# Effects of distributed practice and criterion level on word retrieval in aphasia

**Julia Schuchard**[1], **Katherine A. Rawson**[2], **Erica L. Middleton**[1]

[1]Moss Rehabilitation Research Institute

[2]Kent State University

## Abstract

This study examined how the distribution and amount of practice affect word retrieval in aphasia as well as how such factors relate to the efficiency of learning. The central hypothesis was that factors that enhance the learning of new knowledge also enhance persistent access to existing, but inconsistently available, word representations. The study evaluated the impact of learning principles on word retrieval by manipulating the timing and amount of retrievals for items presented for naming. Nine people with chronic aphasia with naming impairment completed the experiment. Training materials involved proper noun entities assigned to six conditions formed by crossing a 2-level factor of spacing of sessions, i.e., *intersession interval* (1 day versus 7 days between sessions) with a 3-level factor of number of correct retrievals per item per session, i.e., *criterion level* (Criterion-1, Criterion-2, and Criterion-4). Each intersession interval condition comprised three training sessions and a one-month retention test. Increasing the criterion level enhanced naming performance after short (1 day, 7 days) and long (one month) retention intervals, but these advantages came at the cost of many additional training trials. In most cases, later naming success was superior when the same number of correct retrievals of an item was distributed across multiple sessions rather than administered within one session. The substantial advantages for across-session spacing were gained at little cost in terms of additional training trials. At one-month retention, naming accuracy was numerically but not significantly higher in the

7-day versus 1-day intersession interval condition. Implications for theories of lexical access and naming treatment in aphasia are discussed.

## 1. Introduction

Understanding the effects of different *schedules of practice* (i.e., the amount of practice and how it is distributed) is crucial for maximizing learning outcomes. Though experience tells us practice makes perfect, real-world constraints on one's time and effort make consideration of the efficiency of learning (i.e., the cost versus benefit from additional practice) paramount. One way to enhance learning efficiency is to control *when* practice occurs. The cognitive and educational psychology literatures have produced hundreds of studies demonstrating that knowledge acquisition and its retention are profoundly affected by how practice is distributed. The benefits of *distributed practice* are induced by spacing practice events over time compared to massing them in close succession or by increasing the *lag* (i.e., the interval between spaced practice trials for specific items or between learning sessions; for reviews, see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Delaney, Verkoeijen, & Spirgel, 2010; Toppino & Gerbier, 2014). However, only a small number of distributed practice studies in the verbal domain have included multiple training sessions or assessed retention of learning after intervals longer than 24 hours after training (Cepeda et al., 2006). The short timeframe of these studies differs from most real-world learning situations, including the focus of the current paper: language treatment for adults with aphasia, an acquired language disorder caused by stroke or other forms of brain damage. Unlike learning studies in psychology, aphasia treatment studies often involve multiple weeks of training and follow-up testing, but they rarely include well-controlled tests of different schedules of practice (Cherney, 2012; Middleton, Schuchard, & Rawson, 2019; Warren, Fey, & Yoder, 2007). The present study provides groundwork for bridging these two disparate literatures by examining effects of the distribution of naming practice in the context of multi-session training for people with aphasia. In addition to the timing of practice, as described further below, we also examine how the amount of practice defined in terms of the level of mastery attained within and across sessions affects learning and efficiency.

Given the ubiquity of naming impairment in aphasia (Goodglass & Kaplan, 1983), it is common in speech-language treatment of aphasia for the clinician to provide naming treatment. This often involves presenting a set of items to the patient for multiple trials of naming practice per item. The clinician's goal may be to improve word retrieval for a vocabulary set that is personally relevant to the patient, or to increase the number of entities they can fluently name more generally. In this form of treatment, the clinician must make additional decisions such as how many practice trials to dedicate to each item in a session and how to distribute an item's trials across sessions. As reviewed in the following section,

psychological research suggests such decisions can have a dramatic impact on the efficiency and retention of learning.

## 1.1 Criterion Learning and Distributed Practice Effects

The present study examined how the distribution and amount of practice affect long-term naming performance in aphasia as well as how such factors relate to learning efficiency. The type of naming training involved *retrieval practice*, or practice retrieving the names for depicted entities from long-term memory. This type of practice was chosen because a wealth of psychological research has demonstrated that the act of retrieving information from long-term memory persistently strengthens its future retrievability (for recent reviews see Kornell & Vaughn, 2016; Rawson & Dunlosky, 2011; Roediger & Butler, 2011; Roediger, Putnam & Smith, 2011; Rowland, 2014). The manipulation of amount of practice involved varying the *criterion level* (i.e., experimentally-determined level of mastery obtained per item in a session). An item's assigned criterion level in this study was the number of trials in which the participant successfully named the depicted entity before the item was dropped from further practice in that session. Additionally, items were practiced to their assigned criterion level in each of multiple sessions occurring on different days. This method of training individual items to mastery in each of multiple sessions is termed *successive relearning*, which has been shown to promote robust and durable learning in neurotypical populations (Bahrick, 1979; Bahrick, Bahrick, Bahrick, & Bahrick, 1993; Bahrick & Hall, 2005; Rawson & Dunlosky, 2011, 2013; Rawson, Dunlosky, & Sciartelli, 2013; Rawson, Vaughn, Walsh, & Dunlosky, 2018; Vaughn, Dunlosky, & Rawson, 2016).

Using this successive relearning design, we focused on four key effects of interest observed in psychological research. Two effects pertain to the manipulation of criterion level, and two pertain to distributed practice. First, increasing the criterion level of items during an initial practice session typically enhances learning, as measured by retrieval practice success at the first trial of each item in the next session. However, additional increases in criterion level beyond the first few correct retrievals in the initial session confer diminishing returns to initial retrievability at the next session, which can make learning inefficient (Pyc & Rawson, 2009; Rawson & Dunlosky, 2011; Rawson et al, 2018; Vaughn & Rawson, 2011, 2014; Vaughn et al., 2016). For example, Rawson and Dunlosky (2011) observed no benefit of requiring four correct trials per item compared to three, and Vaughn et al. (2016) found little benefit of requiring seven correct trials per item compared to three.

Second, a number of studies have found that the benefits of increasing criterion level in an initial learning session are increasingly attenuated with each additional relearning session (e.g., Rawson & Dunlosky, 2013; Rawson et al., 2018; Vaughn et al., 2016). Although a higher criterion level in session 1 results in better initial performance at session 2, the effect is weaker on performance at session 3, even weaker on performance at session 4, and so on. A possible implication is that at a final test following multiple successive relearning sessions, effects of criterion level may be weak or absent despite great cost in terms of the number of training trials administered to achieve higher criterion levels.

Third, when criterion level is controlled, distributing practice for items across multiple sessions confers more robust and efficient learning compared to focusing practice for items

within a single session (Rawson & Dunlosky, 2011, 2013; Rawson et al. 2018; Vaughn et al., 2016). In other words, increasing the lag between successive practice events for an item from several minutes *within* a session to at least one day *between* sessions enhances the durability and efficiency of learning.

Fourth, the lag between successive learning sessions, or the *intersession interval*, can affect the durability of learning. The vast majority of studies of lag effects have examined lags on the order of minutes or hours rather than days or weeks. The few studies in psychology that have manipulated lags on the order of days provide some evidence for a benefit of longer intersession intervals on learning measured at multiple weeks or months post-training (Bahrick et al., 1993; Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008; Cepeda et al., 2009). A recent study of successive relearning, however, showed little difference between two-day and seven-day intersession intervals on performance one month after training (Rawson et al., 2018). Further research is needed to establish effects of intersession interval on long-term learning.

## 1.2 Learning Effects in Naming Rehabilitation

The central tenet of the present study is that factors that enhance the learning of new knowledge can improve the accessibility of language-based representations. The focus of the present study is *lexical access*, specifically the process of retrieving known words and their constituent speech sounds for oral production. For people with aphasia, impaired lexical access is a meaningful context to examine because it is a prevalent cause of difficulty in naming familiar people, objects, actions, places, etc. (Dell, Schwartz, Martin, Saffran, & Gagnon, 1997; Rapp & Goldrick, 2000; Walker & Hickok, 2016). Treatment for this type of impairment commonly involves discrete training trials in which an individual practices retrieving the names of depicted entities from memory. The scheduling of the practice trials for specific items can be experimentally manipulated, just as practice trials for novel, discrete units of information are manipulated in basic learning paradigms.

To investigate the impact of criterion level and distributed practice variables on lexical access, we adopted two strategies in the current study that feature in our prior work (Middleton, Schwartz, Rawson, Traut, & Verkuilen, 2016; Middleton, Rawson, & Verkuilen, 2019). First, we recruited people with aphasia for whom lexical access issues contributed to their naming deficit, as indicated by neuropsychological testing (see section 2.1). Second, the study used proper noun entities because of their propensity to elicit tip-of-the tongue states in neurotypical older adults (e.g., Burke, Mackay, Worthley, & Wade, 1991) as well as word retrieval failures in individuals with even mild aphasia (e.g., Middleton et al., 2016). To confirm prior familiarity with the proper noun entities selected for training, each participant underwent extensive testing to identify a set of items for that participant in which the entity and its name were known, but the participant experienced difficulty naming (see section 2.3.1).

A nascent but growing literature has examined the influence of learning factors such as retrieval practice and distributed practice on naming impairment in people with aphasia. In the psychological literature, learning from retrieval practice is typically demonstrated by observations that performance on a retention test is greater following training that involves

the retrieval of target information from long-term memory versus training that only involves restudy of target information that is presented in its entirety. Several recent studies have adapted this paradigm to examine the benefit from retrieval practice in the context of language treatment in aphasia. The observation across studies is that practice retrieving names for depicted entities from long-term memory (i.e., retrieval practice naming treatment) confers more durable improvements in naming performance compared to being provided the correct name for the picture and repeating it aloud (Friedman, Sullivan, Snider, Luta, & Jones, 2017; Middleton, Schwartz, Rawson, & Garvey, 2015; Middleton et al., 2016; Middleton et al., 2019).

Concerning distributed practice effects, a ubiquitous finding in the psychological literature is that learning is superior when repeated training trials for a specific item are spaced over time, rather than being presented in close or contiguous succession (i.e., the *spacing effect;* see Cepeda et al., 2006; Delaney et al., 2010; Toppino & Gerbier, 2014). Middleton et al. (2016) provided the first demonstration of spacing effects in a naming treatment context in aphasia. In that study, an item's training trials were separated either by multiple trials for other items (spaced schedule) or by only one intervening trial (massed schedule). On naming tests administered one day and one week following training, performance was superior for items trained in the spaced schedule compared to those trained in the massed schedule. In another study, Middleton et al. (2019) compared items that were trained by either spaced or massed trials in each of multiple sessions. Spaced training promoted superior naming performance on a naming test administered after one week, with a marginal advantage for spacing over massing at a one-month test.

Overall, the literature on retrieval practice and distributed practice effects in aphasia has demonstrated that learning principles that affect the acquisition of knowledge also enhance naming treatment outcomes in aphasia. However, none of the studies to date have manipulated criterion level nor examined the efficiency and durability of learning when training of items is focused within versus across sessions, as in the present study. Only a small number of studies have examined the impact of intersession interval on the efficacy of treatment in aphasia while controlling for other relevant factors (Dignam et al., 2015; 2016; Martins et al., 2013; Ramsberger & Marie, 2007; Raymer, Kohen, & Saffell, 2006; Sage, Snell, & Lambon Ralph, 2011). The findings of this study will be discussed in relation to these prior studies in the final sections of the paper.

## 1.3 Theoretical Frameworks

Current models of lexical access that include mechanisms for learning suggest that each instance of word retrieval induces incremental, persistent changes to the strength of connections between meaning and words, facilitating subsequent access for the target of the retrieval attempt (Howard, Nickels, Coltheart, & Cole-Virtue, 2006; Oppenheim, Dell, & Schwartz, 2010). By testing the effects of retrieval practice in people with aphasia, Schuchard and Middleton (2018a; 2018b) provided evidence that retrieval practice is particularly beneficial for strengthening these meaning-to-word connections, as compared to practice that involves word repetition without word retrieval attempts. To date, however, incremental learning models have not addressed spacing or lag effects, and these models

have focused on effects that persist over intervals of seconds or minutes, rather than long-term learning. The present work is an important first step in evaluating the effects of different schedules of retrieval practice on word retrieval, including retention intervals of multiple days or weeks. This study does not include experimental contrasts to test the specific model and level(s) of lexical access relevant to distributed practice effects. Nevertheless, the results are expected to motivate and inform this work.

Despite the extensive body of research on human memory and learning, the cognitive processes that give rise to distributed practice effects are poorly understood. However, the *relearning attenuates decay* (RAD) model has been successful in accounting for effects observed in studies of successive relearning (Rawson et al., 2018). Grounded in ACT-R's (Adaptive Control of Thought-Rational) theory of declarative memory (Anderson, 2007), the RAD model represents the strength of an item in memory as an activation value. The activation of an item incrementally increases with each practice event for it, which in turn increases the probability of successful future performance for that item. Intuitively, then, more practice is better. However, a practice event that occurs when the item has a lower activation value leads to a slower rate of forgetting. Hence, activation for an item increases after it is initially practiced, and subsequent instances of that item within the same training session induce smaller benefits for retention than an instance in a later session when the item's activation is lower. Although the RAD model differs in significant ways from current models of lexical access, it may be useful for informing the integration of lag effects into theories of lexical learning. Bringing these frameworks together would in turn have implications for scheduling practice for people with aphasia.

## 1.4 Current Design and Research Questions

In the present study, for each of nine participants with aphasia, we assigned separate sets of items (i.e., pictures and their corresponding names) to six conditions formed by crossing a three-level factor of criterion level (Criterion-1, Criterion-2, and Criterion-4) with a two-level factor of intersession interval (1-day versus 7-day). See Figure 1 for a depiction of the study design. In each intersession interval condition, items were trained to their assigned criterion level in each of three sessions. In the 1-day intersession interval condition, the three training sessions occurred on consecutive days. In the 7-day intersession interval condition, the three training sessions occurred on the same day (e.g., Monday) in three consecutive weeks. Naming performance on the items in each intersession interval condition was assessed in a final test administered one month after the final training session (hereafter, *Long-term Retention Test*).

In the present design, we addressed four specific research questions, examining each of the four key effects of interest discussed in section 1.1. First, what is the impact of increasing criterion level in an initial training session on subsequent naming performance in aphasia? By measuring naming performance on the first retrieval practice trial per item at Session 2, we assessed effects of increasing criterion level in the initial session at retention intervals of one day (1-day intersession interval condition) and one week (7-day intersession interval condition). Second, what is the impact of training items to their assigned criterion level in each of multiple sessions on subsequent naming performance in aphasia? Prior work has

shown that the benefit from increasing criterion level in the initial learning session is progressively attenuated with each additional relearning session (e.g., Rawson & Dunlosky, 2013; Rawson et al., 2018; Vaughn et al., 2016). However, these studies only manipulated criterion level in the first session. In contrast, items in this study were trained to their assigned criterion level in all training sessions. We assessed effects of criterion level by comparing naming performance at the Long-Term Retention Test for items trained in the three criterion level conditions.

Third, controlling for criterion level, what is the impact of distributing practice for items across sessions versus within a session on later naming performance in aphasia? To address this question, we compared naming success at key points in the design at which two different item sets had accumulated the same total number of correct responses but differed with regard to whether those responses occurred within one prior training session or across two prior training sessions (see Figure 1). Fourth, what is the impact of distributing practice sessions over shorter versus longer intersession intervals on long-term naming performance in aphasia? To address this, we compared performance at the Long-term Retention Test for items trained in the two intersession interval conditions.

For each learning factor, the primary outcome of interest was naming accuracy. However, we also considered the efficiency of practice by examining the relative benefit of the learning conditions with respect to added cost in terms of number of training trials.

## 2. Method

### 2.1 Participants

Nine native English speakers (six female) completed the protocol. All participants were diagnosed with chronic aphasia from unilateral left-hemisphere stroke with the exception of participant P2. For that participant, clinical imaging showed a small right frontal lobe infarct in addition to a large thrombosis in the left hemisphere. Demographic and language test battery information for the participant sample is provided in Table 1. Table 1 also reports cutoff scores for clinically significant impairment (when available) or in lieu of cutoff scores, average test performance collected on 20 neurotypical controls in a normative study. All participants had completed at least 12 years of education (*M*=15 years) and had no indications of developmental disorder, neurological diagnosis other than stroke, or premorbid major psychiatric illness. Eight of the nine participants were classified as the anomic subtype of aphasia by the Western Aphasia Battery (WAB; Kertesz, 1982). One participant (P7) was classified as the transcortical motor subtype but showed a similar language profile compared to the other participants.

Because the present study focused on word retrieval, we selected individuals from a larger pool of well-characterized, potentially available people with aphasia. Recruitment priority was given to those whose battery test scores pointed to a likely ability to complete the training task. For example, the group exhibited good word repetition ability and generally mild or no apraxia of speech (AOS), which was anticipated to reduce the likelihood of failure to achieve the designated training criterion due to difficulties pronouncing words. Test battery scores were also used to recruit individuals with oral naming impairment due to

lexical access difficulty rather than deficits in processes peripheral to lexical access, although these cannot be completely ruled out. Generally good word repetition and minimal AOS in the sample indicated little impairment in post-lexical stages of production subsequent to word form retrieval (e.g., phoneme buffering, syllabification, or articulation; for discussion, see Goldrick & Rapp, 2007). Nonword repetition was markedly poor for some participants, which could indicate at least some impairment in processes other than lexical access, such as phonological input processing, post-lexical production, or short-term memory deficits. Verbal short-term memory deficits were present in our sample, and performance on this measure appeared to track nonword performance (see STM span in Table 1; Martin, Shelton, & Yaffee, 1994). Relatively good scores on the auditory comprehension subtest of the WAB and on a test of nonverbal semantic comprehension suggest that the participants did not suffer primarily from a central semantic deficit that can impair semantic input to word retrieval. Overall, testing was consistent with a sample in which naming impairment was largely, but not exclusively, attributable to lexical access deficit. Participants gave informed consent under a protocol approved by the Institutional Review Board of Einstein Healthcare Network and were paid at a rate of $15 per hour.

## 2.2 Materials and Design

A 600-item picture corpus of proper noun entities was used in the current study. The 600 pictures were a subset of a 700-item proper noun corpus collected from internet sources that was developed in prior studies (for additional details of corpus development, see Middleton et al., 2016). We selected 600 of the original 700 items that were no more than 5 syllables long to reduce the likelihood that articulatory difficulty would prevent items from reaching criterion. An additional 35 non-experimental pictures were collected to be used as practice items and as fillers during the item selection phase (see section 2.3.1). The pictures depicted famous people (e.g., actors, politicians, historical figures), fictional characters (e.g., Superman), and movie posters (e.g., Forrest Gump) that were generally familiar to people in the age range of our participants. Between 7–10 neurotypical older adults named each of the 600 experimental items to identify alternative pronunciations or lexical variations (e.g., Jimmy/James Stewart) that would be considered correct if produced in the experiment. Audio recordings of the proper nouns were created by a female native English speaker and normalized in volume. All stimuli were presented on a PC computer using E-Prime software.

This study used a within-subject design in which each participant received each of the six conditions formed by crossing a three-level factor of criterion level (Criterion-1, Criterion-2, and Criterion-4) with a two-level factor of intersession interval (1-day versus 7-day).

## 2.3 Procedure

The study consisted of the following three phases: the item selection phase, the 1-day intersession interval (ISI) training phase, and the 7-day ISI training phase. Each participant began the study with the item selection phase, in which a personalized set of familiar but difficult-to-name items was selected and divided among six training conditions. The order of the two training phases was counterbalanced across participants. Each training phase consisted of three training sessions and a long-term retention test (see Figure 1).

**2.3.1    Item Selection Phase**—Item selection procedures were designed to identify items for a participant for which they confirmed knowledge of the entity and its name but experienced difficulty naming the item. Item selection began with one or two initial sessions dedicated to administration of a multiple-choice picture-name matching task. For each picture in the 600-item corpus, participants were instructed to select the correct name from among five written options. Three foils were the names for similar entities (e.g., for Cameron Diaz, foils were Claire Danes, Charlize Theron, and Cate Blanchett) and the fourth foil was a "none of the above" option. The task included 30 filler items for which the correct answer was "none of the above" so that participants would not ignore this option. The experimenter read the options aloud unless she was confident that the participant could accurately read them.

Only items that were answered correctly in the picture-name matching task were included in the subsequent confrontation naming and familiarity rating tasks. These items were presented in a random order two times over multiple sessions, with the two administrations occurring in separate weeks. On a naming trial, a picture was shown, and the participant was asked to try to produce the full name for the proper noun entity. The experimenter ended the trial when the participant indicated that he or she had finished responding by pointing to a thumbs-up picture, or after 20 seconds from trial onset, whichever came first. This procedure was designed to avoid feedback of any kind from the experimenter. Each naming trial was followed by familiarity ratings for that item. First, the picture was shown above the written prompt "Do you recognize this person or thing? 1=yes, 2=not sure, 3=no." After the experimenter recorded the participant's answer, the picture was shown above the written prompt "Even if you can't think of the name right now, would you recognize the name if you saw it? 1=yes, 2=not sure, 3=no." After the experimenter recorded the participant's answer, the next naming trial was presented.

At the end of the item selection phase, items were identified for each participant that met the following criteria: (1) the item elicited a correct response in the picture-name matching task, (2) the item elicited an error naming response or no response on both administrations of the naming test[1], and (3) the participant recognized the entity and the name as indicated by a '1' or '2' response on both administrations of each rating. Across the 864 total items selected for the group, only 17% received a '2' response for name recognition on one or both administrations, and only 2% received a '2' response for person recognition on one or both administrations. For each participant, 144 items that met these criteria were divided evenly among the six training conditions, matching as closely as possible for the length of the names (i.e., numbers of phonemes and syllables) and the numbers of specific item types (i.e., person, character, or movie).

**2.3.2    1-day ISI Training Phase**—The 1-day ISI training phase comprised three training sessions that were administered on three consecutive days. Each training session consisted of approximately 1.5 hours of retrieval practice on 72 items and involved as many breaks as the participant requested. In each training session, an item was dropped from further training

---

[1]P3 and P4 were assigned some items that were named correctly on one of the two administrations. These constituted less than 15% of the participant's items and were distributed evenly across the training conditions.

after the item was accurately named on one, two, or four retrieval practice trials, depending on criterion level condition (hereafter referred to as the *Criterion-1*, *Criterion-2*, and *Criterion-4* conditions, respectively). Items were administered in six blocks of 12 items, with two blocks per criterion level condition per session. The order of criterion level condition blocks was counterbalanced across participants and varied across the three sessions for each participant. In each case, the order of the blocks was designed to minimize differences across conditions with regard to effects of participant fatigue. For example, if one block of Criterion-4 items was administered first, the remaining block of Criterion-4 items was administered last: Criterion-4 – Criterion-2 – Criterion-1 – Criterion-1 – Criterion-2 – Criterion-4.

In Session 1, each block of retrieval practice was preceded by a repetition trial for each of the 12 items in the block so that the correct name for each entity was presented prior to retrieval practice. This design feature is not commonly used in naming treatment for aphasia but was incorporated to parallel the typical paradigm in psychological studies of retrieval practice. On a repetition trial, the picture was presented simultaneously with the written and auditory forms of the corresponding name. The participant was asked to repeat the name once. The picture and written form of the name remained on the screen for eight seconds, and then the experimental software advanced to the next trial. After all pictures in a block were administered for repetition in a randomized order, the pictures were presented in that same order for retrieval practice. On a retrieval practice trial, the picture was presented for eight seconds, and the participant was asked to try to produce the full name for the entity. For each retrieval practice trial, the experimenter pressed a key that recorded the naming response as correct if the correct response was produced within the duration of the trial; otherwise, error was indicated.[2] If the item reached its assigned criterion level (correct naming on 1, 2, or 4 retrieval practice trials), the experimental software automatically dropped the item from further training. Otherwise, an additional retrieval practice trial for the item was added to the end of the block. Regardless of the accuracy of the response, each retrieval practice trial was followed by correct-answer feedback. Feedback was identical to the repetition procedure that preceded each block except that the picture and written name were presented for only five seconds rather than eight seconds. The block ended when every item in the block reached its assigned criterion level, or when naming repeatedly failed for the last one or two items in the block.

The procedures for Session 2 and Session 3 were identical to those in Session 1, with the critical exception that the initial repetition trials were excluded. Without these trials, the participant had no opportunity to see/hear the name of an item prior to his or her first naming attempt for that item in the session. These initial naming attempts on the first training trial for the items in Sessions 2 and 3 served as a test of one-day retention of learning in the prior session(s). Hence, two important tests in this study comprised the initial

---

[2]During the training session, the experimenter attempted to apply the same criterion for correct naming as was used for offline coding (see section 2.4). However, as response coding during the session was necessarily applied "on the fly" without the benefit of transcription or the time for formal analysis, there were deviations between in-session and offline coding of training responses. These discrepancies are reported in section 3.1.

training trial for each item in Session 2 (hereafter, *Session 2 Test*) and in Session 3 (hereafter, *Session 3 Test*).

Four weeks after Session 3, the *Long-term Retention Test* was administered. All 72 items were administered in a random order using the same naming procedures as in the item selection phase (i.e., 20 seconds to name each item, with no feedback).

**2.3.3   7-day ISI Training Phase**—The 7-day ISI training phase followed the same procedures as the 1-day ISI phase, with the exception of the intersession interval. In the 7-day ISI phase, the three training sessions were spaced seven days apart. Hence, the Session 2 Test and Session 3 Test in this phase measured *one-week* retention of learning in the prior session(s).

## 2.4   Naming Accuracy Coding

Here we provide a brief overview of the naming accuracy coding system. For a full description, see Middleton et al. (2016). Participant responses on retrieval practice trials and at the retention tests were digitally recorded and subsequently transcribed by trained coders. Most of the proper noun targets were composed of multiple morphemes, sometimes eliciting a last name prior to the first during naming attempts (e.g., "Seinfeld… Jerry Seinfeld"). Hence, naming accuracy coding began by mapping each of the target name's constituents (e.g., first name and last name) to the best response constituent from among all non-fragmented constituents produced within the given time limit. The "best response constituent" was determined by a phonological overlap formula (Lecours & Lhermitte, 1969). This formula yields a continuous measure of phonological similarity between a response and target, standardized across different word lengths (Formula 1). Shared phonemes were identified independent of position, and credit was assigned only once if a response had two instances of a single target phoneme. After selecting the best response constituents, Formula 1 was applied to calculate the phonological overlap between the whole response (e.g., Jerry+Seinfeld) and the full target name.

$$\text{phonological overlap} = \frac{\#\ \text{shared phonemes in target and response} \times 2}{\Sigma\ \text{phonemes in target and response}} \tag{1}$$

Consistent with Middleton et al. (2016), naming responses were coded as correct if (1) each best response constituent had at least 0.5 phonological overlap with the target constituent and (2) the whole response had at least 0.75 phonological overlap with the full target name. Responses that did not meet both of the criteria were coded as error. The rationale for coding responses with most of the target name as correct was to credit successful name retrieval while disregarding minor deviations in phonological form that can arise during post-lexical processes.

## 2.5   Data Analyses

In the analyses, we report average naming accuracy for each condition of interest and the average number of training trials per item, i.e., the total number of training trials for that condition administered prior to the point of testing divided by the number of items in that

condition (see Figures 2 and 3). This latter measure represents the cost in terms of time invested in training, for comparison with the final naming accuracy achieved. Paired t-tests were used to test differences in the number of trials administered across conditions. However, the primary focus of statistical analyses for this study was naming accuracy on the relevant retrieval practice trials during training and at the Long-Term Retention Tests. Results were analyzed using logistic mixed effects regression applied to the binary naming accuracy variable (correct/error) with alpha=0.05 for tests of significance and dummy coding for fixed effects. These models offer the advantage of evaluating fixed-effects factors to test key predictions at a group level, and at the same time capturing dependencies among observations due to participants providing multiple responses to overlapping item sets (see Baayen et al., 2008). The random effects structure of each model included by-subject and by-item random intercepts unless inclusion of either term caused model nonconvergence, suggestive of overfitting. By-subject slopes were tested but not included in any of the final models due to nonconvergence, lack of improved model fit by a chi-square test of deviance in model log-likelihoods (alpha=0.05), and/or absence of improved model fit indices (i.e., Bayesian information criterion and Akaike information criterion). Key results are reported in the following section. Full details of the final models and effects observed for each participant are reported in the appendix.

## 3. Results

### 3.1 Training Schedule Fidelity

Due to participants' availability and other constraints, there were minor deviations from the assigned training schedules. Between the training sessions, the average intersession interval in the 1-day ISI phase was 1.1 days ($SD$=0.5), and the average intersession interval in the 7-day ISI phase was 7.3 days ($SD$=1.2). The number of days that elapsed between Session 3 and the Long-term Retention Test ranged from 19–31 ($M$=25.8, $SD$=3.5). For each participant, however, the retention interval in the 1-day ISI phase was no more than three days shorter or longer than the retention interval in the 7-day ISI phase. In addition to deviations in the spacing of practice, some items were named correctly on too few or too many trials relative to the assigned criterion level ($M$=14% of items per training session per participant). These instances occurred due to experimenter error in online response coding, or because the participant's repeated failure to name an item resulted in discontinuation of the training block before the item's criterion level was achieved. On average, the actual number of correct responses per item per session closely matched the assigned criterion level for Criterion-1 items ($M$=1.04), Criterion-2 items ($M$=2.02), and Criterion-4 items ($M$=3.98).

### 3.2 Effects of Criterion Level

Table 2 reports average naming accuracy for Criterion-1 items, Criterion-2 items, and Criterion-4 items on initial retrieval practice trials for each item in each training session. In each condition, the first retrieval practice trial for an item in Session 1 occurred 12 trials after the initial repetition trial for that item. Hence, if items were well-matched across the conditions, we would expect similar accuracy on the first retrieval practice trial of Session 1.

As reported in Table 2, Session 1 naming accuracy ranged from 63–69% correct across the six training conditions, suggesting similar item difficulty across the conditions.

**3.2.1 Initial Criterion Level**—On initial trials for items at Session 2 (line 2, Table 2) in the 1-day ISI phase, the Criterion-4 condition ($M$=63.4% correct) significantly outperformed the Criterion-1 condition ($M$=51.9% correct), *Estimate*=0.56, *SE*=0.22, $p$=.010, and the Criterion-2 condition ($M$=52.3% correct), *Estimate*=0.55, *SE*=0.22, $p$=.012. Results did not indicate a significant difference between Criterion-1 and Criterion-2 conditions in the 1-day ISI phase ($p$=.96). (See Appendix Table A.1.) On initial trials for items at Session 2 in the 7-day ISI phase, the Criterion-4 condition ($M$=32.9% correct) significantly outperformed the Criterion-1 condition ($M$=22.7% correct), *Estimate*=0.52, *SE*=0.22, $p$=.018, but the advantage over the Criterion-2 condition ($M$=25.5% correct) was marginal, *Estimate*=0.37, *SE*=0.22, $p$=.088. As in the 1-day ISI phase, results did not indicate a significant difference between the Criterion-1 and Criterion-2 conditions in the 7-day ISI phase ($p$=.49). (See Appendix Table A.2.)

**3.2.2 Criterion Level Across Sessions**—To examine the effect of criterion level after multiple training sessions, Long-term Retention Test performance was compared for the three criterion level conditions within each ISI phase (see Figure 2). In the 1-day ISI phase, the only significant advantage from higher criterion level was Criterion-4 ($M$=48.1% correct) compared to Criterion-2 ($M$=38.0% correct), *Estimate*=0.45, *SE*=0.21, $p$=.029. (See Appendix Table A.3.) However, the difference in means (10.1%) came at a significantly greater cost of approximately 6 more trials per item, or on average 138 more trials total per participant across the training sessions, $t(8)$=18.5, $p$<.001. In the 7-day ISI phase, the only instance in which higher criterion conferred significantly higher performance was in the Criterion-4 ($M$=55.1% correct) compared to the Criterion-1 condition ($M$=40.3% correct), *Estimate*=0.66, *SE*=0.21, $p$=.002. (See Appendix Table A.4.) Here, the additional 14.8% in naming accuracy was obtained at a significantly greater cost of an additional 10 trials per item, or on average 241 more trials total per participant across the training sessions, $t(8)$=23.8, $p$<.001.

## 3.3 Effects of Distributed Practice

**3.3.1 Within- versus Across-session**—Within each ISI phase, we focused on key points in the design at which two different item sets had accumulated the same total number of correct responses but differed with regard to whether those responses occurred *within* one prior session or *across* two prior sessions. To test this effect for items that had been retrieved correctly a total of two times, we compared the Criterion-2 items at the Session 2 Test versus the Criterion-1 items at the Session 3 Test ($W_2$ vs. $A_2$ in Figure 1, where "W" denotes within-session and "A" denotes across-session). To test this effect for items retrieved correctly a total of four times, we compared the Criterion-4 items at the Session 2 Test versus the Criterion-2 items at the Session 3 Test ($W_4$ vs. $A_4$ in Figure 1). Each of these two comparisons were made within each of the two ISI phases, resulting in a total of four separate tests of within-session versus across-session distribution.

Figure 3 displays the results of the four comparisons of within- versus across-session distribution of practice. In the 1-day ISI phase, naming accuracy reflects performance one day after the prior training session. In this phase, across-session distribution was associated with a significant advantage over within-session training (additional 17.6% in naming accuracy) for items that had been retrieved correctly a total of two times, *Estimate*=0.82, *SE*=0.21, *p*<.001. (See Appendix Table A.5.) The advantage for across-session distribution came at the mere cost of less than one additional trial per item, or on average only 12 more trials total per participant across the training sessions compared to the within-session condition, $t(8)=1.5$, *p*=.16. In the 1-day ISI phase, counter to predictions, the within- versus across-session comparison was not significantly different for items that had been retrieved correctly a total of four times, *Estimate*=−0.05, *SE*=0.24, *p*=.84. (See Appendix Table A.6.)

In the 7-day ISI phase, naming accuracy reflects performance seven days after the prior training session. In this phase, across-session distribution produced superior naming accuracy compared to the within-session condition for items that had been retrieved correctly a total of two times, *Estimate*=0.80, *SE*=0.23, *p*<.001. (See Appendix Table A.7.) Here, the advantage of an additional 16.6% in naming accuracy for across- versus within-session distribution came at the statistically significant but small cost of less than one additional trial per item, or on average only 20 more trials total per participant across the training sessions, $t(8)=5.8$, *p*<.001. Similarly, for items retrieved correctly a total of four times in the 7-day ISI phase, there was a significant 20.8% advantage for across- versus within-session training, *Estimate*=0.87, *SE*=0.20, *p*<.001, at the cost of less than one additional trial per item, or on average only 21 more trials total per participant across the training sessions, $t(8)=3.8$, *p*<.01. (See Appendix Table A.8.) In sum, in three of four comparisons, across-session distribution conferred a robust advantage over within-session presentation at retention intervals of one day and seven days, with little additional cost in terms of number of additional trials.

**3.3.2 Intersession Interval**—Figure 2 displays naming accuracy for Criterion-1 items, Criterion-2 items, and Criterion-4 items at the Long-term Retention Test in each training phase. Because the interaction between criterion level and intersession interval did not significantly improve the model of naming accuracy, $\chi^2$=2.53, *p*=.28, we collapsed across criterion level for the analysis of intersession interval. The difference in naming accuracy between the 1-day ISI (*M*=42.4% correct) and the 7-day ISI (*M*=47.5% correct) was marginal, *Estimate*=0.22, *SE*=0.12, *p*=.057. Six of the nine participants scored numerically higher on the Long-term Retention Test in the 7-day ISI condition compared to the 1-day ISI condition. (See Appendix Table A.9.) As Figure 2 shows, the added cost in terms of number of trials per item for the 7-day versus the 1-day ISI condition was minimal at less than 1 additional trial per item, although the consistently greater number of trials required for the 7-day schedule resulted in a statistically significant difference between the two conditions, $t(8)=3.0$, *p*<.05.

## 4. Discussion

Several key findings in the present study demonstrate that learning factors that have been shown to enhance knowledge acquisition also exert long-lasting changes on lexical access in

aphasia, with similar implications for learning efficiency across the two domains. In the following sections, we discuss the details of these findings and their theoretical and clinical implications.

### 4.1 Criterion Level Effects

Individuals with aphasia showed increased naming accuracy at the second session from increasing criterion level during the initial session. Effects of initial criterion level were apparent after retention intervals of 1 day and 7 days, although these effects were only consistently observed in the comparison between the Criterion-4 and the Criterion-1 conditions. After items were trained to their assigned criterion in each of three training sessions, performance showed a significant advantage for Criterion-4 over Criterion-1 at long-term retention in the 1-day ISI phase and for Criterion-4 over Criterion-2 at long-term retention in the 7-day ISI phase. In each case, the advantage was achieved at the cost of large increases in the number of training trials. In addition, the presence and magnitude of benefit from increased criterion levels were variable across participants (see Appendix Tables A.1–A.4 for effects per participant).

It is apparent, but rarely noted in treatment studies, that the benefits of increased amounts of practice incur the costs of increased time spent in training. The results of this study suggest that there may be a point at which increasing the criterion level for an item within a session becomes highly inefficient. This type of inefficiency is suggested in prior studies of naming training in aphasia that included variations in the amount of training trials per item within each of multiple sessions. These studies showed that the number of training trials per item per session had little or no effect on post-training naming accuracy (Laganaro, Pietro, & Schnider, 2006; Off, Griffin, Spencer, & Rogers, 2015). In contrast, increasing the amount of practice by adding additional training sessions does have benefits for the treatment of aphasia (Des Roches, Balachandran, Ascenso, Tripodis, & Kiran, 2015; Friedman et al., 2017). This evidence leads to the testable hypothesis that the amount and spacing of training interact, such that increasing the amount of training has greater benefits when the additional training is scheduled as *more sessions*, rather than more trials per item within sessions.

We aim to draw attention to the goal of efficient training schedules for two primary reasons. First, the current healthcare environment often constrains the amount of time a clinician can work with a patient. More efficient schedules of practice will result in a greater number of items that can be mastered in the same amount of time. Laganaro et al. (2006) demonstrated this principle by showing that participants learned a greater number of words when they doubled the number of items practiced in the same amount of time. Second, we expect that examining efficient schedules of practice will ultimately provide evidence in support of increasing the duration of language training schedules by showing that additional training administered after a sufficient length of time has potent effects on long-term maintenance of gains. This aspect of efficiency was exemplified in the present study using the across-session effect to achieve significantly higher naming accuracy with only a small increase in the number of training trials (as discussed below) as well as numerically higher naming accuracy with *fewer* trials (see the results of two correct retrievals across sessions compared to four correct retrievals within a session in Figure 3).

### 4.2   Distributed Practice Effects

The largest and most consistent effects in the present study were induced by distributing practice across sessions, relative to within a single session (see Appendix Tables A.5–A.8 for effects per participant). Compared to two correct retrievals of an item within a single session, two correct retrievals across two training sessions (i.e., once in the first session and once in the second) resulted in an additional 17–18% in naming accuracy at 1 day or 7 days post-training. When the test was administered 1 day post-training, distributing four correct retrievals of an item across two sessions (i.e., twice in the first session and twice in the second) did not result in superior naming accuracy compared to four correct retrievals within a session. However, the same manipulation yielded an across-session advantage of an additional 21% in naming accuracy at 7 days post-training. This pattern of results suggests that some participants may have approached their ceiling of performance at 1-day post-training when relatively large amounts of training had been administered, whereas the difficulty induced by a longer retention interval revealed an across-session effect, despite the large amount of training per item.

We also observed a trend towards an advantage of the 7-day intersession interval compared to the 1-day intersession interval, but the effect was small and not statistically significant. Few aphasia treatment studies have manipulated the intersession interval while controlling for the type and total amount of treatment and length of treatment sessions. These studies have shown little difference in the outcomes of shorter versus longer intersession intervals (Dignam et al., 2016; Martins et al., 2013; Ramsberger & Marie, 2007; Raymer et al., 2006) or significantly better naming outcomes with longer intersession intervals (Dignam et al., 2015; Sage et al., 2011). The effect of the interval between practice sessions remains poorly understood in both the psychological learning and aphasia treatment literatures. Nevertheless, the results of the present study and prior studies in aphasia challenge a prevalent hypothesis that compressing large amounts of training within relatively short periods is superior to more distributed schedules for aphasia treatment (for a recent review, see Middleton, Schuchard, & Rawson, 2019).

### 4.3   Study Limitations and Future Directions

It is important to note limitations to the clinical implications of the present study. This study used a specific type of stimuli (proper nouns), evaluated results across a group of relatively homogeneous participants, and focused exclusively on naming gains for trained items rather than generalized improvement in lexical access. Further research will be needed to replicate the present findings with other training materials and with individuals with a greater severity and variety of language impairments. The latter is particularly important because the optimal schedule of training likely differs across individuals. Studies with more heterogeneous and larger numbers of participants would have the statistical power to test individual factors that may interact with scheduling factors to affect learning outcomes.

The results of the present study suggest priorities for future research to advance the understanding of learning principles and their applications for language treatment. The results of this study show a powerful effect of distributing practice across sessions on naming performance one day or one week later. An important next step will be testing the

effect of distributing practice across sessions on performance after longer retention intervals. In contrast to the strong effect of distributing practice across sessions, the length of the interval between sessions may have a relatively low impact on naming outcomes. Establishing a clinically meaningful effect of intersession interval may require conditions that rarely occur in applied settings, such as a comparison of daily sessions to sessions that are spaced more than a week apart. An alternative schedule of practice to examine is an expanding schedule in which intersession intervals are initially short and gradually increase. It is possible that this type of schedule could benefit individuals that have high difficulty during practice after a long interval has elapsed. Similarly, the intervals between practice for a particular item *within* each training session could be initially short and gradually increase, as examined by Fridriksson et al. (2005) in people with aphasia. Finally, a better understanding of the neurocognitive mechanisms underlying distributed practice effects is needed to integrate these effects into models of lexical learning.

This study adds to the growing evidence that principles of learning established by psychological research operate in naming rehabilitation in aphasia. In other neuropsychological populations, studies have shown the relevance of distributed practice principles for enhancing learning in standard laboratory tasks (e.g., paired associate learning) and the learning of functional information such as remembering appointments or learning people's names (e.g., in dementia, Balota, Ducheck, Sergent-Marshall, & Roediger, 2006; in multiple sclerosis, Goverover, Basso, Wood, Chiaravalloti, & DeLuca, 2011; Sumowski, Chiaravalloti, & DeLuca, 2010; in traumatic brain injury, Goverover, Arango-Lasprilla, Hillary, Chiaravalloti, & DeLuca, 2009; Hillary et al., 2003). Such observations, along with the results of the present work, suggest significant potential for the application of distributed practice principles to inform and improve neurorehabilitation practices.

## Acknowledgements

## Appendix

## Results of Statistical Models with Observed Effects per Participant

Statistically significant results ($p<.05$) are indicated by an asterisk. Coef. = model estimation of the change in naming accuracy (in log odds) from the reference category for each fixed effect. SE = standard error. p = p-value. Var. = variance. SD = standard deviation. Obs. = number of observations (items or participants) on which the random effect was assessed. Data points = number of trials modeled. ISI = intersession interval. P1-P9 = Observed difference in naming accuracy from the reference category for participants P1 through P9. For example, the first value for P1 in Table A.1 (0.33) may be interpreted as the following: P1's naming accuracy was 33 percentage points higher in the Criterion-2 condition compared to the Criterion-1 condition.

**Table A.1**

Initial Criterion Level (Section 3.2.1): 1-Day ISI Phase

| Model Output | | | | Observed Effects per Participant | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Fixed effects** | **Coef.** | **SE** | **p** | **P1** | **P2** | **P3** | **P4** | **P5** | **P6** | **P7** | **P8** | **P9** |
| Criterion-2[a] | 0.01 | 0.22 | 0.96 | 0.33 | −0.04 | −0.04 | −0.25 | 0.08 | −0.25 | 0.08 | 0.00 | 0.13 |
| Criterion-4[a] | 0.56 | 0.22 | 0.01* | 0.21 | 0.04 | 0.08 | 0.17 | 0.17 | 0.13 | 0.25 | 0.00 | 0.00 |
| Criterion-4[b] | 0.55 | 0.22 | 0.01* | −0.13 | 0.08 | 0.13 | 0.42 | 0.08 | 0.38 | 0.17 | 0.00 | −0.13 |
| **Random effects** | **Var.** | **SD** | **Obs.** | | | | | | | | | |
| Participants | 0.68 | 0.82 | 9 | | | | | | | | | |
| Items | 0.26 | 0.5 | 369 | | | | | | | | | |
| Data points = 648 | | | | | | | | | | | | |

[a]Reference level = Criterion-1.

[b]Reference level = Criterion-2.

**Table A.2**

Initial Criterion Level (Section 3.2.1): 7-Day ISI Phase

| Model Output | | | | Observed Effects per Participant | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Fixed effects** | **Coef.** | **SE** | **p** | **P1** | **P2** | **P3** | **P4** | **P5** | **P6** | **P7** | **P8** | **P9** |
| Criterion-2[a] | 0.16 | 0.23 | 0.49 | −0.13 | 0.13 | −0.08 | 0.04 | −0.08 | 0.04 | 0.13 | 0.04 | 0.17 |
| Criterion-4[a] | 0.52 | 0.22 | 0.02* | 0.04 | 0.00 | −0.13 | 0.17 | 0.08 | 0.13 | 0.21 | 0.04 | 0.38 |
| Criterion-4[b] | 0.37 | 0.22 | 0.09 | 0.17 | −0.13 | −0.04 | 0.13 | 0.17 | 0.08 | 0.08 | 0.00 | 0.21 |
| **Random effects** | **Var.** | **SD** | **Obs.** | | | | | | | | | |
| Participants | 0.11 | 0.34 | 9 358 | | | | | | | | | |
| Items | 0.00 | 0.06 | 358 | | | | | | | | | |
| Data points = 648 | | | | | | | | | | | | |

[a]Reference level = Criterion-1.

[b]Reference level = Criterion-2.

**Table A.3**

Criterion Level Across Sessions (Section 3.2.2): 1-Day ISI Phase

| Model Output | | | | Observed Effects per Participant | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Fixed effects** | **Coef.** | **SE** | **p** | **P1** | **P2** | **P3** | **P4** | **P5** | **P6** | **P7** | **P8** | **P9** |
| Criterion-2[a] | −0.15 | 0.21 | 0.46 | 0.17 | 0.08 | −0.21 | 0.13 | −0.08 | −0.21 | −0.04 | −0.08 | −0.04 |
| Criterion-4[a] | 0.30 | 0.20 | 0.14 | 0.04 | −0.21 | 0.08 | 0.29 | 0.17 | 0.17 | 0.00 | 0.04 | 0.04 |
| Criterion-4[b] | 0.45 | 0.21 | 0.03 * | −0.13 | −0.29 | 0.29 | 0.17 | 0.25 | 0.38 | 0.04 | 0.13 | 0.08 |
| **Random effects** | **Var.** | **SD** | **Obs.** | | | | | | | | | |

| Model Output | | | Observed Effects per Participant |
|---|---|---|---|
| Participants | 0.22 | 0.47 | 9 |
| Items | 0.14 | 0.37 | 369 |
| Data points = 648 | | | |

[a]Reference level = Criterion-1.

[b]Reference level = Criterion-2.

**Table A.4**

Criterion Level Across Sessions (Section 3.2.2): 7-Day ISI Phase

| Model Output | | | | Observed Effects per Participant | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Fixed effects** | **Coef.** | **SE** | **p** | **P1** | **P2** | **P3** | **P4** | **P5** | **P6** | **P7** | **P8** | **P9** |
| Criterion-2[a] | 0.30 | 0.21 | 0.16 | 0.00 | −0.08 | −0.25 | 0.25 | −0.13 | 0.08 | 0.29 | 0.25 | 0.21 |
| Criterion-4[a] | 0.66 | 0.21 | <.01* | 0.04 | 0.17 | −0.25 | 0.25 | 0.00 | 0.46 | 0.21 | 0.29 | 0.17 |
| Criterion-4[b] | 0.36 | 0.21 | 0.09 | 0.04 | 0.25 | 0.00 | 0.00 | 0.13 | 0.38 | −0.08 | 0.04 | −0.04 |
| **Random effects** | **Var.** | **SD** | **Obs.** | | | | | | | | | |
| Participants | 0.08 | 0.28 | 9 | | | | | | | | | |
| Items | 0.33 | 0.57 | 358 | | | | | | | | | |
| Data points = 648 | | | | | | | | | | | | |

[a]Reference level = Criterion-1.

[b]Reference level = Criterion-2.

**Table A.5**

Within- versus Across-session (Section 3.3.1): 1-Day ISI Phase/2 Correct Retrievals

| Model Output | | | | Observed Effect per Participant | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Fixed effects** | **Coef.** | **SE** | **p** | **P1** | **P2** | **P3** | **P4** | **P5** | **P6** | **P7** | **P8** | **P9** |
| Across-session[a] | 0.82 | 0.21 | <.001* | −0.25 | 0.13 | 0.21 | 0.50 | 0.21 | 0.46 | 0.00 | 0.25 | 0.08 |
| **Random effects** | **Var.** | **SD** | **Obs.** | | | | | | | | | |
| Participants | 0.41 | 0.64 | 9 | | | | | | | | | |
| Data points = 432 | | | | | | | | | | | | |

[a]Reference level = Within-session.

**Table A.6**

Within- versus Across-session (Section 3.3.1): 1-Day ISI Phase/4 Correct Retrievals

| Model Output | | | | Observed Effect per Participant | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Fixed effects** | **Coef.** | **SE** | **p** | **P1** | **P2** | **P3** | **P4** | **P5** | **P6** | **P7** | **P8** | **P9** |
| Across-session[a] | −0.05 | 0.24 | 0.83 | 0.00 | 0.00 | 0.13 | −0.29 | 0.13 | −0.08 | −0.08 | 0.00 | 0.17 |
| **Random effects** | **Var.** | **SD** | **Obs.** | | | | | | | | | |
| Participants | 1.20 | 1.09 | 9 | | | | | | | | | |

| Model Output | | | Observed Effect per Participant |
|---|---|---|---|
| Items | 0.53 | 0.73 | 290 |
| Data points = 432 | | | |

[a]Reference level = Within-session.

**Table A.7**

Within- versus Across-session (Section 3.3.1): 7-Day ISI Phase/2 Correct Retrievals

| Model Output | | | | Observed Effect per Participant | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Fixed Effects** | **Coef.** | **SE** | **p** | **P1** | **P2** | **P3** | **P4** | **P5** | **P6** | **P7** | **P8** | **P9** |
| Across-session[a] | 0.80 | 0.23 | <.001* | 0.33 | 0.33 | 0.17 | 0.29 | 0.29 | 0.17 | 0.21 | 0.29 | 0.21 |
| **Random Effects** | **Var.** | **SD** | **Obs.** | | | | | | | | | |
| Items | 0.22 | 0.47 | 279 | | | | | | | | | |
| Data points = 432 | | | | | | | | | | | | |

[a]Reference level = Within-session.

**Table A.8**

Within- versus Across-session (Section 3.3.1): 7-Day ISI Phase/4 Correct Retrievals

| Model Output | | | | Observed Effect per Participant | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Fixed effects** | **Coef.** | **SE** | **p** | **P1** | **P2** | **P3** | **P4** | **P5** | **P6** | **P7** | **P8** | **P9** |
| Across-session[a] | 0.87 | 0.20 | <.001* | 0.21 | 0.46 | 0.33 | 0.13 | 0.08 | 0.29 | 0.29 | 0.04 | 0.04 |
| **Random effects** | **Var.** | **SD** | **Obs.** | | | | | | | | | |
| Participants | 0.06 | 0.24 | 9 | | | | | | | | | |
| Data points = 432 | | | | | | | | | | | | |

[a]Reference level = Within-session.

**Table A.9**

Intersession Interval (Section 3.3.2)

| Model Output | | | | Observed Effects per Participant | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Fixed effects** | **Coef.** | **SE** | **p** | **P1** | **P2** | **P3** | **P4** | **P5** | **P6** | **P7** | **P8** | **P9** |
| 7-day ISI[a] | 0.22 | 0.12 | 0.06 | 0.11 | –0.06 | 0.00 | 0.07 | –0.03 | 0.11 | 0.10 | 0.07 | 0.08 |
| **Random effects** | **Var.** | **SD** | **Obs.** | | | | | | | | | |
| Participants | 0.16 | 0.40 | 9 | | | | | | | | | |
| Items | 0.13 | 0.37 | 478 | | | | | | | | | |
| Data points = 1296 | | | | | | | | | | | | |

[a]Reference level = 1-day ISI.

# References

Anderson JR (2007). How can the human mind occur in the physical universe? New York, NY: Oxford University Press.

Baayen RH, Davidson DJ, & Bates DM (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59(4), 390–412. 10.1016/j.jml.2007.12.005

Bahrick HP (1979). Maintenance of knowledge: Questions about memory we forgot to ask. Journal of Experimental Psychology: General, 108, 296–308. 10.1037/0096-3445.108.3.296

Bahrick HP, & Hall LK (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. Journal of Memory and Language, 52(4), 566–577. 10.1016/j.jml.2005.01.012

Bahrick HP, Bahrick LE, Bahrick AS, & Bahrick PE (1993). Maintenance of foreign language vocabulary and the spacing effect. Psychological Science, 4(5), 316–321. 10.1111/j.1467-9280.1993.tb00571.x

Balota DA., Ducheck JM., Sergent-Marshall SD., & Roediger HL III. (2006). Does expanded retrieval practice produce benefits over equal-interval spacing? Explorations of spacing effects in healthy aging and early stage Alzheimer's disease. Psychology and Aging, 21(1), 19–31. 10.1037/0882-7974.21.1.19 [PubMed: 16594788]

Burke DM, Mackay DG, Worthley JS, & Wade E (1991). On the tip of the tongue: What causes word finding failures in young and older adults. Journal of Memory and Language, 30(5), 542–579. 10.1016/0749-596X(91)90026-G

Cepeda NJ, Coburn N, Rohrer D, Wixted JT, Mozer MC, & Pashler H (2009). Optimizing distributed practice. Experimental Psychology, 56(4), 236–246. 10.1027/1618-3169.56.4.236 [PubMed: 19439395]

Cepeda NJ, Pashler H, Vul E, Wixted JT, & Rohrer D (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. Psychological Bulletin, 132(3), 354–380. 10.1037/0033-2909.132.3.354 [PubMed: 16719566]

Cepeda NJ, Vul E, Rohrer D, Wixted JT, & Pashler H (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. Psychological Science, 19(11), 1095–1102. 10.1111/j.1467-9280.2008.02209.x [PubMed: 19076480]

Cherney LR (2012). Aphasia treatment: Intensity, dose parameters, and script training. International Journal of Speech-Language Pathology, 14(5), 424–431. 10.3109/17549507.2012.686629 [PubMed: 22731660]

Delaney PF, Verkoeijen PPJL, & Spirgel A (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature In Psychology of learning and motivation (Vol. 53, pp. 63–147). Academic Press 10.1016/S0079-7421(10)53003-2

Dell GS, Schwartz MF, Martin N, Saffran EM, & Gagnon DA (1997). Lexical access in aphasic and nonaphasic speakers. Psychological Review, 104(4), 801–838. 10.1037/0033-295X.104.4.801 [PubMed: 9337631]

Des Roches CA., Balachandran I., Ascenso EM., Tripodis Y., & Kiran S. (2015). Effectiveness of an impairment-based individualized rehabilitation program using an iPadbased software platform. Frontiers in Human Neuroscience, 8, 1015 10.3389/fnhum.2014.01015 [PubMed: 25601831]

Dignam JK, Copland D, McKinnon E, Burfein P, O'Brien K, & Farrell A et al. (2015). Intensive versus distributed aphasia therapy: A nonrandomized, parallel-group, dosagecontrolled study. Stroke, 46(8), 2206–2211. 10.1161/STROKEAHA.115.009522 [PubMed: 26106114]

Dignam JK, Copland D, Rawlings A, O'Brien K, Burfein P, & Rodriguez AD (2016). The relationship between novel word learning and anomia treatment success in adults with chronic aphasia. Neuropsychologia, 81, 186–197. 10.1016/j.neuropsychologia.2015.12.026 [PubMed: 26724545]

Fridriksson J, Holland A, Beeson P, & Morrow L (2005). Spaced retrieval treatment of anomia. Aphasiology, 19(2), 99–109. 10.1080/02687030444000660 [PubMed: 16823467]

Friedman RB, Sullivan KL, Snider SF, Luta G, & Jones KT (2017). Leveraging the test effect to improve maintenance of the gains achieved through cognitive rehabilitation. Neuropsychology, 31(2), 220–228. 10.1037/neu0000318 [PubMed: 27732041]

Goldrick M, & Rapp B (2007). Lexical and post-lexical phonological representations in spoken production. Cognition, 102(2), 219–260. 10.1016/j.cognition.2005.12.010 [PubMed: 16483561]

Goodglass H, & Kaplan E (1983). The assessment of aphasia and related disorders, second edition. Philadelphia, PA: Lea & Febiger.

Goverove Y., Arango-Lasprill JC., Hillar FG., Chiaravallot N., & DeLuc J. (2009). Application of the spacing effect to improve learning and memory for functional tasks in traumatic brain injury: A pilot study. American Journal of Occupational Therapy, 63(5), 543–548. 10.5014/ajot.63.5.543 [PubMed: 19785253]

Goverover Y, Basso M, Wood H, Chiaravalloti N, & DeLuca J (2011). Examining the benefits of combining two learning strategies on recall of functional information in persons with multiple sclerosis. Multiple Sclerosis Journal, 17(12), 1488–1497. 10.1177/1352458511406310 [PubMed: 21586486]

Hillary FG, Schultheis MT, Challis BH, Millis SR, Carnevale GJ, Galshi T, DeLuca J (2003). Spacing of repetitions improves learning and memory after moderate and severe TBI. Journal of Clinical and Experimental Neuropsychology, 25(1), 49–58. 10.1076/jcen.25.1.49.13631 [PubMed: 12607171]

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

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. 10.1016/j.cognition.2005.02.006 [PubMed: 16413014]

Kertesz A (1982). Western Aphasia Battery. New York, NY: Grune & Stratton.

Kornell N, & Vaughn KE (2016). How retrieval attempts affect learning: A review and synthesis In Psychology of learning and motivation (Vol. 65, pp. 183–215). Academic Press.

Laganaro M, Pietro MD, & Schnider A (2006). Computerised treatment of anomia in acute aphasia: Treatment intensity and training size. Neuropsychological Rehabilitation, 16(6), 630–640. 10.1080/09602010543000064 [PubMed: 17127569]

Lecours AR, & Lhermitte F (1969). Phonemic paraphasias: Linguistic structures and tentative hypotheses. Cortex, 5(3), 193–228. 10.1016/S0010-9452(69)80031-6 [PubMed: 5391877]

Martin RC, Shelton JR, & Yaffee LS (1994). Language processing and working memory: Neuropsychological evidence for separate phonological and semantic capacities. Journal of Memory and Language, 33(1), 83–111. 10.1006/jmla.1994.1005

Martins IP., Leal G., Fonseca I., Farrajota L., Aguiar M., Fonseca J., et al. (2013). A randomized, rater-blinded, parallel trial of intensive speech therapy in sub-acute post-stroke aphasia: the SP-I-R-IT study. International Journal of Language & Communication Disorders, 48(4), 421–431. 10.1111/1460-6984.12018 [PubMed: 23889837]

Middleton EL, Schuchard J, & Rawson KA (2020). A review of the application of distributed practice principles to naming treatment in aphasia. Topics in Language Disorders.

Middleton EL, Schwartz MF, Rawson KA, & Garvey K (2015). Test-enhanced learning versus errorless learning in aphasia rehabilitation: Testing competing psychological principles. Journal of Experimental Psychology: Learning, Memory, and Cognition, 41(4), 1253–1261. 10.1037/xlm0000091

Middleton EL, Schwartz MF, Rawson KA, Traut H, & Verkuilen J (2016). Towards a theory of learning for naming rehabilitation: Retrieval practice and spacing effects. Journal of Speech, Language, and Hearing Research, 59(5), 1111–13. 10.1044/2016_JSLHR-L-15-0303

Middleton EL, Rawson KA, & Verkuilen J (2019). Retrieval practice and spacing effects in multi-session treatment of naming impairment in aphasia. Cortex, 119, 386–400. 10.1016/j.cortex.2019.07.003 [PubMed: 31408823]

Mirman D, Strauss TJ, Brecher A, Walker GM, Sobel P, Dell GS, & Schwartz MF (2010). A large, searchable, web-based database of aphasic performance on picture naming and other tests of cognitive function. Cognitive Neuropsychology, 27(6), 495–504. 10.1080/02643294.2011.574112 [PubMed: 21714742]

Off CA, Griffin JR, Spencer KA, & Rogers MA (2015). The impact of dose on naming accuracy with persons with aphasia. Aphasiology, 30(9), 983–1011. 10.1080/02687038.2015.1100705

Oppenheim GM, Dell GS, & Schwartz MF (2010). The dark side of incremental learning: A model of cumulative semantic interference during lexical access in speech production. Cognition, 114(2), 227–252. 10.1016/j.cognition.2009.09.007 [PubMed: 19854436]

Pyc MA., & Rawson KA. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? Journal of Memory and Language, 60(4), 437–447. 10.1016/j.jml.2009.01.004

Ramsberger G, & Marie B (2007). Self-administered cued naming therapy: A singleparticipant investigation of a computer-based therapy program replicated in four cases. American Journal of Speech-Language Pathology, 16(4), 343–358. 10.1044/1058-0360(2007/038) [PubMed: 17971494]

Rapp B, & Goldrick M (2000). Discreteness and interactivity in spoken word production. Psychological Review, 107(3), 460–499. 10.1037/TO33-295X.107.3.460 [PubMed: 10941277]

Rawson KA, & Dunlosky J (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? Journal of Experimental Psychology: General, 140(3), 283–302. 10.1037/a0023956 [PubMed: 21707204]

Rawson KA, & Dunlosky J (2013). Relearning attenuates the benefits and costs of spacing. Journal of Experimental Psychology: General, 142(4), 1113–1129. 10.1037/a0030498 [PubMed: 23088488]

Rawson KA, Dunlosky J, & Sciartelli SM (2013). The power of successive relearning: Improving performance on course exams and long-term retention. Educational Psychology Review, 25(4), 523–548. 10.1007/s10648-013-9240-4

Rawson KA, Vaughn KE, Walsh M, & Dunlosky J (2018). Investigating and explaining the effects of successive relearning on long-term retention. Journal of Experimental Psychology: Applied, 24(1), 57–71. 10.1037/xap0000146 [PubMed: 29431462]

Raymer AM., Kohen FP., & Saffell D. (2006). Computerised training for impairments of word comprehension and retrieval in aphasia. Aphasiology, 20(2–4), 257–268. 10.1080/02687030500473312

Roach A, Schwartz MF, Martin N, Grewal RS, & Brecher A (1996). The Philadelphia Naming Test: Scoring and rationale. Clinical Aphasiology, 24, 121–133.

Roediger HL III, & Butler AC (2011). The critical role of retrieval practice in long-term retention. Trends in Cognitive Sciences, 15(1), 20–27. 10.1016/j.tics.2010.09.003 [PubMed: 20951630]

Roediger HL III, Putnam AL, & Smith MA (2011). Ten benefits of testing and their applications to educational practice In Psychology of learning and motivation (Vol. 55, pp. 1–36). Academic Press.

Rowland CA (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. Psychological Bulletin, 140(6), 1432–1463. 10.1037/a0037559 [PubMed: 25150680]

Sage K, Snell C, & Lambon Ralph MA (2011). How intensive does anomia therapy for people with aphasia need to be? Neuropsychological Rehabilitation, 21(1), 26–41. 10.1080/09602011.2010.528966 [PubMed: 21181603]

Schuchard J (2020, 1 28). Distributed Practice in Aphasia. Retrieved from osf.io/z2ndp

Schuchard J, & Middleton EL (2018a). The roles of retrieval practice versus errorless learning in strengthening lexical access in aphasia. Journal of Speech, Language, and Hearing Research, 61(7), 1700–1717. 10.1044/2018_JSLHR-L-17-0352

Schuchard J, & Middleton EL (2018b). Word repetition and retrieval practice effects in aphasia: Evidence for use-dependent learning in lexical access. Cognitive Neuropsychology, 35(5–6), 271–287. 10.1080/02643294.2018.1461615 [PubMed: 29724159]

Sumowski JF., Chiaravalloti N., & DeLuca J. (2010). Retrieval practice improves memory in multiple sclerosis: Clinical application of the testing effect. Neuropsychology, 24(2), 267–272. 10.1037/a0017533 [PubMed: 20230121]

Toppino TC, & Gerbier E (2014). About practice: Repetition, spacing, and abstraction In Psychology of learning and motivation (Vol. 60, pp. 113–189). Academic Press.

Vaughn KE, & Rawson KA (2011). Diagnosing criterion-level effects on memory: What aspects of memory are enhanced by repeated retrieval? Psychological Science, 22(9), 1127–1131. 10.1177/0956797611417724 [PubMed: 21813798]

Vaughn KE, & Rawson KA (2014). Effects of criterion level on associative memory: Evidence for associative asymmetry. Journal of Memory and Language, 75(C), 14–26. 10.1016/j.jml.2014.04.004

Vaughn KE, Dunlosky J, & Rawson KA (2016). Effects of successive relearning on recall: Does relearning override the effects of initial learning criterion? Memory & Cognition, 44(6), 897–909. 10.3758/s13421-016-0606-y [PubMed: 27027887]

Walker GM, & Hickok G (2016). Bridging computational approaches to speech production: The semantic-lexical-auditory-motor model (SLAM). Psychonomic Bulletin & Review, 23(2), 339–352. 10.3758/s13423-015-0903-7 [PubMed: 26223468]

Warren SF, Fey ME, & Yoder PJ (2007). Differential treatment intensity research: A missing link to creating optimally effective communication interventions. Mental Retardation and Developmental Disabilities Research Reviews, 13(1), 70–77. 10.1002/mrdd.20139 [PubMed: 17326112]

## Highlights

Principles of distributed practice operate in naming training in aphasia.

Spacing retrievals of a name across separate sessions yielded substantial benefits.

Increasing retrievals per item per session was beneficial but relatively inefficient.

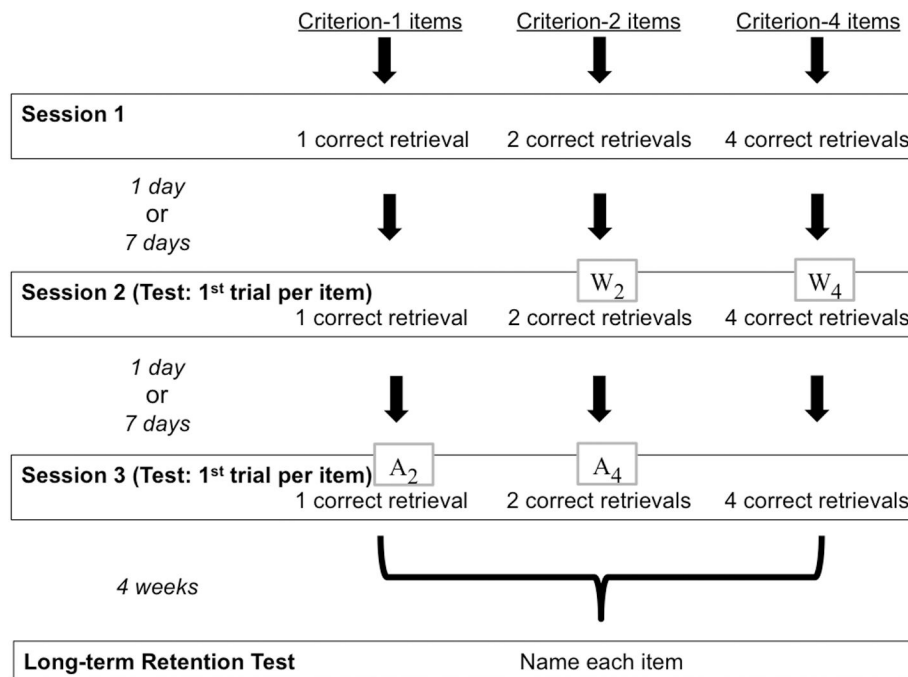A trend favored 7-day versus 1-day intersession intervals for long-term retention.

**Figure 1.**
Study design. $W_2$ and $A_2$ denote key points in the design at which each item has been named correctly a total of two times within one prior session ($W_2$) or a total of two times across two prior sessions ($A_2$). $W_4$ and $A_4$ denote key points in the design at which each item has been named correctly a total of four times within one prior session ($W_4$) or a total of four times across two prior sessions ($A_4$).
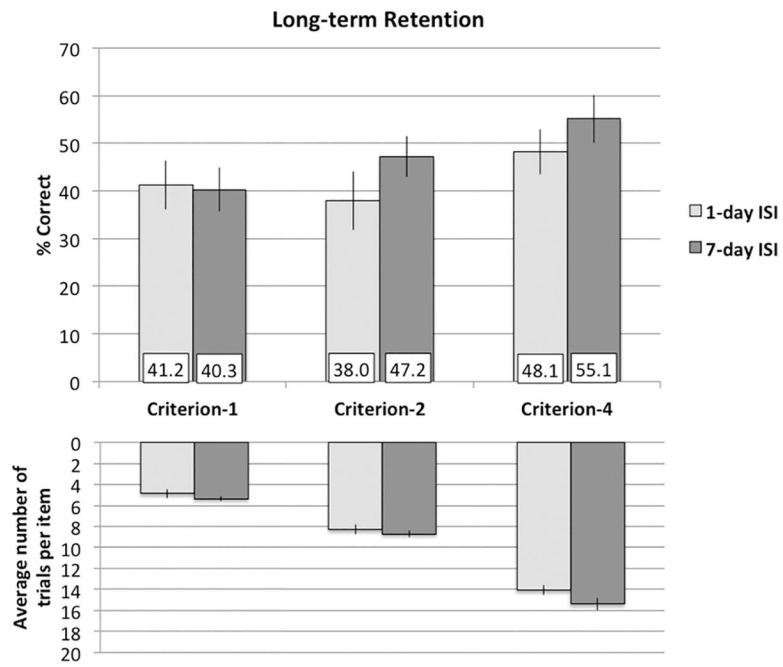
**Figure 2.**
Top graph displays average naming accuracy at the Long-term Retention Test (four weeks post-training). Bottom graph displays the average number of training trials administered per item across all three training sessions. Error bars represent standard error of the mean of the nine participants. ISI = intersession interval.
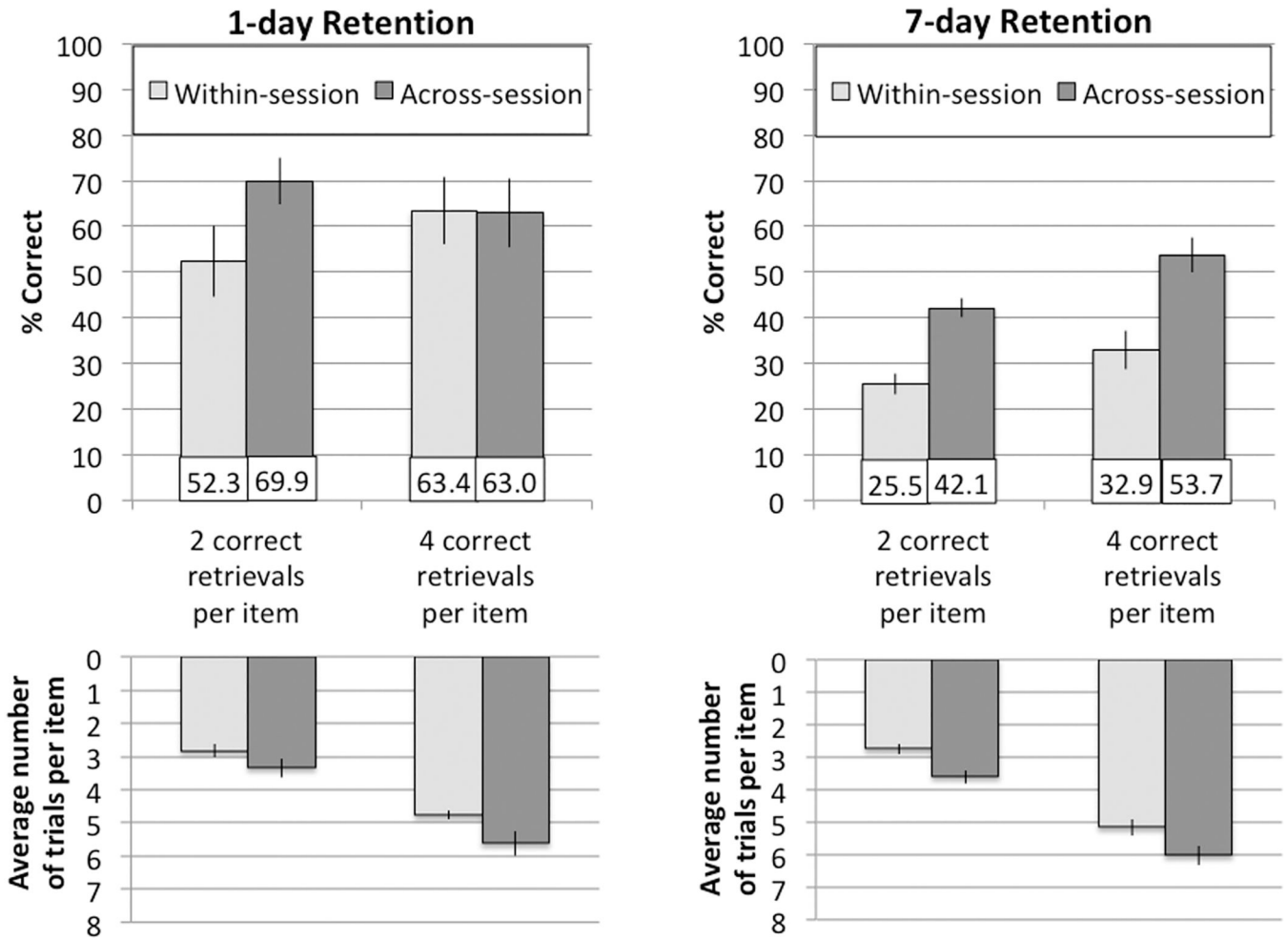
**Figure 3.**

Top graphs display average naming accuracy at key points in the design at which two different item sets had accumulated the same total number of correct responses but differed with regard to whether those responses occurred within one prior session or across two prior sessions. Bottom graphs display the average number of training trials administered per item prior to the test of naming accuracy. Error bars represent standard error of the mean of the nine participants.

**Table 1.**

Participant Demographics and Language Test Scores

| Variable/Test | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | Average | Average (controls)[a] | Cutoff[b] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Age (years) | 66 | 47 | 53 | 77 | 68 | 54 | 37 | 72 | 54 | 58.7 | | |
| Years post-onset | 6 | 3 | 7 | 1 | 1 | 6 | 2 | 2 | 11 | 4.3 | | |
| Western Aphasia Battery scores | | | | | | | | | | | | |
|   AQ | 90.4 | 88.5 | 83.2 | 91.6 | 93.6 | 81.1 | 71.6 | 92.0 | 88.3 | 86.7 | | 93.8 |
|   Subtype | A | A | A | A | A | A | TCM | A | A | | | |
|     Auditory comprehension subtest | 9.1 | 9.4 | 8.0 | 8.3 | 9.3 | 9.3 | 7.7 | 9.9 | 9.9 | 9.0 | | |
|     Repetition subtest | 8.4 | 9.4 | 8.8 | 9.5 | 8.5 | 8.2 | 8.9 | 9.2 | 9.0 | 8.9 | | |
| Apraxia of speech | none | mild-mod | none | none | none | mild | mild | none | none | | | |
| Picture naming | 85 | 94 | 85 | 91 | 96 | 74 | 87 | 82 | 77 | 85.7 | 97 | |
| Word repetition | 89 | 99 | 95 | 66 | 95 | 93 | 100 | 97 | 99 | 92.6 | 100 | |
| Nonword repetition | 37 | 57 | 77 | 23 | 68 | 43 | 48 | 65 | 82 | 55.6 | 83 | |
| STM span | 3 | 2.8 | 3.6 | 3.2 | 3 | 2.6 | 3.2 | 4.2 | 3.8 | 3.2 | 4.8 | |
| Nonverbal comprehension | 85 | 96 | 92 | 94 | 96 | 92 | 100 | 73 | 88 | 90.7 | | 90 |

Note. AQ = Western Aphasia Battery Aphasia Quotient score out of 100 (Kertesz, 1982). WAB subtests have a maximum score of 10. Subtype = aphasia subtype as determined by the WAB where A = anomic and TCM = transcortical motor. Picture naming = Percentage of correct responses on a test of oral picture naming (Philadelphia Naming Test; Roach et al., 1996). Word repetition = A test of immediate word repetition, in percentages (Philadelphia Repetition Test; Mirman et al., 2010). Nonword repetition = A test of immediate repetition of nonwords, in percentages (Philadelphia Nonword Repetition Test; Mirman et al., 2010). STM span = test of verbal short term memory where participants repeat lists of words of increasing lengths; max score = 5 (Martin et al., 1994). Nonverbal comprehension = A picture-picture association test for nonverbal semantic comprehension, in percentages (Pyramids and Palm Trees Test; Howard & Patterson, 1992).

[a] Average performance for neurotypical control sample.

[b] Scores below cutoff indicate clinically significant impairment.

**Table 2.**

Average Naming Accuracy on Initial Retrieval Practice Trials in Each Training Session

| Session | 1-day ISI | | | 7-day ISI | | |
| | Criterion-1 | Criterion-2 | Criterion-4 | Criterion-1 | Criterion-2 | Criterion-4 |
|---|---|---|---|---|---|---|
| Session 1 | 67.1 *(19.3)* | 68.1 *(15.2)* | 69.0 *(14.0)* | 63.9 *(16.4)* | 69.0 *(13.4)* | 63.0 *(19.1)* |
| Session 2 | 51.9 *(17.2)* | 52.3 *(23.2)* | 63.4 *(22.0)* | 22.7 *(13.4)* | 25.5 *(6.7)* | 32.9 *(12.6)* |
| Session 3 | 69.9 *(15.3)* | 63.0 *(22.7)* | 75.5 *(20.0)* | 42.1 *(6.1)* | 53.7 *(11.3)* | 54.2 *(17.2)* |

*Note.* ISI = intersession interval. Values in parentheses indicate standard deviations. For Session 2 and Session 3, values in the 1-day ISI phase represent one-day retention, whereas values in the 7-day ISI phase represent seven-day retention.