The effect of semantic similarity on re-learning of vocabulary in activation and inhibition deficit in aphasia

Nazbanou Nozari [1], Michael C. Freund [3], Brenda Rapp [4]

[1]Department of Psychological and Brain Sciences, Indiana University

[2] Department of Cognitive & Psychological Sciences, Brown University

[3] Department of Cognitive Science, Johns Hopkins University

Corresponding author: Nazbanou Nozari

Address: 1101 E 10th St., Bloomington, IN 47405

Phone number: 812-855-5818

Email: bnozari@iu.edu

Lab website: https://www.nozarilab.com/

Abstract

Semantic similarity between known items (e.g., cat/dog) can interfere with a speaker's ability to name them. Recent studies have shown that this interference extends to the acquisition of new vocabulary in neurotypical adults. However, the influence of such similarity on long-term retention of relearned vocabulary in individuals with aphasia remains unknown. This study examines this issue in two individuals with lexical retrieval deficits, who nevertheless show distinct profiles: one has difficulty activating lexical items (*activation deficit*), the other inhibiting lexical competitors (*inhibition deficit*). Participants were trained for six weeks using blocks of semantically related, unrelated, and mixed items. Learning was assessed during training, immediately after training, one week after each training session during the training period, and finally, two, six, and sixteen weeks after the last training session. We found a detrimental effect of semantic similarity on long-term retention in the inhibition, but not the activation, deficit. This dissociation shows that semantic similarity interacts with processes of activation and selection in a manner predicted by theories of incremental learning in language production. With regard to clinical implications, we found that, similar to neurotypical learners, semantic similarity among items in a training set can hurt the long-term retention, but that this effect can be ameliorated by spreading the related items across different training blocks within a session.

Keywords

Semantic similarity, aphasia, vocabulary learning, training, inhibition, activation

INTRODUCTION

Understanding the cognitive bases of word production and learning is not only a central part of understanding humans' language processing system, but also important for optimizing vocabulary acquisition (in first or second language) or re-acquisition of the lost vocabulary in individuals with aphasia (IWA), for whom difficulty in word retrieval typically remains a debilitating problem (Goodglass & Wingfield, 1997; Wisenburn & Mahoney, 2009). In this vein, cognitive research has contributed significantly to advancing aphasia treatment (Kiran & Bassetto, 2008; Nickels et al., 2010). These studies include, but are not limited to, research investigating the dose and duration of treatment (Basso, 2005; Bhogal et al., 2003; Hinckley & Carr, 2005; Off et al., 2016), massed vs. spaced retrieval treatments (Benigas & Bourgeois, 2016; Fridriksson et al., 2005; Mahendra, 2008; Morrow & Fridriksson, 2006; Rapp & Wiley, 2019; Rogalski et al., 2014) errorless vs. errorful retrieval practice (Clare & Jones, 2008; Fillingham et al., 2005; McKissock & Ward, 2007; Middleton & Schwartz, 2012), and different cueing methods (Cameron et al., 2006; Conroy et al., 2009; Harnish et al., 2014; Loverso et al., 1988; Lowell et al., 1995; Macoir et al., 2015; Marshall et al., 1992; Neumann, 2018; Wambaugh et al., 2002).

Following the cognitive tradition of such studies, the present study focuses on the role of semantic similarity among items within a training set in the re-acquisition of lost vocabulary in IWA. This investigation is motivated by research showing detrimental effects of taxonomic similarity on word production in a variety of paradigms (e.g., Howard et al., 2006; Nozari et al., 2016; Schnur et al., 2006; Schriefers et al., 1990; see Nozari & Pinet, 2020, for a review) and more recently, the extension of such detrimental effects to the acquisition of new vocabulary in neurotypical individuals (Korochkina et al., 2021; Waller et al., 2025).

Here, we investigate whether the same detrimental effect of semantic similarity is found in IWA relearning vocabulary from their first language. Specifically, we study two IWA, XR and QD, reported in Nozari (2019), whose deficits are localized to the stage of mapping semantic features to lexical items (i.e., lexical retrieval). While matched in many respects, XR's difficulty in lexical retrieval is primarily an inability to activate lexical items (*activation deficit*), whereas QD's difficulty is in inhibiting competing lexical items (*inhibition deficit*). We can thus compare the influence of semantic similarity on vocabulary relearning in these two deficits.

In addition to this novel comparison of deficit types, the study investigates two new factors: first, using a longitudinal design, we measure learning and retention at various intervals during training, right after training, one week after each training session, and two, six, and sixteen weeks after the last training session, examining a comprehensive timeline of semantic effects on learning from immediate to long-term. Second, we include two controls for training in blocks of semantically related items: completely unrelated blocks (in which the items do not belong to any of the categories presented in the related blocks) and mixed blocks (in which items within a block do not belong to the same category, but they each belong to one of the categories trained in related blocks). The inclusion of mixed blocks allows us to test whether it is local (i.e., within-block) or global (i.e., across all blocks) similarity that affects learning. Additionally, we implement several techniques known to benefit learning, including repetition (e.g., Fillingham et al., 2005; Nickels, 2002), retrieval practice (Karpicke & Aue, 2015; Karpicke & Roediger III, 2007; Zaromb & Roediger, 2010), and semantic feature analysis (Boyle & Coelho, 1995; Efstratiadou et al., 2018). Inclusion of the last technique, which explicitly activates the semantic features of the target item during training, is important for disentangling the influence of increased semantic activation on the *target* itself from its influence on *non-target* items, which may indirectly affect target selection.

Semantic similarity and its influence on production and learning

Similarity effects have been instrumental in uncovering the nature of representations and processes underlying language production (e.g., Breining et al., 2016; Howard et al., 2006; Oppenheim et al., 2010; Schriefers et al., 1990). Interestingly, semantic effects can differ based on the type of relation between target and context. Taxonomic relations often create interference in language production (e.g., Belke et al., 2005; Nozari et al., 2016; Schnur et al., 2006; 2009), whereas non-taxonomic relations (e.g., thematic relations, whole-part relations, and associations) do not cause interference and may even facilitate production (Alario et al., 2000; Costa et al., 2005; Mahon et al., 2007; Oppenheim & Nozari, 2024). Our focus in the current study is on taxonomic relations. Henceforth, we use the term "semantic" to refer to such relations.

A large body of literature has shown that pictures are harder to name if they appear in the context of semantically similar pictures. This effect has been shown using a variety of paradigms and populations (e.g., Belke et al., 2005; Nozari et al., 2016; Schnur, 2014; Schnur et al., 2006, 2009). Intriguingly, this so-called "semantic interference" effect is long-lasting, surviving passage of time and multiple intervening items between two semantically related pictures (Hepner & Nozari, 2020; Schnur, 2014). This resilience has given rise to learning accounts of semantic interference (Oppenheim, 2018; Oppenheim et al., 2010; Oppenheim & Nozari, 2024). According to such accounts, when naming an item (i.e., producing the word "cat" in response to a picture or a thought) semantic neighbors like "dog" are activated through shared semantic features. As a consequence of the word production episode, the connections between the target lexical item "cat" and its semantic features grow stronger, whereas the connections between the competitor “dog” and its semantic features grow weaker. Such adjustments in the system’s connections, which put “dog” at a disadvantage if it becomes the target in future trials, are long-lasting and thus have been argued to constitute learning (Oppenheim et al., 2010).

Learning accounts of semantic interference in naming known pictures naturally extend to learning new vocabulary. In support of this extension, recent studies have found a detrimental effect of semantic similarity between items in a training set on learning an artificial language in young neurotypical adults (Korochkina et al., 2021; Waller et al., 2025; see also Pérez-Serrano et al., 2022, for a similar finding in natural language learning). For example, Korochkina et al. (2021) trained young adults in two sessions, 24 hours apart, on new labels for familiar objects grouped in semantically related and unrelated conditions. A negative impact of semantic similarity on learning was reported throughout training and immediately after training was over. This impact was evident in several tasks, including picture-word interference and translation tasks. Similarly, in two large-scale studies, Waller et al. (2025) taught young adults new labels for unfamiliar objects, grouped into semantically related, phonologically related, or unrelated blocks. They found a persistent negative impact of semantic similarity on the comprehension of new vocabulary, which persisted for at least 48-72 hours after the training session. Evidence of semantic interference has also been reported for learning new written names. Breining et al. (2019) trained 17 neurotypical adults on semantic features and written names of novel objects, in semantically related, segmentally related, and unrelated blocks, and found that both semantic and segmental similarity interfered with learning.

These results naturally point to the possibility that a similar effect may be at work for relearning lost vocabulary in the first language of IWA. As one of the studies with the largest sample size testing this issue, Martin, Fink, Laine, et al. (2004) tested 11 IWA on blocks of semantically related, phonologically related, and unrelated items. All pictures were presented together. The experimenter pointed to and named each picture, and the participant repeated the name. This procedure was repeated four times. After training, the participant was presented with each picture one at a time and asked to name them (during-training measure). The process was then repeated eight times. Five minutes after the end of training, the set of five trained items and five untrained items were presented to the participant to name (short-term retention measure). The results showed more errors during training in the semantically related compared to the unrelated block. For short-term retention, the authors did not specifically report a statistical comparison between semantically related and unrelated conditions. Nonetheless, the numbers reported in the tables suggest a numerically greater improvement in the unrelated (53%)

compared to the semantically related (45%) condition, a difference that does not reach statistical significance. A series of smaller studies (N = 1-4) found similar patterns of semantic interference during training, without focusing on short-term comparison of the effect of semantically similar and unrelated conditions (Fink et al., 2009; Renvall et al., 2005, 2007). Moreover, none of these studies measured longer-term retention, calling for investigations of the longer-term effects of semantic similarity during training on learning in IWA.

Activation vs. Inhibition deficits in lexical retrieval and their relation to semantic similarity

As alluded to in the earlier sections, semantic interference in production arises from the dynamics of retrieving lexical items from their semantic features. This is a key locus of deficit of a subgroup of IWA whose dominant error type in picture naming tasks is lexical substitutions (i.e., producing the wrong word) rather than segmental substitutions (i.e., producing the wrong sound or letter) (Dell et al., 1997). Accordingly, individuals with this locus of deficit are said to have a deficit in lexical retrieval. When mild to moderate, the lexical substitutions manifest primarily as semantically related (e.g., “dog” for the target “cat”) and mixed semantic and phonological (e.g., “rat” for “cat”) errors. When more severe, they may grow more distant from the target in semantic space (Dell et al., 1997; Rapp & Goldrick, 2000) (see Black & Nozari, 2023, for a similar demonstration in segmental space in segmental retrieval deficits).

Lexical retrieval has two key components: activating lexical items from their semantic features and selecting the correct lexical items among all possible activated items. Either process could be impaired in aphasia; an inability to activate lexical items from semantic features creates an *activation deficit*. Selection mechanisms have been a matter of debate, with some researchers proposing competitive mechanisms (e.g., Abdel Rahman & Melinger, 2009; Roelofs, 1992; 2018), whereas others positing non-competitive mechanisms (e.g., Mahon et al., 2007). The competitive view maintains that a lexical item can only be selected if the difference between its activation and that of its competitors reaches a certain threshold (a *relative* selection threshold). This is usually accomplished through lateral inhibition: as one representation gains activation, it drives down the activation of its competitors (Coultrip, Granger, & Lynch, 1992; Markram *et al.*, 2004). The non-competitive view, on the other hand, claims that the activation of competitors has no bearing on lexical selection. As soon as any lexical item reaches a certain threshold, it will be produced (an *absolute* threshold). As such, the non-competitive position disputes a role for any kind of inhibition in lexical selection. For years, behavioral interference and facilitation effects have been taken to support one view or another, but recent computational work has shown that such evidence does not definitively support or reject either position (Oppenheim & Nozari, 2024).

This leaves open the question: Is inhibition involved in lexical selection? Nozari and Hepner (2019a, b) proposed a "flexible criterion" model of lexical selection, in which the system can flexibly shift its selection criterion based on task goals. If selecting a highly specific word is not absolutely critical to the communication goal, a non-competitive selection mechanism will be good enough. For example, when talking to a friend about a purchase, one may use "couch" or "sofa" interchangeably, with little consequence. Therefore, the first word that comes to the speaker's mind can be produced without regard for the activation of alternatives. On the other hand, the same speaker may be in a very different situation if she is teaching statistics and has to select the correct lexical item between the closely related terms "probability" and "likelihood". Here, rather than producing the first item that comes to mind, the speaker must closely weigh the activation of the two items and, if need be, inhibit the incorrect term to maintain her desired level of accuracy. A completely non-competitive system that has no inhibitory mechanism can explain the former, but not the latter case. On the other hand, a system that is equipped with inhibitory control mechanisms may use such mechanisms to various degrees depending on the task goals. In the couch/sofa case, inhibition can be set to near zero, making selection effectively non-competitive. In the latter case, inhibition can be ramped up continuously, thus making selection competitive to various degrees, depending on how critical accuracy is to the task goal.

For this reason, among others such as neurobiological plausibility (e.g., widespread evidence for inhibitory neurons in various cognitive systems, e.g., Coultrip, Granger, & Lynch, 1992; Markram *et al.*, 2004), we assume that the system is equipped with inhibitory control mechanisms for lexical selection and uses them when accuracy is critical to the task goal. From this assumption, it naturally follows that an impairment in selection may result from an *inhibition deficit*. Nozari (2019) introduced two IWA, XR and QD, who showcased activation and inhibition deficits, respectively. Both XR and QD had good semantic and lexical comprehension, intact auditory word repetition, and mild-to-moderate picture naming deficits, with semantic errors comprising the majority of their commission errors. As such, they exemplified relatively pure lexical retrieval deficits.

However, XR and QD's performance showed subtle but important differences. XR's picture naming was slow, and the majority of her responses were either omissions or single utterances. In contrast, QD's picture naming responses were relatively quick (compared to XR), and he often output several semantically related words in response to a target picture. This difference is compatible with an activation deficit in XR's case and an inhibition deficit in QD's. To further test this hypothesis, Nozari (2019) conducted additional tests on these two IWA's. A clear double dissociation was observed in a number of tests. When a task required activation —but not inhibition— of lexical items, XR's performance suffered, whereas QD's performance was unimpaired. This was showcased with the modified version of the Category Probe test (Freedman & Martin, 2001). This test presented participants with a list of auditorily presented words, followed by a probe word. Participants were asked to indicate whether the probe shared a semantic category with any of the items in the list. Activating lexical items and their links to semantic categories was necessary. However, selection amongst the listed words was not required, because any match would qualify as a "yes". XR's performance in the Category Probe test was markedly impaired, whereas QD performed at the level of healthy controls.

In contrast, when the task imposed higher demands on selection than activation, the pattern was reversed. For example, in the Miscue test (e.g., Lambon Ralph et al., 2000), participants were presented with a picture to name, together with an onset phoneme, presented aurally (e.g., a picture of a lion presented with the correct onset /l/ or the misleading onset /t/, which triggers the semantically related word "tiger"). They were told that the presented onset was only correct 50% of the time, so they should ignore it if they did not believe it matched the picture's name. Phonological cues are known to facilitate the activation of lexical items (e.g., Meteyard & Bose, 2018); therefore, the activation demands of this task are relatively low. However, since half of the phonological cues were misleading, participants had to inhibit the incorrectly activated responses. For both participants, naming accuracy benefited from phonological cues when they were correct, as expected from IWA. However, with incorrect cues, they showed different profiles: XR correctly rejected the majority of the incorrect cues. QD, on the other hand, often produced the incorrect name elicited by the misleading phonological cue, showing his difficulty in inhibiting the incorrect response.

A similar pattern was observed for the influence of semantic similarity on picture naming in a name reversal task (Nozari et al., 2016; Pinet & Nozari, 2023). In this task, participants were presented with several blocks of pictures to name. Each block contained only two pictures, which could be unrelated or semantically related (e.g., cat/pen or cat/dog). In the first half of the block, participants saw each picture several times in randomized order and named it. In the second half, when presented with each picture, they were instructed to name the other picture (say "cat" in response to the picture of dog, thus creating a need to inhibit a strongly activated non-target item). XR's performance in this task was better on related than unrelated blocks. She was helped by semantic similarity, presumably because activating lexical items was precisely the difficulty she experienced, and the semantic overlap among items increased the activation of the target words. Conversely, QD's performance suffered in the semantically related condition, compatible with a challenge in inhibiting the non-target items that had gained greater activation through their shared semantic features with the target.

It is reasonable to ask whether QD's inhibitory deficit is language-specific or extends to nonlinguistic tasks that require the inhibition of distractors. When tested on the classic Simon spatial-motor task (Lu & Proctor, 1995), which requires inhibition of prepotent motor responses, XR's performance fell within normal range, whereas QD's performance was impaired, similar to the linguistic tasks with similar demands. This implies that QD's deficit in suppressing competitors went beyond language. While often taken in the literature as evidence for a unitary impairment of domain-general inhibitory control, we prefer to view this problem more precisely as damage to domain-general neural regions involved in implementing control in several domains. The distinction is subtle but important (Nozari & Martin, 2024; Nozari & Novick, 2017). If there is a unitary control mechanism, it can be either healthy across all systems or damaged across all systems; there can be no selective control deficits in some systems but not others. On the other hand, if some parts of the control system are domain-general, but others are not, one may have global or selective control deficits if domain-general or domain-specific parts are damaged, respectively.

The evidence points to the latter possibility: Although a network of frontoparietal regions and the underlying white matter is involved in implementing control in multiple domains, including language production (Nozari, 2022, 2025; Shekari & Nozari, 2023; Wessel & Anderson, 2024), there is evidence that control is implemented through the connections between these regions and domain-specific task representations (e.g., Freund & Nozari, 2018; Zhang et al., 2025; Zhu et al., 2025). This means that if domain-general neural regions are healthy, control can still be selectively impaired in one domain but not another, if domain-specific representations or their connections to the domain-general centers are impaired. In this sense, control is task- and domain-specific, meaning that its damage does not need to be global. However, if domain-general control regions are damaged, as in the case with QD, the problem naturally manifests as a control problem that encompasses multiple domains, hence QD's impairment in the non-linguistic Simon task, along with control-demanding linguistic tasks.

One might also wonder if inhibition always requires the involvement of the domain-general control regions mentioned above. The answer depends on the task demands. In neurotypical speakers, if the stimulus activates the target more than the distractors (e.g., looking at the picture of a cat activates the target "cat" more than other animals, even if such animals have been previously named), those distractors could potentially be suppressed through lateral inhibition between the target and the distractors. This inhibition can be implemented locally (i.e., in the middle temporal gyrus) without involving the frontoparietal cortex. However, here too, domain-general regions can help processing by strengthening the connections between input and output (Burgess & Nozari, 2022; Nozari, Mirman, & Thompson-Schill, 2016). On the other hand, if the target is, for whatever reason, less potent than the distractors (e.g., meaning to call one child when your other child is standing in front of you), lateral inhibition alone may not be sufficient for arriving at the correct response. In such cases, domain-general control regions are needed to boost the information supporting the target in a goal-oriented fashion (Pinet & Nozari, 2023; Nozari, 2025a, b). This is why domain-general control regions are almost always involved when target selection requires overriding a prepotent distractor, but their involvement in target selection in semantically related contexts is inconsistent (e.g., de Zubicaray et al., 2013; Piai et al., 2016). In short, selection of the target over distractors can be accomplished by either lateral inhibition alone or by combining top-down control from domain-general regions with such inhibition.

The double dissociation observed in the performance of XR and QD in task conditions emphasizing activation or inhibition is particularly intriguing with regard to the influence of semantic similarity on performance. As described above, more semantic similarity was generally helpful to XR, as it helped her activate lexical items. In contrast, semantic similarity often had a detrimental effect on QD's performance, as it exacerbated his problem with inhibiting non-target lexical items. The current study was designed to examine the long-term consequences of semantic similarity on the re-learning of lost word forms in these two types of deficits.

Current study

This study trained XR and QD to relearn lost word forms in order to assess the influence of semantic similarity on learning in activation and inhibition deficits. We first discuss the general design features and then focus on the specific manipulations and predictions.

General design features. We integrated several well-known principles of learning in our design. These included repetition, retrieval practice, and semantic feature analysis (SFA). Repetition is known to facilitate lexical retrieval (Howard et al., 2006) and is a routine feature of aphasia rehabilitation (e.g., Fillingham et al., 2005; Nickels, 2002). Retrieval practice refers to improved learning when the learner is encouraged to retrieve the answer rather than study it (Karpicke & Aue, 2015; Karpicke & Roediger III, 2007; Zaromb & Roediger, 2010). While not necessarily more effective in incidental vocabulary learning (Waller et al., 2024) encouraging retrieval is often beneficial for learning in pedagogical settings both for children learning a language and for IWA relearning their lost language (Middleton & Schwartz, 2012; Saxton, 1997; Saxton et al., 1998).

Additionally, SFA is a technique in which semantic cues are used to aid lexical retrieval (Boyle & Coelho, 1995; Efstratiadou et al., 2018). For example, to help retrieve the word "pen", the learner is encouraged to answer questions such as "What do you do with it?" or "Where do you usually find it?". The idea is for such questions to activate semantic knowledge, e.g., writing, school, office, etc., which would, in turn, activate the target word. SFA has been successful in aphasia rehabilitation (Antonucci, 2009, 2014; Boyle, 2004, 2010; Coelho et al., 2000; Falconer & Antonucci, 2012; Kiran & Roberts, 2010; Peach & Reuter, 2010) with the exception of IWA with more profound semantic impairments (van Hees et al., 2013). Note that SFA is different from the manipulation of semantic context, in that semantic activation in SFA does not target the activation of a competitor, but the target itself.

The study incorporated these three techniques. Each participant was trained individually for six weeks, one hour a week. Each training session contained three conditions -related, unrelated, and mixed- each presented in three blocks consisting of three items, for a total of 27 training items. Items in each block were trained in two cycles of the three items to take advantage of repetition. Each training trial contained two retrieval attempts, taking advantage of retrieval practice. Finally, each training item was introduced along with two semantic features. Semantic features benefit learning most when they are distinctive for individual items in a category, as opposed to being shared between many members of a category (Breining et al., 2019; Hillis, 1998). For example, "banana" benefits more from a fairly unique feature such as "Monkeys like it." than a more generic feature "It is a fruit", especially if the training set contains other fruits. We included this variant of SFA to enhance target-specific semantic activation as much as possible without activating many semantic competitors.

We implemented customized training sets for each participant for two reasons. The first reason was methodological. To ensure that results reflected training, and not the well-known fluctuations in item accuracy among IWA, we used a double-baseline design and chose only items that participants could not name correctly on both occasions. No single set of well-balanced items fulfilled this criterion across both participants. The second reason was pragmatic. To keep individuals motivated to continue with a longitudinal study, we targeted items that they needed training on. Since XR and QD had trouble with different items, we opted for customized training sets, well-controlled for lexical frequency, length, and morphological complexity across conditions. We also ensured that the related categories were comparable across the two participants (birds, kitchen items, and fruits in XR, and mammals, home items, and food in QD).

Specific manipulations and predictions. Given the dissociations described earlier in the profiles of XR and QD, indicating activation and inhibition deficits, respectively, there is reason to expect semantic similarity to affect learning differently in these two individuals. Specifically, semantic similarity may be beneficial to XR and detrimental to QD. However, several factors may mediate the influence of similarity on learning in these IWA.

First, the timeline and longevity of the effect of semantic similarity on relearning the lost word forms are not well understood. While the detrimental effects of semantic similarity in neurotypical young adults persist for at least 48-72 hours after study (Waller et al., 2025), research on IWA has not investigated longer-term retention. Investigation this is important because short-term difficulties may turn into long-term benefits (E. L. Bjork & Bjork, 2011; R. A. Bjork, 1994). However, this is not always the case, which is why researchers have used the term “desirable difficulties” to refer to conditions where greater difficulty during learning results in benefits on retention (e.g., E. L. Bjork & Bjork, 2011), but identifying such conditions remains challenging even in neurotypical individuals (R. A. Bjork & Bjork, 2020). This study addresses this issue by measuring retention of trained vocabulary over various intervals: during training, immediately after training, between training sessions one week apart, and finally, two, six, and sixteen weeks after the last training session. Doing so allows us to map out the trajectory of the influence of semantic similarity on learning for each participant over time. Based on the current literature, we would expect semantic similarity to confer greater benefits to XR than to QD in short- and perhaps medium-term intervals. Long-term effects depend on whether semantic similarity during learning constitutes a desirable difficulty or not. Waller et al. (2025) showed that, for neurotypical young adults, it does not. What is hard during training remains hard during the test. Insofar as QD’s overall profile shows semantic interference in production tasks similar to neurotypical speakers, we would expect him to continue to show detrimental effects of semantic similarity on learning in longer retention intervals. As for XR, if initially, semantic similarity facilitates performance in the semantically related condition, the relatively difficult condition should be the unrelated one. If desirable difficulty is at work, then items in the unrelated condition should benefit from it and be better retained over time. However, if the benefits of semantic similarity for semantically related items persist over time, then we would expect better long-term retention of items in the related condition for XR.

Another factor of interest is the potential modulation of the influence of semantic similarity on learning through the clustering of related items. As mentioned earlier, psycholinguistic studies have shown that semantic interference is, to some extent, immune to the passage of time and intervening items (Hepner & Nozari, 2020; Schnur, 2014). However, these results come from studies of naming and not from training studies. Moreover, even in simple picture naming studies, there appears to be a limit to the number of intervening items between two semantically related items, beyond which semantic interference dissipates (Schnur, 2014). In training, the issue gains practical importance: are the effects of semantic similarity only experienced for items trained together as a unit, e.g., within a block of training trials? Or are such effects farther-reaching and encompass any trained item in the category, even when they are trained in different units and blocks within a session? We address this issue by including two different control conditions in our study. In addition to related and unrelated blocks, we also include mixed blocks. Mixed blocks contain items that are locally unrelated (i.e., do not belong to the same category as any other item within the same block) but globally related (i.e., they share categories with items in the related blocks). If local similarity is required for learning to be affected, we would expect mixed blocks to behave more like unrelated blocks. If, on the other hand, the global similarity is sufficient to influence learning, mixed blocks would be expected to behave more like related blocks.

In sum, this study allows for the further evaluation of the distinct nature of activation and inhibition mechanisms in word production. We do so by evaluating the response profiles of these deficit types to a behavioral intervention protocol that evaluates the time-course of semantic similarity manipulations during training.

METHODS

Participants

Participants were the two IWA, XR and QD, reported in Nozari (2019). At the time of this study, XR was a 56-year-old female with a high-school education, 22 years post-stroke. QD was a 65-year-old male, also with a high school education, 12 years post-stroke. XR had been premorbidly left-handed and had extensive damage to her right hemisphere, encompassing temporal lobe, as well as parts of parietal and prefrontal cortex, although parts of her anterior, superior and ventrolateral prefrontal cortex had been spared (indicating right hemisphere language lateralization). QD had been premorbidly right-handed and had a large middle cerebral artery stroke encompassing almost all of the left prefrontal cortex, as well as significant parts of his parietal and temporal cortex (see Nozari, 2019, for structural MRIs). As explained in the Introduction, XR and QD were selected for this study because they both had lexical retrieval deficits, but showed activation and inhibition deficits, respectively.

Table 1 summarizes the similarities and differences in their language profiles. All background tests are described in detail in Nozari (2019), but we provide a conceptual overview of each test here. Both participants had good semantic comprehension, as measured by a task that requires participants to match a probe picture to the one of two pictures that is semantically closer to it. Since the task did not require any language knowledge, a high score indicates good conceptualization abilities and an intact structure of semantic knowledge. Participants also showed good lexical comprehension measured by a word-to-picture matching task based on the Philadelphia Naming Test (PNT, Roach et al., 1996), without semantic and phonological lures. In a more challenging task with both types of lures (Nozari, 2019), they showed slightly worse performance, but accuracy remained above 80%, indicating a good, if not perfect, ability to map lexical items onto their meaning. Both participants were almost perfect in auditory word repetition measured by the Philadelphia Repetition Test (PRT, Roach et al., 1996), indicating intact lexical-to-phonological mapping (Dell et al., 2013; Nozari et al., 2010; Nozari & Dell, 2013). Picture naming abilities showed a mild impairment on the PNT and a moderate impairment in the larger 444-picure set (Nozari, 2019) administered on two separate occasions. Relative performance was consistent, with both participants showing ~25% drop in accuracy from the PNT to the larger set, with XR maintaining a small advantage over QD. A closer inspection of their error profiles in picture naming revealed that the majority of their commission errors were semantically related (79% and 82% on the PNT, and 85% and 84% on the larger set, for XR and QD, respectively; less than 7% omission errors in both). Collectively, the profiles described above point to a lexical retrieval deficit in word production. Despite these similarities, XR and QD showed different picture naming profiles. XR's response times (RTs) were significantly slower than QD's, and she often produced a single utterance in response to a picture. In contrast, QD responded more quickly and produced twice as many multiple-utterance responses as XR, with the majority of these utterances being semantically related to one another.

Table 2 shows the double dissociation[1] in the performance of XR and QD, leading to the identification as cases of activation and inhibition deficits, respectively. Aside from the differences described in their picture naming, XR was impaired in the Category Probe test which taps into activation of lexical items and their links to semantic categories, but scored within the normal range in tasks that measure inhibitory control (Simon, Miscue, and Word Stroop; see Nozari, 2019, for details of each task and Introduction for a summary). Conversely, QD was unimpaired on the Category Probe test, but was impaired on all of the tasks that required the suppression of a prepotent response.

Table 1. Descriptive statistics showing similarities and differences in the language profiles of XR and QD. Multi-response trials refer to trials where more than one response was produced. Comp = comprehension; Lex = lexical; PNT = Philadelphia Naming Test; PRT = Philadelphia Repetition Test; RT = response time; Sem = semantic. See text and Nozari (2019) for more details on the tasks.

	Similarities						***Differences***	
	Sem Comp	**Lex Comp (no lures)**	**Lex Comp (with lures)**	**PRT**	**PNT**	**Expanded naming test**	**PNT RTs (ms)**	**Multi-response trials**
XR	96%	99%	83%	98%	88%	63%	2064 (SE = 142)	14%
QD	96%	99%	90%	98%	76%	50%	1658 (SE = 128)	28%

Table 2. Profiles across multiple tasks of activation deficit (XR) and inhibition deficit (QD). See text and Nozari (2019) for task details

Task	***XR (Activation deficit)***	***QD (Inhibition deficit)***
Picture naming	Long RTs, many omissions, single responses on most trials	Short RT, multiple semantically related responses on most trials
Category probe task	Impaired	Unimpaired
Simon task	Unimpaired	Impaired
Miscue task	Few miscue errors	Many miscue errors
Name reversal task	Semantic similarity: beneficial	Semantic similarity: detrimental

Materials

[1] Statistical tests for differences between XR's and QD's performance and their differences from control samples have been reported in Nozari (2019) and are not repeated here. The reported dissociations are supported by those statistical analyses.

Materials were customized to each participant. The goal was to choose items that the participant had persistent difficulty naming, thus making it less likely that the item is named correctly during/after training by chance. To this end, we used a double-baseline design. Each participant named a large set of 444 pictures twice (Nozari, 2019) over several sessions. We selected only items that the participant failed to name on both occasions. From this subset, we selected 27 items for training each participant. These items formed three conditions: *related, unrelated*, and *mixed*, each containing three blocks of three items. Items in each related block were semantically related (e.g., fruits: mango/raspberry/apricot). For XR, related blocks were birds, kitchen items, and fruits; for QD, mammals, home items, and food. Items in the unrelated blocks were not from the same category as each other or any items in the related blocks (e.g., chimney/notebook/jet). Items in the mixed blocks were not related to each other, but each was from one of the three related categories (e.g., dove, tongs, and grapefruit, from the related categories birds, kitchen items, and fruits, respectively). We also ensured that items were matched in lexical frequency (Brysbaert & New, 2009), length (number of phonemes), and morphological complexity (number of morphemes) across the three conditions (see Table 3).

Table 3. Characteristics of the training materials for XR and QD.

	Frequency (log)	**Phonemes**	**Morphemes**
XR			
Related	1.82	6.11	1.56
Unrelated	2.1	6.22	1.33
Mixed	1.84	5.78	1.56
Mann-Whitney U Test			
Related vs. Unrelated	z = 0.92, p = 0.35	z ~ 0, p ~ 1.00	z= 0.45, p = 0.62
Related vs. Mixed	z= 0.09, p = 0.93	z = 0.27, p= 0.79	z = 0.04, p= 0.97
QD			
Related	2.37	5.11	1.22
Unrelated	2.44	5.11	1.33
Mixed	2.10	5.33	1.33
Mann-Whitney U Test			
Related vs. Unrelated	z = 0.47, p = 0.64	z = 0.13, p = 0.90	z = 0.04, p = 0.97
Related vs. Mixed	z = 1.63, p = 0.10	z ~ 0, p ~ 1.00	z = 0.35, p = 0.73

Each training item comprised a colored picture from the 444-item large naming test (Nozari, 2019). For the current study, two features were identified for each item to act as semantic cues for retrieving the name. Rather than features widely shared between the members of a taxonomic category, we attempted to choose features that were either unique to the item or at least constrained to a small subset of items within the taxonomic category. For example, if the trained item was “banana”, features would be “Monkeys like this” and “One may slip on its skin”. Importantly, the selected features were not shared with any of the other items within the training set, so as to increase the semantic distinctiveness of training items (Breining et al., 2019).

In addition to the training items, a set of 18 items was set aside as the untrained set. The untrained set consisted of items either semantically related to those trained in the study or not, but since this set was never trained, there could not be a “mixed” condition. For the related and unrelated conditions, we selected 9 related and 9 unrelated items, matched to the participants’ trained items in these two conditions in lexical frequency, length, and morphological complexity (all p values > 0.1). Untrained items were selected using the same baselining procedures as the trained items, but were never encountered after baselining until the three long-term retention tests.

Procedures

Training

The same procedure was implemented in the training and testing of both participants. Figure 1 shows the structure of a training session (Fig. 1a), a block (Fig. 1b), and a trial (Fig. 1c). Each training session contained all nine blocks, presented in pseudorandomized order, such that no two blocks from the same condition were adjacent. At the beginning of each block, the participant was explicitly told the block category. For example, “In this block, we will learn about fruits”. For unrelated and mixed blocks, the participant was told that “This block contains a variety of items from different categories”. Then the first item within the block was introduced.

Each block contained three items, trained in two cycles (Fig. 1b). Cycle 2 only began after all three items had been trained in cycle 1. The order of items was pseudorandomized across the two cycles, with the restriction that the same item was never trained twice in a row (i.e., the last item in cycle 1 was never the first item in cycle 2). Each item within a cycle was trained using the following procedure (Fig. 1c): first, the participant saw a colored picture of the item and heard a prompt, “What is this?” to encourage them to attempt a retrieval (retrieval practice). After 3000 ms, the two written descriptions of the item’s features appeared next to the picture and were read aloud to the participant. The participant was told to think about the features and the object’s name together, but they did not have to repeat the features. The item’s name then appeared underneath it in capital letters and was read aloud to the patient. The participant was asked to repeat the name. Next, the participant was asked whether they remembered the item's features. No verbal response was required, but the participant was encouraged to think about the item's features. Finally, the picture appeared one last time, and the participant was asked to name it. If no name was given within 3000 ms, the features were repeated. Participants could name the picture at any point while they listened to the features. The participant's accuracy was registered at this point. The label then appeared underneath the picture and was read aloud. If the participant had failed to name the object spontaneously, they were asked to repeat the name before moving on to the next trial.

The experimenter was present throughout training to monitor and regulate the session. While he did not name the item or the features, he guided the participant through the session by prompting them when it was time to name an object or by encouraging them to think about the item's features. Participants worked at their own pace and could take breaks between blocks. None of the training sessions lasted more than an hour. Training sessions were audio-recorded, and participants' responses on the last naming attempt of each trial were transcribed and coded. This provided two data points per item (one from cycle 1 and one from cycle 2) during each training session.

Each participant received one-on-one training with the experimenter, with the structure described above, once a week. A predetermined stopping criterion was set for stopping training when at least 2/3 of items were produced accurately during a "pre-training" test (see below for details of tests). This value was selected to ensure a reasonably strong starting point for examining long-term retention, while taking into account the difficulties faced by IWA in consistently accessing labels, even when they know them. Note that this criterion did not take experimental conditions into account and was thus orthogonal to the hypotheses of the study. Both XR and QD reached the criterion in week 6.

Testing

Figure 2 shows the overall structure of the study, including testing points. Testing was conducted in three phases to assess retention in three time scales: *short-term retention, medium-term retention,* and *long-term retention*. At each test, 27 items were presented to the participant, one at a time, in randomized order, and the participant had 10 seconds to name each item. Short-term retention measured learning 5 minutes after training in each session (post-training). Medium-term retention followed a similar testing format but was administered at the beginning of each training session, before training on that session (pre-training). As such, it measured retention from the previous training session, one week prior. Finally, long-term retention measured retention after training had ended, i.e., week 6. Long-term retention was assessed at three time points: 2, 6, and 16 weeks after the last training session. In addition to the trained items, we also measured naming accuracy for untrained items at the long-term retention time-points to evaluate possible generalization. The test set for the 27 trained items was administered separately from that of the 18 untrained items, to keep the trained-item test settings identical across all retention intervals. No feedback was provided on any of the tests.

To recap, each training session started with a medium-term retention test (pre-training), followed by training, and ended with a short-term retention test (post-training). On long-term retention tests, participants only completed two naming tests, one on the 27 trained items and one on the 18 untrained items.

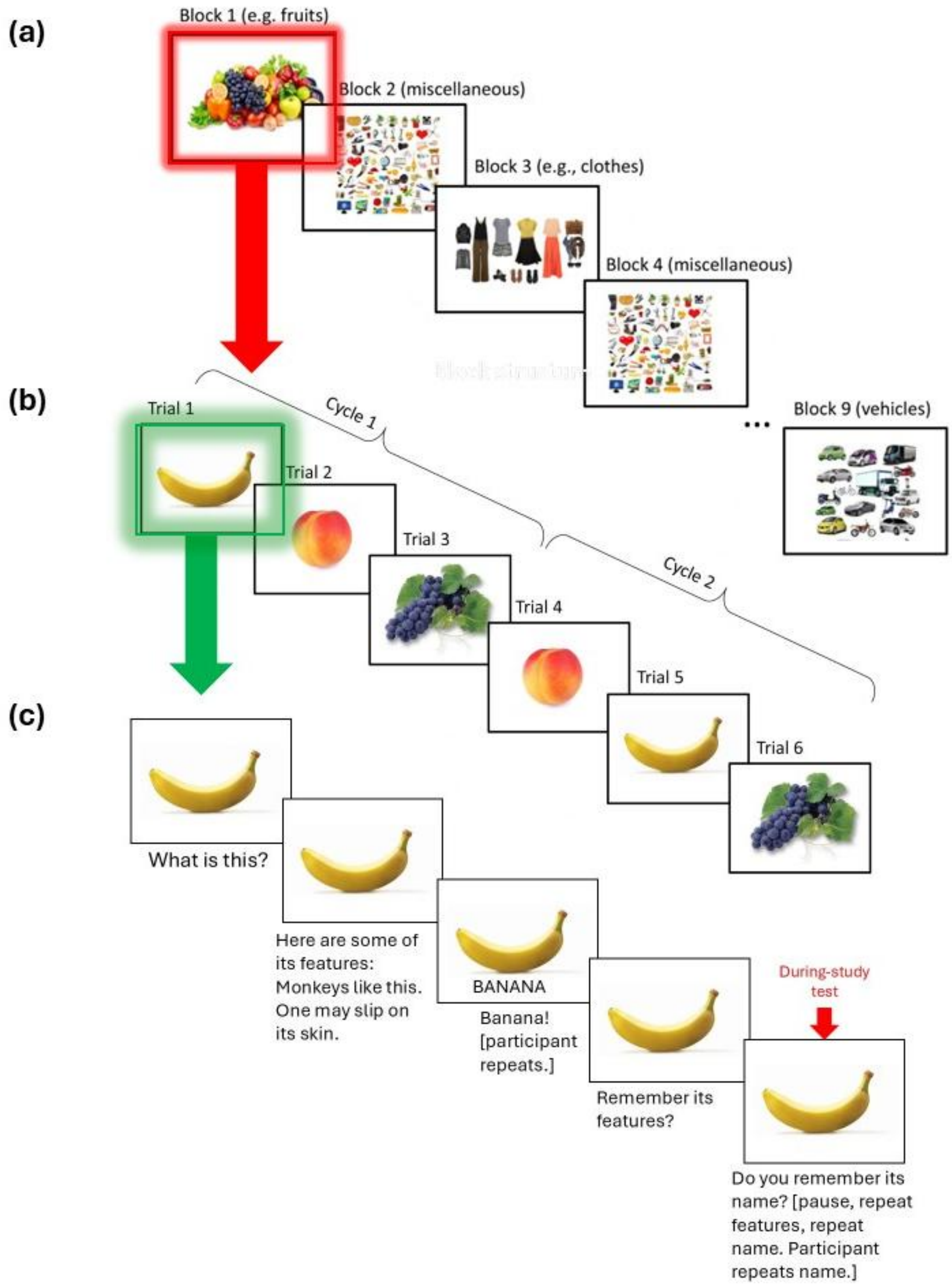

Figure 1. Structure of a training session (a), a block (b), and a trial (c). Categories and items are given as examples.

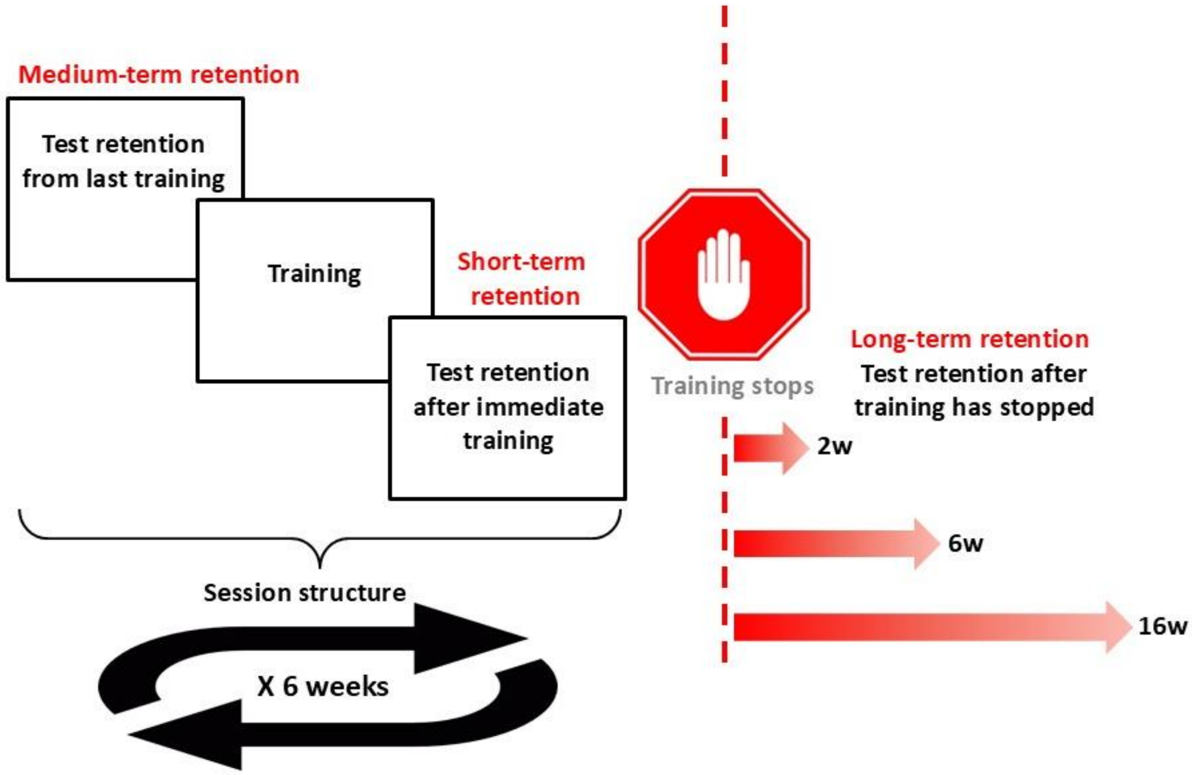

Figure 2. Overall structure of the study. Six training sessions (by which time the pre-determined stopping criterion had been reached), one week apart, were followed by three long-term retention tests, 2, 6, and 16 weeks after the last session. Short-term retention was assessed 5 minutes after training in each session. Medium-term retention was assessed at the beginning of each session (i.e., one week after the last training session, except for session 1, which was not preceded by training).

Transcription and coding

All of the training and testing sessions were audio-recorded. Participants' responses to critical trials, i.e., the naming attempts of cycle 1 and cycle 2 trials, as well as all the test sessions, were transcribed and coded by a trained native-speaking RA. A second coder checked transcriptions for consistency, and disagreements were reconciled. The first response, identified as defined by PNT rules, was coded for accuracy. Accuracy was coded leniently, i.e., minor phonological and phonetic aberrations, such as simplification of consonant clusters (e.g., /naɪtsænd/ for "nightstand") were accepted as correct, as they demonstrated correct access and ordering of the morphological components of the target word. Other responses, including no responses, descriptions without a naming attempt, and substitutions, were coded as errors. Given that, as described in their profiles, XR was known to be a slower responder than QD, response times were not a targeted measurement of the current study.

Analysis

Data were analyzed using the *lme4* package (Bates, 2014) in R (version 4.2.2, R Core Development Team, 2022). P-values were based on Satterthwaite approximates using the *LmerTest* package (version 3.1-3, Kuznetsova, Brockhoff, & Christensen, 2016). Four sets of analyses were conducted on the data. The first set examined accuracy during **training**. This test averaged accuracy over the two naming attempts at the end of cycles 1 and 2 of training each item, for all items within a condition across the six training sessions[2]. Condition was coded as two contrasts (related vs. mixed and related vs. unrelated), and a general linear model was fit to each participant's data to compare accuracy for the two contrasts. The second set of analyses tested accuracy for **short-term retention**, conducted on the post-training data. This set contained two levels of analyses: a first-level general analysis on all data and a second-level subject-specific analysis. The goal of the first-level analysis was to uncover any main effects of semantic similarity, as well as its potential modulation by participants and/or sessions. We used a generalized linear model with binary accuracy as the dependent variable (i.e., a logistic regression model) and Condition, Session Number, Participant, and their 2- and 3-way interactions as independent variables of interest. We also included lexical frequency and number of phonemes as control variables. As before, Condition was coded as two contrasts (related vs. mixed and related vs. unrelated). The goal of the second-level analysis was to examine the effect of semantic similarity and its possible dependency on session within individual participants. This analysis used similar models, but without Participant as a variable, and evaluated each participant's data individually. The third set of analyses evaluated accuracy for **medium-term retention**. These models were identical to those described for assessing short-term retention but used pre-training data. Finally, the fourth set of analyses tested accuracy for **long-term retention** conducted on data obtained 2, 6, and 16 weeks after the final training session. This set followed the same 2-level pipeline described above. However, since there was no active manipulation between the three follow-up testing sessions, Session was only included as a control variable to capture possible decay, as opposed to a variable actively interacting with similarity manipulations. In addition, we also

---

[2] On some trials, after responding in cycle 1, participants continued to repeat the response until the end of cycle 2. Therefore, the two responses could not be counted as independent, hence the decision to average the two.

measured the accuracy of untrained items. Since very few of the untrained items were accurately named, there was not enough variability in the data to carry out statistical analyses on the untrained set.

RESULTS

Analysis 1: Training (during training)

Table 4 presents the accuracy results during training. Both participants showed the lowest accuracy in the related condition. XR's accuracy in the related condition was significantly lower than the unrelated condition (t = 2.087, p = 0.039), but not significantly different from the mixed condition (t = 0.96, p = 0.340). QD's accuracy in the related condition was significantly lower than both the unrelated (t = 2.03, p = 0.044) and the mixed (t = 2.84, p = 0.005) conditions[3].

Table 4. Percentage of accurate trials (SE) during training for each condition.

	Related	Unrelated	Mixed
XR	0.75 (0.03)	0.93 (0.02)	0.81 (0.02)
QD	0.59 (0.03)	0.65 (0.03)	0.79 (0.02)

Analysis 2: Short-term retention (end of session)

Figure 3a shows short-term retention accuracy (i.e., post-training) in each of the three conditions at the end of sessions 1-6 for each participant. Table 5 shows the full results of the first-level general model. This model revealed that learning improved with session, that accuracy was higher for higher-frequency items, and that XR had higher overall accuracy than QD for short-term retention. Of interest is the finding of significantly higher accuracy in the related compared to the unrelated (t = 2.76, p = 0.006) but not the mixed (t = -0.36, p = 0.721) condition. This effect interacted with participant (t = -2.40, p = 0.016) and session number (-2.59, p = 0.01). There was also a three-way interaction between condition, participant, and session number (t = 2.63, p = 0.009).

[3] These results were double checked with the non-parametric Wilcoxon signed-rank test, which is more conservative than a linear regression model, but makes no assumptions regarding the distribution. Accuracy in the related condition was still significantly lower than the unrelated condition for XR (z = 2.90, p = .004) and the mixed condition for QD (z = 2.64, p = .008). However, QD's accuracy was not significantly different between the related and unrelated conditions when tested by the conservative non-parametric test.

To unpack the interactions, the second-level analyses were conducted separately on each participant. Tables 6 and 7 show the full results of these analyses for XR and QD, respectively. While accuracy improved with session and was higher for items with higher lexical frequency, no other effect was significant in XR. In contrast, for QD, learning was significantly better in the related compared to the unrelated ($t = 2.73$, $p = 0.006$) but not the mixed ($t = -0.75$, $p = 0.45$) condition. The advantage of relatedness also interacted with session ($t = -2.59$, $p = 0.01$), such that it decreased by session.

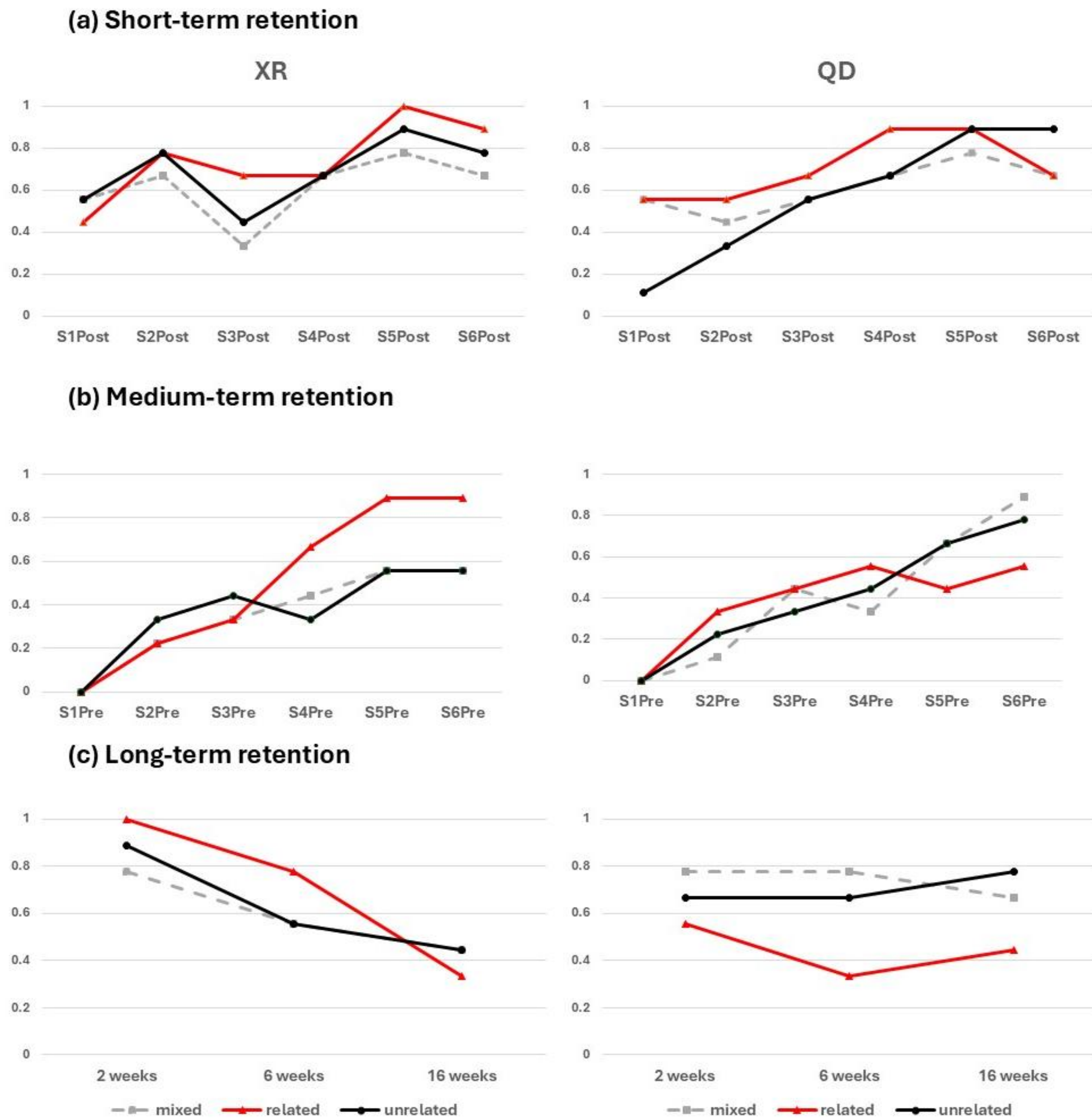

Figure 3. Observed accuracy for short-term, medium-term, and long-term retention.

To summarize, the results showed that short-term retention was better for items in the related condition than those in the unrelated condition, but only for QD, and that this advantage decreased in later sessions.

Table 5. Results of the general model examining short-term retention.

	Coefficient	SE	z	p value
Intercept	-2.886	0.961	-3.003	0.003
Condition 1* (Related vs. Mixed)	-0.354	0.989	-0.357	0.721
Condition 2 (Related vs. Unrelated)**	3.489	1.262	2.764	0.006
Participant	1.262	0.627	2.013	0.044
Session Number	0.586	0.141	4.149	< .001
Lexical Frequency	0.583	0.279	2.091	0.037
Number of phonemes	0.056	0.085	0.659	0.51
Condition1 x Participant	0.759	1.366	0.556	0.578
Condition2 x Participant	-3.794	1.582	-2.398	0.016
Condition1 x Session Number	0.107	0.268	0.398	0.69
Condition2 x Session Number	-0.979	0.378	-2.589	0.01
Participant x Session Number	-0.299	0.177	-1.687	0.092
Condition1 x Participant x Session Number	-0.438	0.38	-1.151	0.25
Condition2 x Participant x Session Number	1.233	0.469	2.631	0.009

* Condition 1 represents the contrast (Related vs. Mixed) within the factor Condition.

** Condition 2 represents the contrast (Related vs. Unrelated) within the factor Condition.

Table 6. Results of the model examining short-term retention for XR.

	Coefficient	SE	z	p value
Intercept	-4.855	1.402	-3.464	0.001
Condition 1 (Related vs. Mixed)	0.449	1.011	0.445	0.657
Condition 2 (Related vs. Unrelated)	-0.009	1.018	-0.009	0.993
Session Number	0.323	0.115	2.799	0.005
Lexical Frequency	1.891	0.488	3.874	< .001
Number of phonemes	0.17	0.115	1.475	0.14
Condition1 x Session Number	-0.39	0.289	-1.349	0.177
Condition2 x Session Number	0.316	0.294	1.078	0.281

Table 7. Results of the model examining short-term retention for QD.

	Coefficient	SE	z	p value
Intercept	-0.335	1.306	-0.257	0.798

Condition 1 (Related vs. Mixed)	-0.749	0.994	-0.753	0.451
Condition 2 (Related vs. Unrelated)	3.429	1.258	2.725	0.006
Session Number	0.58	0.14	4.131	< .001
Lexical Frequency	-0.298	0.377	-0.79	0.43
Number of phonemes	-0.065	0.146	-0.445	0.657
Condition1 x Session Number	0.099	0.266	0.374	0.708
Condition2 x Session Number	-0.976	0.377	-2.586	0.01

Analysis 3: Medium-term retention (week to week)

Figure 3b shows medium-term retention accuracy (i.e., pre-training) in each of the three conditions for sessions 1-6 for each participant. Table 8 shows the full results of the first-level general model. As with short-term retention, the general model revealed that learning improved with session and that accuracy was higher for higher-frequency items. There was no main effect of condition for either contrast, but both contrasts showed a significant three-way interaction with participant and sessions ($t = -2.67$, $p = 0.007$ for related vs. mixed, and $t = 2.60$, $p = 0.009$ for related vs. unrelated). These significant three-way interactions point to different effects of semantic similarity in participants across the sessions and are unpacked in the second-level analyses.

Table 8. Results of the general model examining medium-term retention

	Coefficient	**SE**	**z**	**p value**
Intercept	-5.518	1.197	-4.61	< .001
Condition 1 (Related vs. Mixed)	-1.656	1.375	-1.204	0.228
Condition 2 (Related vs. Unrelated)	1.58	1.275	1.239	0.215
Participant	0.497	0.79	0.629	0.53
Session Number	0.77	0.143	5.377	< .001
Lexical Frequency	1.179	0.328	3.597	< .001
Number of phonemes	-0.057	0.096	-0.598	0.55
Condition1 x Participant	3.1	1.957	1.584	0.113
Condition2 x Participant	-3.203	1.856	-1.726	0.084
Condition1 x Session Number	0.689	0.343	2.007	0.045
Condition2 x Session Number	-0.486	0.315	-1.544	0.123
Participant x Session Number	-0.017	0.195	-0.087	0.931
Condition1 x Participant x Session Number	-1.331	0.498	-2.674	0.007
Condition2 x Participant x Session Number	1.227	0.471	2.603	0.009

Tables 9 and 10 show the full results of second-level analyses for XR and QD, respectively. For XR, both condition contrasts showed a significant interaction with session (t = -2.13, p = 0.033 for related vs. mixed, and t = 2.38, p = 0.017 for related vs. unrelated), showing higher accuracy in the related condition compared to the other two in later sessions (see Fig. 2b). QD only showed a marginal interaction between related vs. mixed condition and session (t = 1.86, p = 0.063) for medium-term retention.

Table 9. Results of the model examining medium-term retention for XR

	Coefficient	**SE**	**z**	**p value**
Intercept	-8.739	1.921	-4.55	< .001
Condition 1 (Related vs. Mixed)	2.028	1.51	1.343	0.179
Condition 2 (Related vs. Unrelated)	-1.76	1.455	-1.21	0.226
Session Number	0.887	0.161	5.527	< .001
Lexical Frequency	2.591	0.594	4.359	< .001
Number of phonemes	0.022	0.133	0.164	0.87
Condition1 x Session Number	-0.836	0.393	-2.129	0.033
Condition2 x Session Number	0.921	0.387	2.38	0.017

Table 10. Results of the model examining medium-term retention for QD

	Coefficient	**SE**	**z**	**p value**
Intercept	-3.203	1.423	-2.252	0.024
Condition 1 (Related vs. Mixed)	-1.801	1.336	-1.348	0.178
Condition 2 (Related vs. Unrelated)	1.503	1.251	1.201	0.23
Session Number	0.718	0.137	5.238	< .001
Lexical Frequency	0.31	0.4	0.775	0.438
Number of phonemes	-0.093	0.155	-0.598	0.55
Condition1 x Session Number	0.611	0.329	1.856	0.063
Condition2 x Session Number	-0.475	0.309	-1.539	0.124

To summarize, the results showed that similarity had significantly different effects on medium-term retention for XR and QD. For XR, relatedness gradually benefited medium-term retention accuracy from session 1 to 6. No such benefit was observed in QD.

Analysis 4: Long-term retention (2-16 weeks)

Figure 3c shows long-term retention accuracy (i.e., 2, 6, and 16 weeks after the last training session) in each of the three conditions for each participant. Table 11 shows the full results of the first-level general model. As expected, overall retention was better in shorter intervals after training, and lexical frequency remained a predictor of retention. But important for our purpose, there was a main effect of similarity in the comparison between related and mixed conditions (t = -2.86, p = 0.004), which significantly interacted with participant (t = 2.74, p = 0.006). Second-level analyses unpack this interaction.

Table 11. Results of the general model examining long-term retention

	Coefficient	SE	z	p value
Intercept	-0.197	1.149	-0.172	0.864
Condition 1 (Related vs. Mixed)	-2.014	0.704	-2.862	0.004
Condition 2 (Related vs. Unrelated)	-0.309	0.727	-0.425	0.671
Participant	-0.616	0.641	-0.961	0.337
Session Number	-0.092	0.031	-2.983	0.003
Lexical Frequency	0.897	0.394	2.278	0.023
Number of phonemes	0.052	0.121	0.435	0.664
Condition1 x Participant	2.57	0.94	2.735	0.006
Condition2 x Participant	0.222	0.916	0.243	0.808

Tables 12 and 13 show the full results of the second-level analyses for XR and QD, respectively. XR showed a gradual loss of retained vocabulary over sessions, and her retention remained sensitive to lexical frequency, with better retention of high-frequency vocabulary. However, long-term influences of semantic similarity were not evident in her performance (t = -1.19, p = 0.24; t = 1.35, p = 0.18, for the comparison of related vs. mixed and unrelated conditions, respectively). QD showed a very different long-term retention profile. His retention was relatively stable over the three follow-up testing sessions and not sensitive to lexical frequency. Importantly, semantic similarity had a clear and detrimental effect on his long-term retention; accuracy was significantly lower for items trained in the related condition compared to both mixed (t = 2.40, p = 0.016) and unrelated (t = -2.63, p = 0.009) conditions.

To summarize, the main finding of these analyses was an interaction between the influence of semantic similarity on long-term retention and participant, whereas XR's performance did not suffer from semantic similarity, a clear detrimental effect was evident on QD's performance.

Table 12. Results of the model examining long-term retention for XR

	Coefficient	SE	z	p value
Intercept	-0.492	1.763	-0.279	0.780
Condition 1 (Related vs. Mixed)	-0.806	0.680	-1.185	0.236
Condition 2 (Related vs. Unrelated)	0.984	0.728	1.352	0.176
Session Number	-0.168	0.048	-3.503	< .001
Lexical Frequency	1.818	0.711	2.557	0.011
Number of phonemes	-0.141	0.166	-0.847	0.397

Table 13. Results of the model examining long-term retention for QD

	Coefficient	SE	z	p value
Intercept	-2.699	1.892	-1.426	0.154
Condition 1 (Related vs. Mixed)	1.722	0.717	2.403	0.016
Condition 2 (Related vs. Unrelated)	-1.745	0.665	-2.625	0.009
Session Number	-0.024	0.045	-0.54	0.589
Lexical Frequency	0.516	0.5	1.031	0.303
Number of phonemes	0.476	0.261	1.826	0.068

Untrained items

Accuracy was also compared for untrained items that were either from the semantic categories in the related blocks or not. Not surprisingly, accuracy was low on these items and was therefore averaged over the three follow-up sessions. The average accuracy for untrained items was 13% for XR and 19% for QD. In both participants, this accuracy was significantly lower than that of the trained items (XR: $z = 3.67$, $p < .001$; QD: $z = 3.64$, $p < .001$). Within the untrained set, XR's accuracy was 15% and 11% for related and unrelated items, respectively. QD's accuracy was 19% for both. While the data are too sparse for a formal comparison of different conditions, it is clear from these results that the benefits or harms of similarity do not readily extend to untrained items.

Error profiles

Errors made during pre-training, post-training, and follow-up tests were divided into four groups: (1) within-block errors, where the name of another object trained within the same block was given, (2) between-block errors, where the name of an object trained in a different block was given, (3) semantic errors, where a semantically related object was named that was not part of this study, and (4) others including no responses. Most of XR's errors were single semantic errors (57%), whereas the majority of QD's errors were semantic errors with multiple attempts (60%), compatible with their original profiles of inhibition and activation deficits, respectively. Between-block errors comprised 9% and 3% of XR's and QD's responses, respectively, and were mostly caused by the overlap between the categories of items in the semantic and mixed blocks. Most relevant to the hypotheses of the study are within-block errors, which were infrequent (3% and 2% of all errors in XR and QD, respectively), most likely because of the small number of items per block. However, they were observed exclusively in the related blocks, showing that semantic similarity in temporally adjacent items promotes confusion.

DISCUSSION

We set out to examine the influence of semantic similarity on relearning of lost word forms in two IWA, with activation and inhibition deficits in lexical retrieval. We trained them weekly to a predefined criterion, which both participants reached in week 6, after which we stopped training. In addition to comparing learning in related vs. unrelated and mixed blocks in a longitudinal manner, our training protocol took advantage of repetition, retrieval practice, and SFA to maximize learning. We mapped out learning and retention systematically, from during training all the way to 16 weeks after the last training session. While results of single-case studies are invariably noisy, there were clear patterns observed in these data that match previous findings and our theoretical predictions. Below, we present a summary of the findings and discuss their implications for theory and practice.

With regard to the effects of similarity on word relearning, the key findings are as follows: (1) We replicated the finding of immediate semantic interference during learning reported in past studies (Fink et al., 2009; Martin et al., 2004; Renvall et al., 2005, 2007). XR showed worse performance in the related vs. unrelated blocks. QD also had the lowest accuracy in the related condition, although using the more conservative analysis, the difference was only statistically significant from the mixed condition. (2) In contrast, in the short term (in testing at the end of each training session), both participants showed higher accuracy for items in the related compared to the unrelated blocks, although the effect was statistically significant only for QD. However, this advantage grew weaker in later sessions. (3) Results diverged for XR and QD in the medium retention interval (one week after the last training session). While both participants showed a gradual increase in learning in all conditions, XR's slope of learning showed a steeper learning slope for items in the related condition than in the unrelated and mixed conditions. For QD, medium-term retention showed a reversed pattern, with learning in related blocks showing a disadvantage in later sessions.

Finally, (4) with regard to the long-term retention, overall we found that the training was effective in both participants, with approximately 60% of the trained items retained at least 6 weeks post-training. While XR's long-term accuracy dropped gradually from 2 to 16 weeks post-training, QD was able to sustain his acquired vocabulary over this long period. Most salient, however, was the finding that the influence of semantic similarity on retention was distinct in the two participants. Semantic similarity did not strongly modulate long-term retention for XR. If anything, the numerical advantage seen in the short and medium term for items trained in semantically related blocks was sustained during follow-up. In contrast, for QD, the detrimental effect of similarity seen in the final sessions of the medium-term testing was clearly observed during the follow-up period. This effect was present when items in the related block were compared to either unrelated or mixed blocks.

The dissociable influence of semantic similarity in activation vs. inhibition deficits

We designed this study expecting different effects of semantic similarity on XR and QD's performance. Our predictions were mostly borne out in the data. Here, we offer a psycholinguistic explanation of the observed dissociation, rooted in the dual influence of semantic similarity on lexical retrieval. Psycholinguistic studies have shown us that semantic similarity is a well-known double-edged sword, both facilitating and hindering performance through different mechanisms (e.g., Abdel Rahman & Melinger, 2009; Alario et al., 2000; Costa et al., 2009; Mahon et al., 2007; Oppenheim & Nozari, 2024; Wheeldon & Monsell, 1994). In a nutshell, semantic activation is crucial for activating the target lexical item. As such, any factor that increases the activation of target-related semantic features helps target lexical retrieval (a facilitation effect). At the same time, any non-target item that contributes to the activation of a target will also activate its own lexical representation. If the non-target representation gains too much activation, it can slow down or even overtake target production (an interference effect; Levelt et al., 1999; Nozari, 2025).

There has been much debate in the literature on the factors that ultimately swing the balance towards facilitation or interference. For example, Abdel Rahman and Melinger consider the number of lexical competitors to play a critical role, with more competitors creating more interference (Abdel Rahman & Melinger, 2009). Others have argued that the type of semantic similarity is critical, with taxonomic relations causing considerably greater interference than thematic relations (Oppenheim & Nozari, 2024). Ultimately, all of these proposals converge on the conclusion that if there is room for improving the transmission of activation from semantic features to the target lexical item (e.g., when semantic features are not sufficiently activated, or when the connections between semantic features and lexical items are not functioning properly), semantic similarity is beneficial. If not, it is detrimental. Translating this principle to XR and QD's deficits is straightforward: XR can benefit from a stronger semantic-lexical activation of the target, as she is often unable to retrieve any lexical items in response to a stimulus. QD does not have that problem; he often retrieves too many lexical items.

This is the basis for the prediction that, overall, XR's performance should benefit from semantic similarity and QD's should suffer from it. The overall pattern of the data is compatible with this prediction, although the detrimental effect on QD's performance is clearer and more reliable. In short, we can draw the general conclusion that for IWA, in whom the main problem is activating lexical representation from semantic knowledge, similarity between the training materials is not a major concern. However, other individuals, including neurotypical learners, who do not have this problem, semantic similarity among training items hinders learning and should probably be avoided.

Is semantic similarity during vocabulary training a desirable difficulty?

Previous studies in neurotypical adults learning a new language have answered this question in the negative (Korochkina et al., 2021, 2021; Pérez-Serrano et al., 2022). Our findings generally match this conclusion and extend it to the relearning of lost vocabulary in the first language in IWA. However, some patterns in the results suggest that semantic similarity could be a desirable difficulty in a limited way. For example, after experiencing greater difficulty with the items in the semantic blocks during training, both XR and QD showed better short-term retention of those items at the end of each training session, although only QD's results were statistically significant. This pattern provides some limited support for semantic similarity as a desirable difficulty for vocabulary relearning in the short term.

Interpreting the results of long-term retention in light of this debate is more difficult. If we take during-study accuracy as a measure of interference, then the persistence of such interference for items from the related blocks in long-term retention for QD goes against similarity or dissimilarity as a desirable difficulty; what was difficult to begin with, remained difficult in the long run. However, if we consider short-term retention tests as our difficulty index, then QD's pattern of results suggests that items that were more difficult to retain in short-term intervals (in this case, the unrelated items) were more durable in the long run, which constitutes support for *dissimilarity* being a form of desirable difficulty. In other words, QD might be showing a 2-step pattern of desirable difficulty: first, difficulty during training leads to better short-term retention. Second, difficulty in short-term retention leads to better long-term retention. However, even if taken as supporting some version of the desirable difficulty hypothesis cognitively, the practical implications do not match a simple recommendation of the theory: training items in semantically similar blocks does *not* lead to better long-term retention. Moreover, little in XR's pattern of results provides support for semantic similarity being a desirable difficulty for vocabulary learning. We must, therefore, conclude that we do not find evidence that semantic similarity qualifies as a desirable difficulty for vocabulary acquisition in the long term.

Temporal dynamics of semantic similarity

In this section, we discuss the effects of two variables that we earlier anticipated may modulate the influence of semantic similarity over learning: changes to learning over time and the clustering or spreading of the related items within the training session. Since both variables reflect temporal dynamics, we discuss them under this banner.

One of the most clear-cut findings of the current study is that the effect of semantic similarity on vocabulary learning is not uniform over time. Both participants showed initial interference from semantic similarity during training, which turned into at least some degree of facilitation in shorter intervals. For QD, this transient facilitation switched to long-term interference, whereas for XR it did not. The discussions in the earlier sections offer some insight into these timelines. During training, an item's name is repeated over and over, along with its features, providing ample activation in the semantic-lexical representations of the target, even for XR. However, repetition increases activation of semantically related items, increasing interference. Thus, the immediate effect of semantic similarity is one of interference in both participants, regardless of their deficit. As items build up activation over time with more training, the pattern gradually changes in a steady and predictable way. Activating lexical items remains challenging for XR due to her activation deficit; thus, the extra activation conferred by other items in the related blocks continues to benefit her. In contrast, similar to neurotypical learners, QD can build up sufficient activation of the trained items after a few training sessions, after which, he shows a steady detrimental effect of semantic similarity on his performance as competitors gain in strength.

Aside from the overall trajectory of the semantic similarity effect, the inclusion of the mixed blocks allows us to examine how clustering similar items into units vs. spreading them across different units may affect learning. The overall pattern of the results (Fig. 3) shows that, in the majority of cases, the mixed blocks patterned more closely with the unrelated than the related blocks. Statistical analyses also showed differential learning effects between related and mixed conditions during training, as well as in medium-term and long-term intervals. This pattern implies that the close grouping of the related items is critical, especially for inducing the detrimental effects of semantic similarity on learning. On the one hand, this finding appears at odds with learning accounts of semantic interference according to which, as discussed earlier, the connection strengths between concepts and word forms are changed during word production. According to such accounts, these changes (learning) are supposed to be stable and largely insensitive to the passage of time and intervening items (Oppenheim et al., 2010; Schnur et al., 2006). Empirical data provide some support for this idea. For example, IWA can perseverate on semantically related items despite very long lags (Hepner & Nozari, 2020). On the other hand, there is also evidence that semantic interference dissipates after a certain lag or requires a “reminder” to survive longer lags (Schnur, 2014).

Given the amount of evidence in favor of the learning accounts of semantic interference, we do not believe the current data falsify these accounts. Rather, we take our results to indicate that such models currently lack nuances that are critical for modeling learning in real life. Specifically, research has shown that individuals frequently segment ongoing activity into meaningful events (see Zacks & Swallow, 2007, for a review). Neurally, too, there is much evidence showing that hippocampal neurons show distinct patterns of firing during delay periods of a task, leading to the creation of distinct “episodes” (Gill et al., 2011; Pastalkova et al., 2008). None of the learning models of semantic interference currently take this into account. It is not difficult to imagine why learning episodes would be critical to learning. Without episodes, a measure as continuous as semantic similarity becomes meaningless; *everything* is semantically similar to *something* in a person’s life. The question is how the brain partitions events into episodes; one could imagine an entire training session as one episode, or one could instead imagine a hierarchy of episodes, in which blocks create smaller, distinct episodes nested under the larger episode of the current training session. The present results support the latter possibility.

On the pragmatic side, this finding implies that the detrimental effects of semantic similarity in training studies may arise locally, when related items are tightly grouped together and registered as one learning episode. This problem can be ameliorated by spreading the related items across different blocks and thus folding them into separate learning episodes. On the theoretical side, the current findings call for the integration of learning episodes into learning models of semantic interference, and perhaps more broadly, into models of language production.

Generalization

As customary in aphasia training studies, we included untrained items in this study. However, we had no a priori hypotheses regarding the influence of semantic similarity on untrained items, because the hypotheses of the study are directly about similarity effects *during training*. Some studies have reported improvement in naming untrained items with SFA training (e.g., Wallace & Kimelman, 2013). A modest improvement was observed here, but we hesitate to attribute this effect to the extension of SFA benefits from trained items, because (a) all items in this study were subject to SFA. Therefore, we do not have a non-SFA baseline for comparison. It is possible that this modest improvement may have been caused by other factors. Moreover, the current study employed feature *verification* as opposed to feature *generation*, which is less likely to cause generalization of learning to untrained items (Boyle, 2010). (b) We intentionally selected more distinctive features for our SFA for reasons explained in the Introduction. As such, we do not expect a broad range of activation in other items from the same category. This design feature is also compatible with the absence of a relatedness effect on untrained items.

Limitations and future directions

The obvious limitation of the current study is that it contains only two participants. It is important to note that the majority of IWA with lexical retrieval deficits do not fall squarely within the boundaries of an activation or inhibition deficit; many of them show a mixed profile. In the current study, we deliberately chose two extreme cases to be able to study the issue from a theoretical, rather than an applied, point of view. For this reason, we have not reported “effect size” when discussing findings, as such a measure is primarily relevant when considering broader clinical implications. Instead, our focus was on uncovering any statistically validated differences indicating facilitatory or inhibitory effects of semantic similarity on learning. This goal is better achieved by working with a small but clearly dissociated sample, as opposed to a larger and noisier sample.

However, despite the small sample size, the pattern of results showed effects in the directions predicted by theories of language production and learning. Moreover, using a clean sample allowed us to draw conclusions that can be applied to larger groups of IWA with a combination of deficits. In a nutshell, the current findings predict that if a person has greater difficulty in selection vs. activation of lexical items, semantic similarity between items in a learning block should hurt their long-term retention. Since this general principle does not require IWA to have a pure selection deficit, it can be put to the test in larger samples in future studies. Such studies can then strive to study more clinically relevant factors, such as the magnitude of semantic interference effects.

A similar limitation applies to the training items. Due to limitations in time, attention span, and learning abilities of our participants, we were limited to only 27 items. Our interest in including a mixed condition further limited the number of items to 9 in each condition. Despite this limitation, we opted for the inclusion of three distinct related categories within each participant, which were comparable across the two participants. We also controlled for lexical frequency, length, and morphological complexity of items across conditions to decrease the undue influence of irrelevant factors as much as possible. However, we cannot rule out at least partial influences of all such factors. This is a general challenge for all training studies in IWA, although future studies can easily investigate generalization to different categories and items by using new materials.

Finally, we only measured naming abilities because we were specifically interested in semantic interference in production, and because our IWA did not have pronounced lexical comprehension problems. However, the proposed design lends itself well to measuring semantic interference in comprehension as well. For example, recently, Waller et al. (2025) showed a robust semantic interference in the comprehension of newly acquired vocabulary trained in semantically related vs. unrelated blocks. We predict a similar effect in IWA with deficits in lexical comprehension (as long as semantic knowledge is largely intact), which future studies can put to the test.

Conclusions

The findings of this study have implications for our basic understanding of word learning as well as for clinical practice. In terms of the dynamic processes underlying the reacquisition of vocabulary, we found that semantic similarity interacts with processes of activation and selection in a manner predicted by theories of incremental learning in language production. We further argued that such theories need to integrate a notion of learning episodes to better capture the dynamics of learning. With regard to clinical implications, we found that, similar to neurotypical learners, semantic similarity among items in a training set can hurt the long-term retention of those items, but this effect can be ameliorated by spreading the semantically related items across different training blocks within a session.

ACKNOWLEDGMENTS

[Removed for anonymization]

DECLARATION OF INTEREST

The authors do not have a conflict of interest.

REFERENCES

Abdel Rahman, R., & Melinger, A. (2009). Semantic context effects in language production: A swinging lexical network proposal and a review. *Language and Cognitive Processes*, *24*(5), 713–734. https://doi.org/10.1080/01690960802597250

Alario, F.-X., Segui, J., & Ferrand, L. (2000). Semantic and associative priming in picture naming. *The Quarterly Journal of Experimental Psychology: Section A*, *53*(3), 741–764.

Antonucci, S. M. (2009). Use of semantic feature analysis in group aphasia treatment. *Aphasiology*, *23*(7–8), 854–866.

Antonucci, S. M. (2014). What matters in semantic feature processing for persons with stroke-aphasia: Evidence from an auditory concept–feature verification task. *Aphasiology*, *28*(7), 823–839.

Basso, A. (2005). How intensive/prolonged should an intensive/prolonged treatment be? *Aphasiology*, *19*(10–11), 975–984.

Bates, D. (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.

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

Benigas, J. E., & Bourgeois, M. (2016). Using spaced retrieval with external aids to improve use of compensatory strategies during eating for persons with dementia. *American Journal of Speech-Language Pathology*, *25*(3), 321–334.

Bhogal, S. K., Teasell, R. W., Foley, N. C., & Speechley, M. R. (2003). Rehabilitation of aphasia: More is better. *Topics in Stroke Rehabilitation*, *10*(2), 66–76.

Bjork, E. L., & Bjork, R. A. (2011). *Making Things Hard on Yourself, But in a Good Way: Creating Desirable Difficulties to Enhance Learning*.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. *Metacognition: Knowing about Knowing*, *185*(7.2), 185–205.

Bjork, R. A., & Bjork, E. L. (2020). Desirable difficulties in theory and practice. *Journal of Applied Research in Memory and Cognition*, *9*(4), 475.

Black, J., & Nozari, N. (2023). Precision of phonological errors in aphasia supports resource models of phonological working memory in language production. *Cognitive Neuropsychology*, *40*(1), 1–24. https://doi.org/10.1080/02643294.2023.2206012

Boyle, M. (2004). Semantic feature analysis treatment for anomia in two fluent aphasia syndromes. *American Journal of Speech-Language Pathology*.

Boyle, M. (2010). Semantic feature analysis treatment for aphasic word retrieval impairments: What's in a name? *Topics in Stroke Rehabilitation*, *17*(6), 411–422.

Boyle, M., & Coelho, C. A. (1995). Application of semantic feature analysis as a treatment for aphasic dysnomia. *American Journal of Speech-Language Pathology*, *4*(4), 94–98.

Breining, B., Nozari, N., & Rapp, B. (2016). Does segmental overlap help or hurt? Evidence from blocked cyclic naming in spoken and written production. *Psychonomic Bulletin & Review*, *23*(2), 500–506. https://doi.org/10.3758/s13423-015-0900-x

Breining, B., Nozari, N., & Rapp, B. (2019). Learning in complex, multi-component cognitive systems: Different learning challenges within the same system. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(6), 1093–1106. https://doi.org/10.1037/xlm0000630

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.

Burgess, J., & Nozari, N. (2022). A conflict-based model of speech error repairs in humans. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society*. Cognitive Science Society. (pp. 2140-2146).

Cameron, R. M., Wambaugh, J. L., Wright, S. M., & Nessler, C. L. (2006). Effects of a combined semantic/phonologic cueing treatment on word retrieval in discourse. *Aphasiology*, *20*(02–04), 269–285.

Cherney, L. R., Patterson, J., Raymer, A., Frymark, T., & Schooling, T. (2010). Updated evidence-based systematic review: Effects of intensity of treatment and constraint-induced language therapy for individuals with stroke-induced aphasia. *Rockville, MD: American Speech-Language-Hearing Association*.

Clare, L., & Jones, R. S. P. (2008). Errorless Learning in the Rehabilitation of Memory Impairment: A Critical Review. *Neuropsychology Review*, *18*(1), 1–23. https://doi.org/10.1007/s11065-008-9051-4

Coelho, C. A., McHugh, R. E., & Boyle, M. (2000). Semantic feature analysis as a treatment for aphasic dysnomia: A replication. *Aphasiology*, *14*(2), 133–142.

Conroy, P., Sage, K., & Lambon Ralph, M. A. (2009). The effects of decreasing and increasing cue therapy on improving naming speed and accuracy for verbs and nouns in aphasia. *Aphasiology*, *23*(6), 707–730.

Costa, A., Alario, F. X., & Caramazza, A. (2005). On the categorical nature of the semantic interference effect in the picture-word interference paradigm. *Psychonomic bulletin & review*, *12*(1), 125-131.

Costa, A., Strijkers, K., Martin, C., & Thierry, G. (2009). The time course of word retrieval revealed by event-related brain potentials during overt speech. *Proceedings of the National Academy of Sciences*, *106*(50), 21442–21446. https://doi.org/10.1073/pnas.0908921106

Coultrip, R., Granger, R. & Lynch, G. A cortical model of winner-take-all competition via lateral inhibition. *Neural Networks* **5**, 47–54 (1992)

Markram, H. *et al.* Interneurons of the neocortical inhibitory system. *Nat Rev Neurosci* **5**, 793–807 (2004).

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), 801–838.

Dell, G. S., Schwartz, M. F., Nozari, N., Faseyitan, O., & Branch Coslett, H. (2013). Voxel-based lesion-parameter mapping: Identifying the neural correlates of a computational model of word production. *Cognition*, *128*(3), 380–396.

de Zubicaray, G. I., Hansen, S. & McMahon, K. L. (2013) Differential processing of thematic and categorical conceptual relations in spoken word production. *Journal of Experimental Psychology: General* 142, 131–142.

Efstratiadou, E. A., Papathanasiou, I., Holland, R., Archonti, A., & Hilari, K. (2018). A systematic review of semantic feature analysis therapy studies for aphasia. *Journal of Speech, Language, and Hearing Research*, *61*(5), 1261–1278.

Falconer, C., & Antonucci, S. M. (2012). Use of semantic feature analysis in group discourse treatment for aphasia: Extension and expansion. *Aphasiology*, *26*(1), 64–82.

Fillingham, J., Sage, K., & Ralph, M. L. (2005). 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.

Fink, R., Martin, N., & Berkowitz, R. (2009). *Interference and facilitation effects of semantic and phonological contextual priming: A treatment case study* [Clinical Aphasiology Paper].

Freedman, M. L., & Martin, R. C. (2001). Dissociable components of short-term memory and their relation to long-term learning. *Cognitive Neuropsychology*, *18*(3), 193–226.

Freund, M. & Nozari, N. (2018). Is adaptive control in language production mediated by learning? *Cognition, 176*, 107–130.

Fridriksson, J., Holland, A., Beeson, P., & Morrow, L. (2005). Spaced retrieval treatment of anomia. *Aphasiology*, *19*(2), 99–109.

Gill, P. R., Mizumori, S. J. Y., & Smith, D. M. (2011). Hippocampal episode fields develop with learning. *Hippocampus*, *21*(11), 1240–1249.

Goodglass, H., & Wingfield, A. (1997). Word-finding deficits in aphasia: Brain–behavior relations and clinical symptomatology. In H. Goodglass & A. Wingfield (Eds.), *Anomia: Neuroanatomical and cognitive correlates* (pp. 3–27). New York: Academic Press.

Harnish, S. M., Morgan, J., Lundine, J. P., Bauer, A., Singletary, F., Benjamin, M. L., Gonzalez Rothi, L. J., & Crosson, B. (2014). Dosing of a cued picture-naming treatment for anomia. *American Journal of Speech-Language Pathology*, *23*(2), S285–S299.

Hepner, C. R., & Nozari, N. (2020). The dual origin of lexical perseverations in aphasia: Residual activation and incremental learning. *Neuropsychologia*, *147*, 1-14.

Hillis, A. E. (1998). Treatment of naming disorders: New issues regarding old therapies. *Journal of the International Neuropsychological Society*, *4*(6), 648–660.

Hinckley, J., & Carr, T. (2005). Comparing the outcomes of intensive and non-intensive context-based aphasia treatment. *Aphasiology*, *19*(10–11), 965–974.

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.

Karpicke, J. D., & Aue, W. R. (2015). The Testing Effect Is Alive and Well with Complex Materials. *Educational Psychology Review*, *27*(2), 317–326.

Karpicke, J. D., & Roediger, H. L. (2007a). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33, 704 –719.

Kiran, S., & Bassetto, G. (2008). Evaluating the Effectiveness of Semantic-Based Treatment for Naming Deficits in Aphasia: What Works? *Seminars in Speech and Language*, *29*(1), 71–82.

Kiran, S., & Roberts, P. M. (2010). Semantic feature analysis treatment in Spanish–English and French–English bilingual aphasia. *Aphasiology*, *24*(2), 231–261.

Korochkina, M., Bürki, A., & Nickels, L. (2021). Apples and oranges: How does learning context affect novel word learning? *Journal of Memory and Language*, *120*, 104246.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of statistical software, 82,* 1-26.

Lambon Ralph, M. A., Sage, K., & Roberts, J. (2000). Classical anomia: A neuropsychological perspective on speech production. *Neuropsychologia*, *38*(2), 186–202.

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

Loverso, F. L., Prescott, T. E., & Selinger, M. (1988). Cueing verbs: A treatment strategy for aphasic adults (CVT). *Journal of Rehabilitation Research and Development*, *25*(2), 47–60.

Lowell, S., Beeson, P. M., & Holland, A. L. (1995). The efficacy of a semantic cueing procedure on naming performance of adults with aphasia. *American Journal of Speech-Language Pathology*, *4*(4), 109–114.

Lu, C., & Proctor, R. W. (1995). The influence of irrelevant location information on performance: A review of the Simon and spatial Stroop effects. *Psychonomic Bulletin & Review*, *2*(2), 174–207.

Macoir, J., Leroy, M., Routhier, S., Auclair-Ouellet, N., Houde, M., & Laforce Jr, R. (2015). Improving verb anomia in the semantic variant of primary progressive aphasia: The effectiveness of a semantic-phonological cueing treatment. *Neurocase*, *21*(4), 448–456.

Mahendra, N. (2008). Fixed-interval and randomized-interval spaced retrieval training seem to have comparable outcomes for improving confrontation naming performance in three persons with aphasia, but the jury is still out. *Evidence-Based Communication Assessment and Intervention*, *2*(1), 6–10.

Mahon, B. Z., Costa, A., Peterson, R., Vargas, K. A., & Caramazza, A. (2007). Lexical selection is not by competition: A reinterpretation of semantic interference and facilitation effects in the picture-word interference paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 503–535.

Marshall, R. C., Neuburger, S. I., & Phillips, D. S. (1992). Effects of facilitation and cueing on labelling of «novel» stimuli by aphasic subjects. *Aphasiology*, *6*(6), 567–583.

Martin, N., Fink, R., Laine, M., & Ayala, J. (2004). Immediate and short-term effects of contextual priming on word retrieval in aphasia. *Aphasiology*, *18*(10), 867–898.

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

Meteyard, L., & Bose, A. (2018). What Does a Cue Do? Comparing Phonological and Semantic Cues for Picture Naming in Aphasia. *Journal of Speech, Language, and Hearing Research*, *61*(3), 658–674.

Middleton, E. L., & Schwartz, M. F. (2012). Errorless learning in cognitive rehabilitation: A critical review. *Neuropsychological Rehabilitation*, *22*(2), 138–168.

Morrow, K. L., & Fridriksson, J. (2006). Comparing fixed- and randomized-interval spaced retrieval in anomia treatment. *Journal of Communication Disorders*, *39*(1), 2–11.

Neumann, Y. (2018). A case series comparison of semantically focused vs. Phonologically focused cued naming treatment in aphasia. *Clinical Linguistics & Phonetics*, *32*(1), 1–27.

Nickels, L. (2002). Improving word finding: Practice makes (closer to) perfect? *Aphasiology*, *16*(10–11), 1047–1060.

Nickels, L., Kohnen, S., & Biedermann, B. (2010). An untapped resource: Treatment as a tool for revealing the nature of cognitive processes. *Cognitive Neuropsychology*, *27*(7), 539–562.

Nozari, N. (2019). The dual origin of semantic errors in access deficit: Activation vs. inhibition deficit. *Cognitive Neuropsychology*, *36*(1–2), 31–53.

Nozari, N. (2022). Neural basis of word production. In L. R. Gleitman, A. Papafragou, & J. C. Trueswell (Eds.), *The Oxford Handbook of The Mental Lexicon* (pp. 552-574). Oxford: Oxford University Press.

Nozari, N. (2025). Monitoring, control and repair in word production. *Nature Reviews Psychology,* 4, 222–238.

Nozari, N., & Dell, G. S. (2013). How damaged brains repeat words: A computational approach. *Brain and Language*, *126*(3), 327–337.

Nozari, N., Freund, M., Breining, B., Rapp, B., & Gordon, B. (2016). Cognitive control during selection and repair in word production. *Language, Cognition and Neuroscience*, *31*(7), 886–903.

Nozari, N., & Hepner, C. (2019a). To select or to wait? Response to the commentaries. *Cognitive Neuropsychology, 36*(5-6), 226-233.

Nozari, N. & Hepner, C. (2019b). To select or to wait? The importance of criterion setting in debates of competitive lexical selection. *Cognitive Neuropsychology, 36*(5-6), 193-207.

Nozari, N., Kittredge, A. K., Dell, G. S., & Schwartz, M. F. (2010). Naming and repetition in aphasia: Steps, routes, and frequency effects. *Journal of Memory and Language*, *63*(4), 541–559.

Nozari, N., & Martin, R. C. (2024). Is working memory domain-general or domain-specific?. *Trends in cognitive sciences*, *28*(11), 1023-1036.

Nozari, N., & Novick, J. (2017). Monitoring and control in language production. *Current Directions in Psychological Science*, *26*(5), 403-410.

Nozari, N., & Pinet, S. (2020). A critical review of the behavioral, neuroimaging, and electrophysiological studies of co-activation of representations during word production. *Journal of Neurolinguistics*, *53*, 1–19.

Off, C. A., Griffin, J. R., Spencer, K. A., & Rogers, M. A. (2016). The impact of dose on naming accuracy with persons with aphasia. *Aphasiology*, *30*(9), 983–1011.

Oppenheim, G. M. (2018). The paca that roared: Immediate cumulative semantic interference among newly acquired words. *Cognition*, *177*, 21–29.

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.

Oppenheim, G. M., & Nozari, N. (2024). Similarity-induced interference or facilitation in language production reflects representation, not selection. *Cognition*, *245*, 105720. https://doi.org/10.1016/j.cognition.2024.105720

Pastalkova, E., Itskov, V., Amarasingham, A., & Buzsáki, G. (2008). Internally Generated Cell Assembly Sequences in the Rat Hippocampus. *Science*, *321*(5894), 1322–1327.

Peach, R. K., & Reuter, K. A. (2010). A discourse-based approach to semantic feature analysis for the treatment of aphasic word retrieval failures. *Aphasiology*, *24*(9), 971–990.

Pérez-Serrano, M., Nogueroles-López, M., & Dunabeitia, J. A. (2022). Effects of semantic clustering and repetition on incidental vocabulary learning. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.997951

Piai, V., Riès, S. K. & Swick, D. (2016). Lesions to Lateral Prefrontal Cortex Impair Lexical Interference Control in Word Production. *Frontiers in Human Neuroscience,* 9. https://doi.org/10.3389/fnhum.2013.00832

Pinet, S., & Nozari, N. (2023). Different Electrophysiological Signatures of Similarity-induced and Stroop-like Interference in Language Production. *Journal of Cognitive Neuroscience*, *35*(8), 1329–1349.

Rapp, B. & Goldrick, M. (2000). Discreteness and interactivity in spoken word production. *Psychological Review, 107,* 460-499.

Rapp, B. & Wiley, R. (2019). Re-learning and remembering in the lesioned brain. *Neuropsychologia,* 132, 107126.

Renvall, K., Laine, M., & Martin, N. (2005). Contextual priming in semantic anomia: A case study. *Brain and Language*, *95*(2), 327–341.

Renvall, K., Laine, M., & Martin, N. (2007). Treatment of anomia with contextual priming: Exploration of a modified procedure with additional semantic and phonological tasks. *Aphasiology*, *21*(5), 499–527.

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.

Roelofs, A. (1992). A spreading activation theory of lemma retrieval in speaking. *Cognition, 42(1–3)*, 107–142.

Roelofs, A. (2018). A unified computational account of cumulative semantic, semantic blocking, and semantic distractor effects in picture naming. *Cognition*, *172*, 59-72.

Rogalski, Y., Altmann, L. J., & Rosenbek, J. C. (2014). Retrieval practice and testing improve memory in older adults. *Aphasiology*, *28*(4), 381–400.

Saxton, M. (1997). The Contrast Theory of negative input. *Journal of Child Language*, *24*(1), 139–161.

Saxton, M., Kulcsar, B., Marshall, G., & Rupra, M. (1998). Longer-term effects of corrective input: An experimental approach. *Journal of Child Language*, *25*(3), 701–721.

Schnur, T. T. (2014). The persistence of cumulative semantic interference during naming. *Journal of Memory and Language*, *75*, 27–44.

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.

Schnur, T. T., Schwartz, M. F., Kimberg, D. Y., Hirshorn, E., Coslett, H. B., & Thompson-Schill, S. L. (2009). Localizing interference during naming: Convergent neuroimaging and neuropsychological evidence for the function of Broca's area. *Proceedings of the National Academy of Sciences*, *106*(1), 322–327.

Schriefers, H., Meyer, A. S., & Levelt, W. J. M. (1990). Exploring the time course of lexical access in language production: Picture-word interference studies. *Journal of Memory and Language*, *29*(1), 86–102.

Shekari, E., & Nozari, N. (2023). A narrative review of the anatomy and function of the white matter tracts in language production and comprehension. *Frontiers in Human Neuroscience*, *17*. https://doi.org/10.3389/fnhum.2023.1139292

van Hees, S., Angwin, A., McMahon, K., & Copland, D. (2013). A comparison of semantic feature analysis and phonological components analysis for the treatment of naming impairments in aphasia. *Neuropsychological Rehabilitation*, *23*(1), 102–132.

Wallace, S. E., & Kimelman, M. D. Z. (2013). Generalization of word retrieval following semantic feature treatment. *NeuroRehabilitation*, *32*(4), 899–913.

Waller, M., Yurovsky, D., & Nozari, N. (2024). Of Mouses and Mans: A Test of Errorless Versus Error-Based Learning in Children. *Cognitive Science*, *48*(11), e70006. https://doi.org/10.1111/cogs.70006

Waller, M., Yurovsky, D., & Nozari, N. (2025). Production training and contextual similarity hurt the comprehension of new vocabulary. *Language, Cognition and Neuroscience*, *40*(3), 291–314.

Wambaugh, J. L., Doyle, P. J., Martinez, A. L., & Kalinyak-Fliszar, M. (2002). Effects of two lexical retrieval cueing treatments on action naming in aphasia. *Journal of Rehabilitation Research and Development*, *39*(4), 455–466.

Wessel, J. R., & Anderson, M. C. (2024). Neural mechanisms of domain-general inhibitory control. *Trends in Cognitive Sciences*, *28*(2), 124-143.

Wheeldon, L. R., & Monsell, S. (1994). Inhibition of Spoken Word Production by Priming a Semantic Competitor. *Journal of Memory and Language*, *33*(3), 332–356.

Wisenburn, B., & Mahoney, K. (2009). A meta-analysis of word-finding treatments for aphasia. *Aphasiology*, *23*(11), 1338–1352.

Zacks, J. M., & Swallow, K. M. (2007). Event Segmentation. *Current Directions in Psychological Science*, *16*(2), 80–84.

Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition*, *38*(8), 995–1008.

Zhang, S., Kang, K., Zhong, J., Yu, H., Ip, C. T., & Zhang, H. (2025). Cognitive control states and traits modulate lexical competition during word production. *Scientific Reports*, *15*(1), 25381.

Zhu, D., Wang, X., Zhao, E., Nozari, N., Notebaert, W., & Braem, S. (2025). Cognitive control is task specific: Further evidence against the idea of domain-general conflict adaptation. Journal of Experimental Psychology: Learning, Memory, and Cognition. Advance online publication. https://dx.doi.org/10.1037/ xlm0001480