
Random Effect	s^2			
Participant	0.44			
Item	0.42			
Participant-level analyses: Effect of Phase for homogeneous items				
Participant	Coef.	SE	*Z*	*p*
1	1.63	0.23	7.17	<.001***
2	1.11	0.19	5.75	<.001***
3	1.64	0.21	7.64	<.001***
4	1.05	0.17	6.29	<.001***
5	1.50	0.17	8.77	<.001***
6	1.18	0.20	5.74	<.001***
Group analysis: Effect of Phase for mixed items				
Fixed Effect	Coef.	SE	*Z*	*p*
Intercept	-5.24	1.40	-3.73	<.001***
Test[a]	1.55	0.09	16.58	<.001***
Log Frequency	0.41	0.16	2.51	.01*
Name Agreement	5.77	1.52	3.79	<.001***
Random Effect	s^2			
Participant	0.44			
Item	0.68			
Participant-level analyses: Effect of Phase for mixed items				
Participant	Coef.	SE	*Z*	*p*
1	1.26	0.16	7.63	<.001***
2	1.50	0.23	6.64	<.001***
3	1.70	0.23	7.45	<.001***
4	1.05	0.17	6.16	<.001***
5	1.45	0.17	8.59	<.001***
6	1.22	0.21	5.90	<.001***

Note. Sum coding was used for Phase (Test= +1, Item-selection = -1). Excluding the intercepts, Coef. = model estimation of the change in naming accuracy (in log odds) from the reference category for each fixed effect; SE = standard error of the estimate; *Z* = Wald Z test statistic, two-tailed; s^2 = Variance for by-participant random intercepts and by-items random intercepts.

[a]Reference is Item-selection.

Appendix C.

Mixed logistic regression model results on the rate of change in naming accuracy across Time (i.e., going from training to test) for each Condition

Effect of Time for Homogenous condition				
Fixed Effect	Coef.	SE	*Z*	*p*
Intercept	-0.06	0.93	-0.06	.95
Effect of Time				
Training[a]	-0.09	0.07	-1.26	.20
Log Frequency	0.37	0.13	2.88	.003**
Syllable Length	-0.22	0.07	-2.93	.003**
Name Agreement	2.20	0.99	2.21	.03*
Random Effect	s^2			
Participant	0.15			
Item	0.21			

Effect of Time for Mixed condition				
Fixed Effect	Coef.	SE	*Z*	*p*
Intercept	-0.46	0.92	-0.49	.62
Effect of Time				
Training[a]	0.12	0.06	1.85	.06
Log Frequency	0.37	0.12	2.89	.004**
Syllable Length	-0.14	0.06	-2.16	.03*
Name Agreement	2.44	0.99	2.45	.01*
Random Effect	s^2			
Participant	0.14			
Item	0.08			

Note. Sum coding was used for Time (Training = +1, Test = -1). Excluding the intercepts, Coef. = model estimation of the change in naming accuracy (in log odds) from the reference category for each fixed effect; SE = standard error of the estimate; *Z* = Wald Z test statistic, two-tailed; s^2 = Variance for by-participant random intercepts and by-items random intercepts.

[a] Reference is Test timepoint.

Appendix D

Mixed logistic regression model results on naming accuracy: Cycle by Condition interaction and effect of Condition for the first cycle

Cycle by Condition interaction (all cycles included)

Fixed Effect	Coef.	SE	*Z*	*p*
Intercept	-0.05	0.83	-0.07	.94
Cycle	2.91	0.13	22.29	<.001***
Mixed[a]	0.29	0.06	4.77	<.001***
Interaction of Cycle and Condition				
Cycle x Mixed[a]	0.49	0.13	3.93	<.001***
Log Frequency	0.40	0.11	3.54	<.001***
Syllable Length	-0.22	0.06	-3.48	<.001***
Name Agreement	2.78	0.88	3.14	<.001***
Random Effect	s^2			
Participant	0.15			
Item	0.22			

Cycle by Condition interaction (first cycle excluded)

Fixed Effect	Coef.	SE	*Z*	*p*
Intercept	3.41	0.39	8.64	<.001***
Cycle	0.29	0.12	2.28	.02*
Mixed[a]	0.32	0.07	4.76	<.001***
Interaction of Cycle and Condition				
Cycle x Mixed[a]	-0.04	0.12	-0.30	.76
Log Frequency	0.55	0.18	3.13	.002**
Syllable Length	-0.35	0.09	-3.72	<.001***
Random Effect	s^2			
Participant	0.37			
Item	0.59			

Effect of Condition for the first cycle only

Fixed Effect	Coef.	SE	*Z*	*p*
Intercept	-8.35	1.30	-6.41	<.001***
Effect of Condition				
Mixed[a]	-0.02	0.07	-0.35	.72
Log Frequency	0.50	0.15	3.35	<.001***
Name Agreement	8.61	1.41	6.12	<.001***
Random Effect	s^2			
Participant	0.28			
Item	0.32			

Note. Linear contrasts coding was used for Cycle and Sum coding was used for Condition (Mixed = +1, Homogenous = -1). Excluding the intercepts, Coef. = model estimation of the change in naming accuracy (in log odds) from the reference category for each fixed effect; SE = standard error of the estimate; *Z* = Wald Z test statistic, two-tailed; s^2 = Variance for by-participant random intercepts and by-items random intercepts.

[a] Reference is Homogenous condition.

Appendix E1

Mixed linear regression model results on naming latencies: Time by Condition interaction

Interaction Model: Time by Condition				
Fixed Effect	Coef.	SE	*t*	*p*
Intercept	7.23	0.09	78.48	<.001***
Training[a]	-0.04	0.01	-4.77	<.001***
Mixed[b]	-0.01	0.01	-1.92	.054
Interaction of Time and Condition				
Training[a] x Mixed[b]	0.01	0.01	1.11	.27
Log Frequency	-0.06	0.01	-3.95	<.001***
Random Effect	s^2			
Participant	0.04			
Item	0.01			

Note. Sum coding was used for Time (Training = +1, Test = -1) and Condition (Mixed = +1, Homogenous = -1). Excluding the intercepts, Coef. = model estimation of the change in latency for each fixed effect; SE = standard error of the estimate; *t* = Satterthwaite's method, two-tailed; s^2 = Variance for by-participant random intercepts and by-items random intercepts.

[a] Reference is Test timepoint.

[b] Reference is Homogeneous condition.

Appendix E2

Mixed linear regression model results on naming latencies: Time by Item-to-Set Semantic Similarity interaction for homogenous items only

Interaction Model: Time by Item-to-Set Semantic Similarity				
Fixed Effect	Coef.	SE	*t*	*p*
Intercept	7.22	0.09	74.39	<.001***
Training[a]	-0.04	0.02	-2.22	.03*
Item-to-Set Semantic Similarity	0.19	0.13	1.43	.15
Interaction of Time and Item-to-Set Semantic Similarity				
Training[a] x Item-to-Set Semantic Similarity	-0.02	0.11	-0.25	.80
Log Frequency	-0.06	0.02	-2.68	.008**
Syllable length	0.04	0.01	2.44	.01*
Random Effect	s^2			
Participant	0.04			
Item	0.02			

Note. Sum coding was used for Time (Training = +1, Test = -1). Excluding the intercepts, Coef. = model estimation of the change in latency for each fixed effect; SE = standard error of the estimate; *t* = Satterthwaite's method, two-tailed; s^2 = Variance for by-participant random intercepts and by-items random intercepts.

[a] Reference is Test timepoint.

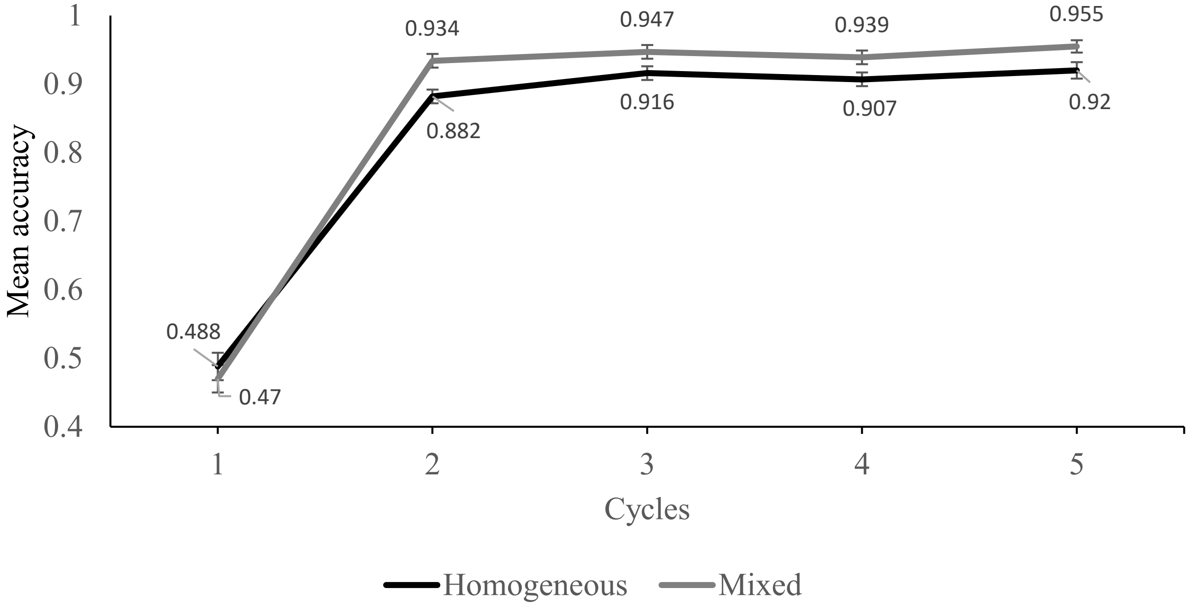
Appendix E3

Mixed linear regression model results on naming latencies: *Cycle by Condition interaction*

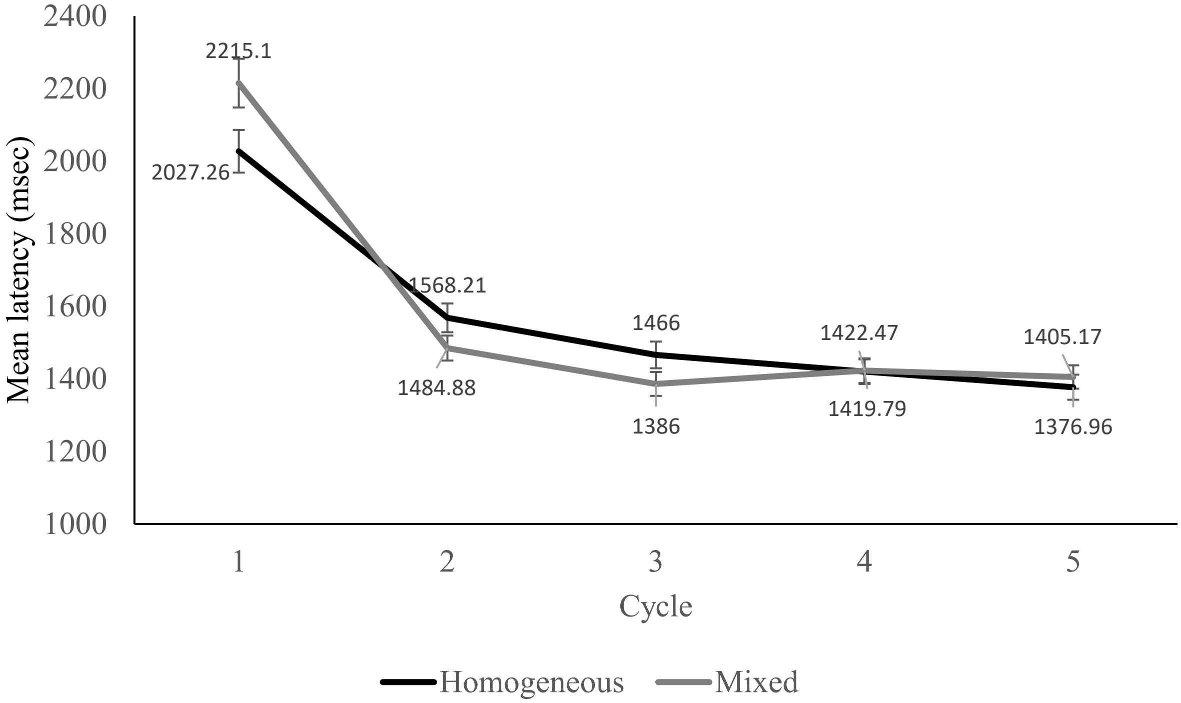
Cycle by Condition interaction (all cycles included)				
Fixed Effect	Coef.	SE	*t*	*p*
Intercept	7.21	0.09	72.42	<.001***
Cycle	-0.24	0.01	-16.98	<.001***
Mixed[a]	-0.01	0.01	-1.00	.32
Interaction of Cycle and Condition				
Cycle x Mixed[a]	0.01	0.01	0.15	.88
Log Frequency	-0.06	0.02	-3.53	<.001***
Syllable Length	0.03	0.01	3.10	.002**
Random Effect	s^2			
Participant	0.05			
Item	0.01			

Note. Linear contrasts coding was used for Cycle and Sum coding was used for Condition (Mixed = +1, Homogenous = -1). Excluding the intercepts, Coef. = model estimation of the change in latency for each fixed effect; SE = standard error of the estimate; *t* = Satterthwaite's method, two-tailed; s^2 = Variance for by-participant random intercepts and by-items random intercepts.

[a]Reference is Homogenous condition.



Appendix F. Mean accuracy for the two conditions (homogeneous and mixed) across five cycles during training. Error bar represents standard error of the mean across participants.



Appendix G. Mean latency in milliseconds for the two conditions (homogeneous and mixed) across five cycles during training. Error bar represents standard error of the mean across participants.