AMERICAN PSYCHOLOGICAL ASSOCIATION

Neuropsychology

© 2022 American Psychological Association
ISSN: 0894-4105

2022, Vol. 36, No. 8, 730–752
https://doi.org/10.1037/neu0000848

# An Examination of Retrieval Practice and Production Training in the Treatment of Lexical–Semantic Comprehension Deficits in Aphasia

Erica L. Middleton[1], Krysta L. Duquette[1], Katherine A. Rawson[2], and Daniel Mirman[3]
[1] Research Department, Moss Rehabilitation Research Institute, Elkins Park, Pennsylvania, United States
[2] Department of Psychology, Kent State University
[3] Department of Psychology, University of Edinburgh

***Objective:*** Little research has addressed the treatment of lexical–semantic comprehension deficits (i.e., difficulty retrieving the meanings of words) in people with aphasia (PWA). Research suggests that practice retrieving names for depicted objects from long-term memory (production-based retrieval practice) more strongly benefits word retrieval for production in PWA compared to errorless learning (i.e., word repetition), which eschews retrieval practice. This study assessed whether production-based and comprehension-based retrieval practice enhance performance on errorful word-comprehension items in PWA measured relative to nonretrieval forms of training and untrained control items. ***Method:*** In a within-participant group study of PWA, errorful comprehension items were assigned to (a) a production-based training module (retrieval practice vs. errorless learning); (b) a comprehension-based training module (a receptive form of retrieval practice vs. restudy). Each module comprised one training session and a 1-day and 1-week comprehension posttest on the module's trained items and an untrained item set. ***Results:*** The comprehension module conditions produced similar and superior posttest performance relative to untrained items. Both production module conditions improved posttest performance relative to untrained items, with retrieval practice conferring more durable learning and generalization indicative of refinement of semantic representations compared to errorless learning. ***Conclusions:*** Results suggest comprehension- and production-based forms of training are both beneficial for improving lexical–semantic deficits in aphasia, with production-based retrieval practice conferring additional benefits to the targeted deficit compared to errorless learning. Future studies should examine these learning factors in schedules of training more commensurate with clinical practice and in other neurological populations (e.g., semantic dementia).

***Key Points***
***Question:*** Do both comprehension-based and production-based forms of training improve lexical–semantic deficits in people with stroke aphasia (i.e., problems understanding words), and does retrieval practice (retrieval of target information from long-term memory) enhance the training benefit? ***Findings:*** Lexical–semantic deficits in people with aphasia are improved with both comprehension- and production-based forms of training, with production-based retrieval practice conferring enhanced durability of learning and generalized improvement relative to errorless learning, a form of practice that eschews retrieval practice. ***Importance:*** This work sets the stage for research aimed at understanding which forms of treatment to prioritize when a patient with aphasia demonstrates problems with both production and comprehension. ***Next Steps:*** Future studies should further examine the clinical relevance of these learning factors for treating lexical–semantic deficits by adopting schedules of training more commensurate with clinical practice.

*Keywords:* retrieval practice, lexical–semantic treatment, aphasia, transfer of learning, naming

This article was published Online First September 1, 2022.

Erica L. Middleton https://orcid.org/0000-0002-2507-1273

The authors thank Myrna F. Schwartz and members of the Language and Learning Lab for important discussions regarding the design of the present study. The authors thank Taylor Foley and Adelyn Brecher for their assistance with data collection, coding, and neuropsychological characterization. A portion of this work was presented at the 59th Annual Meeting of the Academy of Aphasia (virtual conference, October 2021).

This work was supported by the National Institutes of Health (Grant R01 DC015516) awarded to Erica L. Middleton. The authors declare no potential conflicts of interest.

Analysis code and data are available at https://osf.io/wbs6c/?view_only=dc607edf4c1745009bc0e7ed4e300a9f.

Erica L. Middleton played lead role in conceptualization, formal analysis, funding acquisition, methodology, project administration, resources, supervision, visualization and writing of original draft and supporting role in data curation, investigation and software. Krysta L. Duquette played lead role in data curation, investigation and software and supporting role in methodology, project administration, visualization and writing of original draft. Katherine A. Rawson played supporting role in conceptualization, investigation, methodology, resources and writing of original draft. Daniel Mirman played supporting role in formal analysis and writing of review and editing.

Correspondence concerning this article should be addressed to Erica L. Middleton, Research Department, Moss Rehabilitation Research Institute, 50 Township Line Road, Elkins Park, PA 19027, United States. Email: middleer@einstein.edu

There is growing interest in translating from basic research on learning and memory to explicate the treatment process for cognitive and language deficits (for reviews, see Clare & Jones, 2008; Dignam et al., 2016; Fillingham et al., 2003; Middleton & Schwartz, 2012; Oren et al., 2014). A growing body of work has examined the application of the vast psychological literature on retrieval practice effects (a.k.a. test-enhanced learning) to inform naming treatment in aphasia (e.g., Friedman et al., 2017; Middleton et al., 2015, 2016, 2019; Schuchard et al., 2020; Schuchard & Middleton, 2018a, 2018b; for recent review, see de Lima et al., 2020). Retrieval practice, the act of retrieving information from long-term memory, strengthens future access to that information. In the present study, we take a next important step in examining the clinical relevance of retrieval practice in aphasia rehabilitation by studying its application to *word-comprehension deficits* in aphasia, that is, problems reliably accessing the meaning of words.

## Word-Comprehension Deficits in Aphasia and Their Treatment

Word-comprehension deficits exist to some degree in most people with aphasia (PWA) but are understudied because of under-characterization by formal aphasia batteries and underemphasis in light of the typically more obvious spoken output deficits (Morris & Franklin, 2017). Word-comprehension problems can arise from deficits involving speech sound recognition, word recognition, or meaning retrieval (for reviews, see Morris & Franklin, 2017; Semenza, 2020). The targeted deficit in the present study is *lexical–semantic deficit*, or difficulty retrieving the meanings of words. Lexical–semantic deficit in aphasia has been described as arising from weak or noisy connections from words to semantics (e.g., for review, see Mirman & Britt, 2014), damage to central semantic representations (e.g., Hillis et al., 1990), or dysregulated control of semantic processes (e.g., Jefferies & Lambon Ralph, 2006). We accept potential contributions from any of these sources in our study sample. The behavior of interest in the present study pertains to participants' ability to discriminate subtle semantic distinctions between closely related word pairs (i.e., *semantic minimal pairs*, e.g., spider vs. scorpion, goggles vs. sunglasses; pairs are listed in Appendix Table A1) in a word–picture verification (WPV) task (see Figure 1).

As argued by Nickels (2000), semantic-based treatments are different from treatments that target semantics. The typical semantic-based treatment involves practice selecting from an array of pictures to match a stimulus (e.g., a word or picture). In a recent review of studies of semantic-based treatments for acquired language impairment (Casarin et al., 2014), most of the treatments aimed to improve word production rather than semantic processing. Some notable exceptions are reviewed by Nickels (2000; see also Knollman-Porter et al., 2018; Morris & Franklin, 2012). Joining this small literature, the present study examines the effects of a semantics-targeted treatment on word comprehension, along with whether and how production training confers improvements in word comprehension in aphasia (i.e., task transfer).

## Retrieval Practice and Aphasia Treatment

In research on human learning and memory, enumerable studies have documented *retrieval practice effects*, in which practice retrieving from memory (e.g., practice retrieving the target "frog" from the cue "pond") confers superior performance on later tests than *restudy* (e.g., studying the word pair *pond-frog*; for recent reviews, see Kornell & Vaughn, 2016; Rawson & Dunlosky, 2011, 2013; Roediger & Butler, 2011; Roediger et al., 2011; Rowland, 2014). Retrieval practice can take the form of cued recall (as in the example above), or free recall, such as practice retrieving words in a studied list. Retrieval practice can also take the form of a recognition test (e.g., discriminate previously studied words from new words) or multiple choice.

Recent research has also shown that retrieval practice is beneficial for naming impairment, a ubiquitous disorder in aphasia that manifests as difficulty reliably and fluently retrieving familiar words for production. In the first demonstration of this benefit (Middleton et al., 2015), for each of eight PWA in a within-participant design, errorful naming items were administered for one trial of retrieval practice in which the participant attempted to name the picture with a cue (i.e., word onset was provided) or without a cue. These conditions were compared to errorless learning, in which at picture onset, the object's name was presented and the participant repeated the name. Both types of retrieval practice outperformed errorless learning at a next-day test of naming, with the advantage persisting for the cued retrieval practice condition after 1 week. Retrieval practice effects in aphasic naming have since been observed when items are trained in multiple trials within a single session (Middleton et al., 2016), when items are trained in multiple trials in multiple sessions (Middleton et al., 2019), and when items are first trained to mastery followed by retrieval practice versus restudy, with retrieval practice conferring more durable benefits (Friedman et al., 2017). Schuchard and Middleton (2018a, 2018b) showed that these effects arise because, compared to errorless learning, retrieval practice is more effective at strengthening the mapping from semantics to words.

Of major interest in the present work is (a) whether the benefits from naming training extend beyond production, that is, whether naming training improves comprehension, and (b) if retrieval-based naming practice confers superior benefits to word comprehension compared to errorless learning. Both of these possibilities, to our knowledge, have yet to be examined in aphasia. From a theoretical standpoint, we may expect a relative advantage for retrieval practice over errorless learning for improving word-comprehension performance in PWA. This could result from targeted strengthening of the mapping from semantics to words if there is some degree of overlap in the processes that pull from the lexicon in the course of both comprehension and production (e.g., Dell & Chang, 2014). Additionally or otherwise, retrieval practice could be more effective at refining semantic distinctions with implications for comprehension, based on an assumption of overlap in the semantic system(s) underpinning comprehension and production (Chen et al., 2017; Gambi & Pickering, 2017; Pylkkänen, 2019). From a clinical standpoint, examining the impact of different kinds of production training on word comprehension is important because it can reveal whether the benefits of a retrieval-based naming treatment extend beyond naming impairment.

We also examined a receptive version of retrieval practice training and compared it to a receptive training control condition (restudy). Though retrieval practice effects are generally weaker with receptive tests such as multiple choice or recognition (Rowland, 2014), this comparison is of interest given the novelty of the manipulation in this applied domain. Furthermore, given the paucity of research on different treatment methods for lexical–semantic deficits (Nickels, 2000), the current design contributes to an evidence base particularly relevant for people with profound aphasia such as those who are nonverbal and thus, the prioritization of comprehension-based training may be more appropriate than production-focused practice.

## Overview of Current Research

The behavior of interest in the present study concerned a PWA's ability to discriminate a target and foil comprising a *minimal semantic pair* in a WPV task (Hillis et al., 1990; Rapp & Caramazza, 2002). Discrimination of a minimal pair required both accepting as correct a target picture (*van*) with the target word ("van") and rejecting a semantically related foil picture (*bus*) as correct for the target word ("van") on different, nonconsecutive trials. This task is designed to be sensitive to even subtle semantic deficits. First, requiring both acceptance of the target and rejection of the foil assures the test cannot be completed by superficial semantic processing (Breese & Hillis, 2004). Second, the difference between targets and foils centers on distinctive features between closely related category members (e.g., backpack vs. lunchbox, ant vs. cricket, scarf vs. tie).

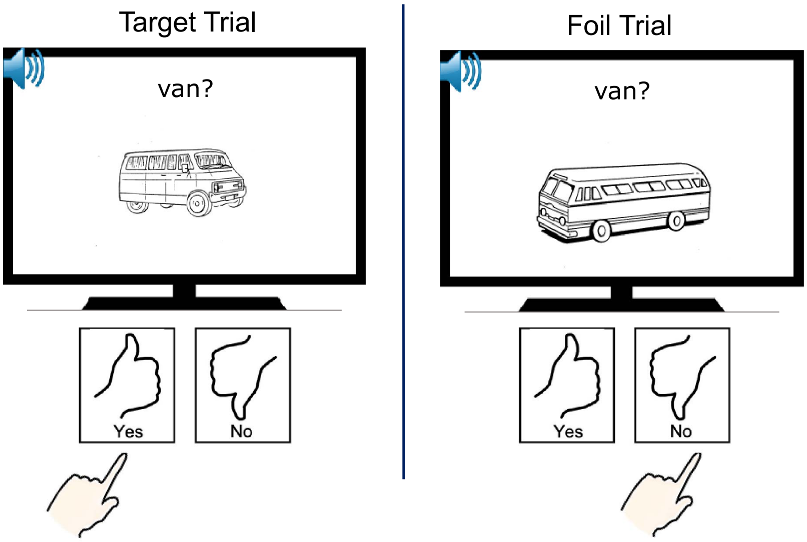
Twelve PWA completed the study, which employed a within-participant design. The primary focus was on results across the group. Prior to training, each PWA engaged in a WPV task for a large set of minimal semantic pairs developed in our lab (see Figure 1 and Appendix Table A1). From this *item selection phase*, a PWA's errorful items were assigned into the different conditions. The study design comprised two training modules (comprehension training and production training; see Figures 2–4) administered in different weeks. Each module comprised one training session and then tests administered the next day and 1 week following training (hereafter, *retention tests*) to probe WPV performance on trained and untrained items for that module.

Items assigned to the production module were presented for noncued retrieval practice versus errorless learning as in Middleton et al. (2015, 2016, 2019; Figure 3). Items assigned to the comprehension module were trained with a receptive form of retrieval practice versus restudy (Figure 2). All 12 participants engaged in the comprehension module. Of the 12 participants, eight demonstrated sufficiently errorful performance during item selection to populate the training conditions in the production module, as well (Figure 4). A retrieval practice effect would be reflected in superior performance at the retention tests for the retrieval practice condition compared to the control training condition (errorless learning or restudy) in each module. Retrieval practice effects arise because retrieval practice more durably strengthens target information compared to restudy (for discussion, see Roediger & Karpicke, 2006). If so, retrieval practice effects may be more pronounced as the memory demands of a retention test increase (e.g., with longer retention intervals). Thus, in the present design, retrieval practice effects may be stronger at the longer retention test (1-week vs. 1-day). A more direct way to assess the durability of learning is to measure forgetting, that is, loss of accuracy between test timepoints. The durability of learning will be assessed by modeling the proportion of correct responses at the 1-day test that persist as correct at the 1-week test in the different conditions within each module (Mettler et al., 2016).

Generalized improvement in a task or skill from treatment is often measured in aphasia studies by probing performance on

**Figure 1**

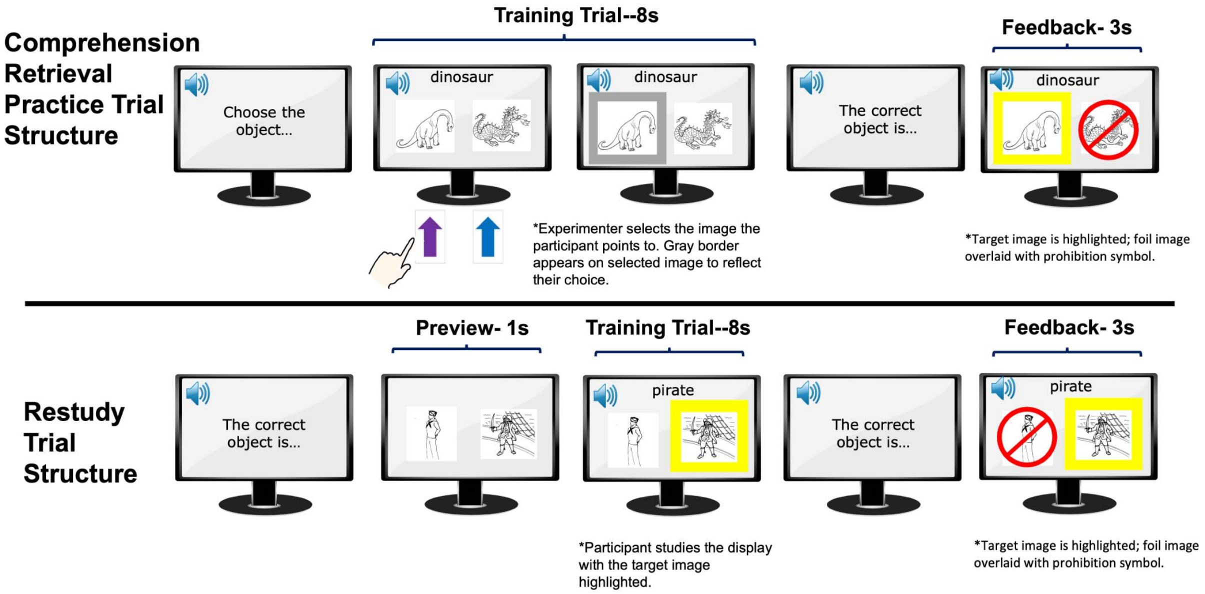
*Word–Picture Verification Was Employed During Item Selection and at the Retention Tests*



*Note*. Participant responses were made on Yes and No cards on the desk in front of them. The participant had 20 s to make their response, which was inputted by the experimenter. An accurate word–picture verification response for the item ("yes" to target image and "no" to foil image) is shown in this example (see Item Selection and Item Assignment section, for more details). Target and foil images from *The Philadelphia Naming Test* by Moss Rehabilitation Research Institute (https://mrri.org/philadelphia-naming-test/). Copyright 1996 Moss Rehabilitation Research Institute. Reprinted with permission. See the online article for the color version of this figure.

**Figure 2**

*Comprehension Training Module Trial Structure Showing the Sequence of Events Displayed on the Computer Screen (Left-to-Right Temporal Order) per Training Type*



*Experimenter selects the image the participant points to. Gray border appears on selected image to reflect their choice.

*Target image is highlighted; foil image overlaid with prohibition symbol.

*Participant studies the display with the target image highlighted.

*Target image is highlighted; foil image overlaid with prohibition symbol.

*Note.* Comprehension retrieval practice training involved asking the participant to choose between the target and foil picture given the target name. For restudy, both the target and foil picture were briefly previewed (1 s) after which the target word was presented and the target picture was identified for the participant. Training trials ended in correct-answer feedback. A sound icon indicates auditory presentation of text (see Comprehension Training Trials section, for more details). Target and foil pictures from *The Philadelphia Naming Test* by Moss Rehabilitation Research Institute (https://mrri.org/philadelphia-naming-test/). Copyright 1996 Moss Rehabilitation Research Institute. Reprinted with permission. See the online article for the color version of this figure.
* Indicates details specific to that event.

untrained items. However, inferring generalization from improved performance on untrained items can be problematic (for discussion, see Howard et al., 2015; Webster et al., 2015). In the present study, the untrained items served as a reference condition against which improvements on the trained items were assessed in order to establish direct treatment effects while controlling for improvements that can arise from other factors, for example, enhanced familiarity with the task. However, in the present study, we will measure generalization in the form of (a) task transfer, or improved comprehension performance from production training, and (b) improvements on foils in the production module. In contrast to comprehension training, during production training, only the target items of target–foil pairs are experienced. Thus, changes in performance on foils in the production module will reflect semantic refinement, which may be greater for retrieval practice versus errorless learning because of greater engagement of the lexical–semantic stage of mapping.

Last, the present study is situated along a research pipeline that ultimately seeks to inform and optimize clinical practice. Our aim is to investigate the applicability of retrieval practice learning factors to a clinical problem, that is, lexical–semantic word-comprehension deficits in aphasia. To do this, we adopted a multipronged strategy developed by Middleton et al. (2015, 2016, 2019, 2020). This involves manipulating aspects of common treatment experiences in controlled, experimental comparisons to provide "proof of concept" of the relevance of retrieval practice factors in the present domain. We invited PWA for participation who generally exhibited the targeted deficit to test the theory that retrieval practice impacts the mapping from words to semantics. Furthermore, the sample was selected to be relatively homogeneous to reduce variability due to comorbidities. Because of the selection of this relatively small group of PWA, the study was designed to maximize the number of observations per participant per condition, to enhance experimental sensitivity. Given the substantial resources required to study each participant, we aimed for a study sample size similar or greater to that in prior proof-of-concept studies of retrieval practice effects in aphasia treatment ($N$ = 8 PWA in Middleton et al., 2015; $N$ = 4 PWA in Middleton et al., 2016). Though in the resulting design, each item was treated a small number of times in its assigned condition, conclusions based on the present work will inform later phases of research examining dosage levels and retention intervals more commensurate with current clinical practice.

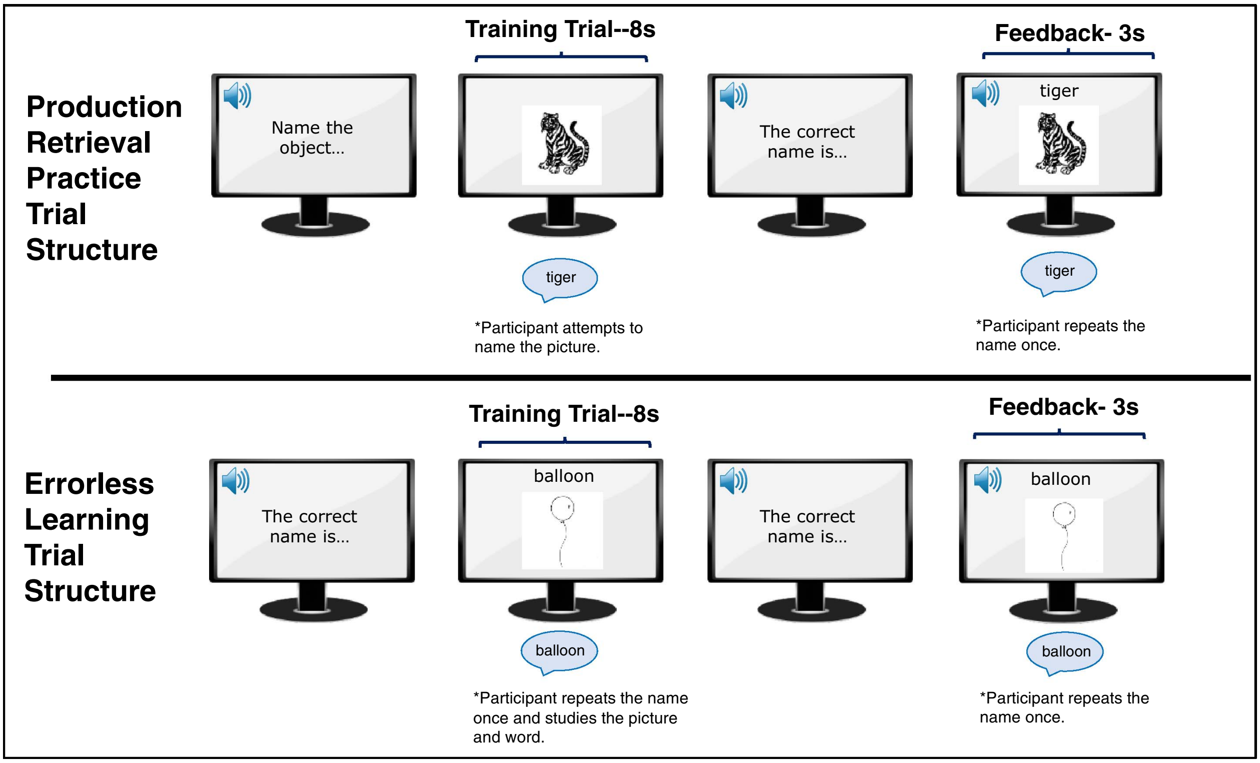
# Method

## Participants

The participants were 12 adults (5 female) with chronic aphasia secondary to left-hemisphere stroke. See Table 1 for demographic and neuropsychological traits of the participants. Reflective of the demographics of the larger metropolitan area of Philadelphia served by Moss Rehab, our sample included five Black participants and seven White participants. All participants consented to study protocol 4526EXP "Word retrieval in aphasia" approved by the institutional review board of Einstein Healthcare Network. Eleven participants were able to provide informed consent; a

**Figure 3**

*Production Training Module Trial Structure Showing the Sequence of Events Displayed on the Computer Screen (Left-to-Right Temporal Order) per Training Type*


Training Trial--8s
Feedback- 3s
Production Retrieval Practice Trial Structure
Name the object...
The correct name is...
tiger
tiger
tiger
*Participant attempts to name the picture.
*Participant repeats the name once.
Training Trial--8s
Feedback- 3s
Errorless Learning Trial Structure
The correct name is...
balloon
The correct name is...
balloon
balloon
balloon
*Participant repeats the name once and studies the picture and word.
*Participant repeats the name once.

*Note.* Production retrieval practice training involved asking the participant to attempt to name the picture. For errorless learning, target name and target picture were simultaneously displayed, and the participant repeated the name. Training trials ended in correct-answer feedback. A sound icon represents auditory presentation of text. Callout graphic represents oral response by the participant (see Production Training Trials section, for more details). Target pictures from *The Philadelphia Naming Test* by Moss Rehabilitation Research Institute (https://mrri.org/philadelphia-naming-test/). Copyright 1996 Moss Rehabilitation Research Institute. Reprinted with permission. See the online article for the color version of this figure.
* Indicates details specific to that event.

12th participant, who was determined to have diminished decisional capability, was consented by their legally authorized representative. Participants were paid $15 per hour of testing.

## *Prescreening and Selection of Participants*

Participants were recruited from a large pool of available research volunteers (>100) with stroke aphasia who had undergone extensive cognitive and linguistic characterization. The study design required that each participant produce a sufficient number of errors in the item selection task to populate the conditions for at least one training module (i.e., 30 errors). All participants except one[1] met this threshold. From the large pool of available research volunteers, we prioritized recruitment of individuals who showed notable impairment in one or more measures of semantic comprehension or word comprehension because we anticipated such individuals would produce a sufficient number of errors at item selection. This guided our recruitment approach except for one case (P3), who was invited to participate because he communicated a strong desire for additional research participation after completion of the cognitive–linguistic test battery. Despite mild impairment on semantic tasks, P3 produced a sufficient number of errorful items during item selection to participate. All participants passed a hearing assessment appropriate for their age group (i.e., below or above age 65) except one individual (P5), who was not asked to complete test battery tasks that solely rely on auditory input. All participants responded correctly to 75% or greater ($M = 95\%$; $SD = 7.4\%$) of the 20 questions on the yes/no questions portion of the comprehension subtest of the Western Aphasia Battery (WAB; Kertesz, 2007).

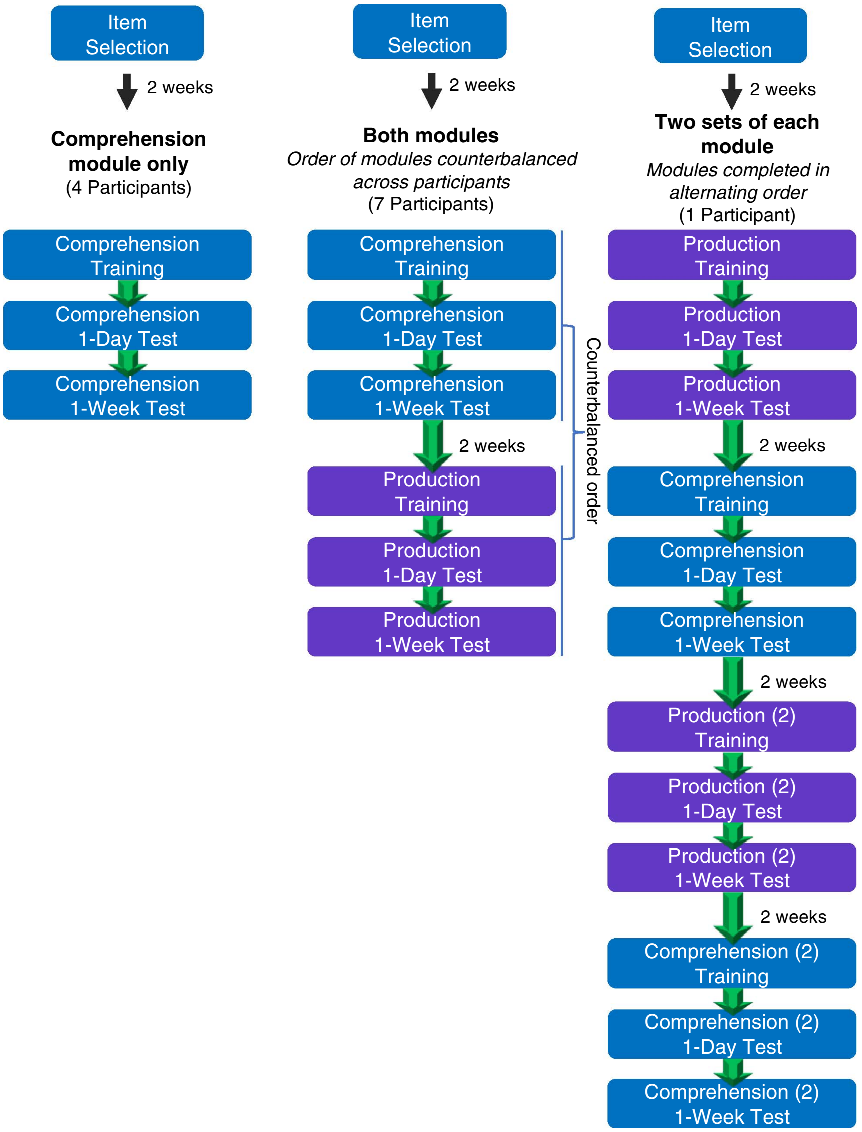
## *Neuropsychological Profile of the Participant Sample*

Neuropsychological characterization of our sample points to lexical–semantic origins of their single-word-comprehension problems (i.e., problems mapping to the meaning of words, and/or in central semantic processing). As shown in Table 1, compared to a reference sample of 262 volunteers with aphasia subsequent to left-hemisphere stroke, our participants generally performed below the larger aphasia sample mean on one of the measures of verbal semantic comprehension (Peabody Picture Vocabulary Test, Dunn & Dunn, 2007; synonym matching task, Saffran et al., 1988) or nonverbal semantic comprehension (Camel & Cactus Test, Bozeat et al., 2000).

[1] One participant produced only 29 errorful items during item selection, but they were invited to continue the study. One correct item was randomly selected to fill the design for that participant.

**Figure 4**

*Experiment Timeline for 12 Participants*



*Note*. The number of modules completed by a participant determined by number of errorful items during item selection. Four participants completed the comprehension training module only. Seven completed both the comprehension and production modules in counterbalanced order. One participant had sufficient errorful items from item selection to populate two cycles of the comprehension and production modules. See the online article for the color version of this figure.

Available scores (mean/standard deviation) for 20 neurotypical controls on the receptive tests are also provided in Table 1 as a benchmark for unimpaired performance. Comparison to those scores reveals that multiple PWA (P1, P2, P9) showed severe impairment (3 *SD* or more lower than the control sample) on the nonverbal semantic comprehension and synonymy matching tasks, indicative of multimodal semantic impairment. Furthermore, faulty phonological input processing and word recognition were likely not major contributors to low performance on the verbal semantic comprehension measures because most participants exhibited scores around or above the reference sample mean for phoneme discrimination and auditory lexical decision, and because speech perception impairment tends to have little impact on standard word-comprehension measures (e.g., Basso et al., 1977; Blumstein, Baker, & Goodglass, 1977; Blumstein, Cooper, et al., 1977; Dial & Martin, 2017; Miceli et al., 1980). Word repetition ability, which is sensitive to word recognition deficits, was mildly or very mildly impaired with a few exceptions; those with more severe word repetition deficits tended to also produce high rates of phonological errors in naming, indicating a phonological output rather than input problem.

Furthermore, among the input measures in Table 1, Spearman rank correlations revealed that performance in the WPV task we

**Table 1**
*Demographic and Neuropsychological Characteristics of the Participants*

| Variable/test | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 | P12 | *M* | Reference sample (*M*/*SD*) | Controls sample (*M*/*SD*) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Age (years) | 76 | 69 | 73 | 62 | 66 | 74 | 69 | 58 | 55 | 49 | 54 | 56 | 63 | | |
| Gender | M | F | M | M | F | M | F | M | M | F | F | M | | | |
| Ethnicity | W | B | W | W | B | W | B | W | W | W | B | B | | | |
| Occupation | Court clerk | Telecommunications manager | Accountant | Police officer | Care manager | Teacher | Housekeeper | Carpenter | Attorney | Teacher | Certified nursing assistant | Nurse | | | |
| Years education | 18 | 13 | 19 | 12 | 14 | 19 | 11 | 12 | 19 | 18 | 13 | 14 | | | |
| Months postonset | 37 | 20 | 63 | 13 | 8 | 90 | 48 | 56 | 38 | 13 | 135 | 127 | 54 | | |
| Handedness | R | R | R | R | R | R | R | L | R | L | R | R | | | |
| Western Aphasia Battery scores | | | | | | | | | | | | | | | |
| Aphasia subtype | A | A | A | C | A | TCM | B | B | A | A | A | A | | | |
| AQ | 92.7 | 89.2 | 91.2 | 73.8 | 90.1 | 69.8 | 68.6 | 46.7 | 80.7 | 79.4 | 88 | 89.7 | 80 | | |
| Philadelphia Naming Test scores | | | | | | | | | | | | | | | |
| % accuracy | 79 | 83 | 70 | 64 | 66 | 70 | 69 | 35 | 43 | 74 | 70 | 87 | 68 | | |
| % sem. errors | 7 | 7 | 2 | 8 | 14 | 2 | 6 | 16 | 18 | 7 | 5 | 4 | 8 | | |
| % phono. errors | 9 | 3 | 22 | 21 | 3 | 21 | 13 | 22 | 18 | 14 | 25 | 9 | 15 | | |
| Auditory lexical decision task d′ | 3.2 | 2.1 | 2.3 | 2.9 | N/A | 3.4 | 1.7 | 2.7 | 3.2 | 2.7 | 2.7 | 2.3 | 2.7 | 3/.74 | |
| Phoneme discrimination | 85 | 90 | 85 | 93 | N/A | 98 | 88 | 93 | 98 | 93 | 88 | 83 | 90 | 89/10 | 97/3 |
| Word repetition | 91 | 97 | 79 | 83 | N/A | 85 | 83 | 69 | 98 | 90 | 57 | 95 | 84 | 84/19 | |
| Semantic comprehension | 67 | 69 | 91 | 89 | 77 | 78 | 77 | 81 | 66 | 84 | 73 | 77 | 77 | 75/14 | 90/6 |
| Receptive vocabulary ability | 87 | 71 | 91 | 91 | 72 | 92 | 66 | 78 | 75 | 79 | 70 | 63 | 78 | 75/17 | |
| Synonym matching | 67 | 67 | 97 | 97 | 70 | 100 | 60 | 70 | 57 | 73 | 77 | 70 | 75 | 79/16 | 97/9 |
| Word–picture verification task accuracy | 21 | 59 | 88 | 91 | 52 | 86 | 55 | 30 | 50 | 84 | 73 | 64 | 63 | | |

*Note*. Occupation = occupation at time of stroke; Aphasia subtype = aphasia subtype as determined by the Western Aphasia Battery–Revised (Kertesz, 2007); A = anomic; B = Broca's; C = conduction; TCM = transcortical motor; AQ = Western Aphasia Battery Aphasia Quotient, score out of 100. Philadelphia Naming Test scores, scores on an oral picture naming test (Roach et al., 1996), with % accuracy = percent correct responses; % sem. errors = percent semantic-based errors consisting of morpheme omissions, semantic errors, and mixed errors, with or without phonological errors; % phono. errors = percentage of phonological errors on the Philadelphia Naming Test, consisting of real-word errors and nonword errors; auditory lexical decision task d′ = measure of discrimination on an auditory word judgment task from the *Psycholinguistic Assessments of Language Processing in Aphasia* (Kay et al., 1992); phoneme discrimination = percentage of correct responses on the Auditory Discrimination Test, an assessment of minimal pair phoneme perception (Martin et al., 1994); word repetition = percentage of correct responses on the Philadelphia Repetition Test, a test of oral repetition of words (Dell et al., 1997); semantic comprehension = percentage of correct responses on the Camel and Cactus Test, a picture–picture association test measuring nonverbal semantic comprehension (Bozeat et al., 2000); receptive vocabulary ability = percentage of correct responses on The Peabody Perceptive Vocabulary Test, a word–picture association test assessing receptive vocabulary (Dunn & Dunn, 2007); synonym matching = percentage of correct responses on the Synonymy Triplets Test, which requires selecting 2 of 3 written and spoken synonyms presented in triads (Saffran et al., 1988); reference sample (*M*/*SD*) = mean (standard deviation) for 262 volunteers with aphasia subsequent to left-hemisphere stroke on select tests of the cognitive–linguistic test battery; controls sample (*M*/*SD*) = mean (standard deviation) for 20 neurotypical controls. Scores for any auditory-only task are not reported for P5 due to hearing loss; word–picture verification accuracy = % correct during item selection (see Item Selection and Item Assignment section, for details).

used for item selection was strongly associated with measures of lexical–semantic processing and not with measures of phonological processing. Specifically, there was a strong, positive association between WPV item selection accuracy (see Item Selection and Item Assignment section, for details) and the semantic comprehension measure, $r(10) = .68$, $p = .02$, two-tailed, and synonym matching, $r(10) = .81$, $p = .002$, two-tailed, but not with tests tapping input phonology and word recognition: auditory lexical decision, word repetition, and phoneme discrimination (all $ps > .38$). The receptive vocabulary test (Peabody Vocabulary Test; Dunn & Dunn, 2007), which is likely influenced by premorbid verbal ability, also did not correlate with WPV accuracy ($p = .24$). Overall, these correlations provide evidence that our WPV procedure taps lexical–semantic processing, or more specifically, problems making refined verbal and nonverbal semantic distinctions, rather than deficits in phoneme or word recognition. Last, as described in the Procedure and Design section, we attempted to minimize the contribution of speech sound or word recognition processes by providing target words in both auditory and written form and allowing participants to complete trials at their own pace.

## Materials

### Image Properties

The materials involved a corpus of 816 pictures constituting 408 minimal semantic pairs of common, everyday objects collected from published image corpora and various internet sources (Brodeur et al., 2010, 2014; Duñabeitia et al., 2018; Roach et al., 1996; Rossion & Pourtois, 2004; Snodgrass & Vanderwart, 1980; Szekely et al., 2004). The corpus included black-and-white pictures and color pictures in the format of drawings/graphics and photos. One of the two images in each minimal semantic pair was designated as the target image (hereafter, target), which matched the target name, and the other image was designated as the foil image (hereafter, foil). During stimuli development, effort was made to match each pair of target–foil images in terms of format (drawing vs. photo) and color status (color vs. black-and-white); 92% of pairs matched on both dimensions. For a full list of the 408 minimal semantic pairs, see Appendix Table A1.

### Norming Studies for Development of Materials

Early phases of development of the corpus involved iterative cycles of pair development and norming. The goal was to develop a large set of minimal semantic pairs in which each target and foil concept were known to American English speakers generally as reflected in high name agreement and/or high familiarity.

In cases in which values were not already available from published image corpora (Brodeur et al., 2010, 2014; Rossion & Pourtois, 2004; Szekely et al., 2004), variables were gathered for the final set of images in normative studies, including visual complexity of each image (values range from 1 = *very simple* to 5 = *very complex*), and visual similarity of each target–foil image pair (values range from 0 = *no similarity at all* to 10 = *very similar*; De Groot et al., 2016). Correlation values between the target and foil name in each pair were estimated with latent semantic analysis (LSA; Landauer et al., 1998) to provide a measure of semantic relatedness. The second author (a female native speaker of mainstream American English) created audio recordings of all target names.

It was important for this project to verify that errors on the WPV task in participants with aphasia were not attributable to issues with the stimuli, for example, poorly rendered or unrepresentative pictures for the target or foil concept. In two waves of norming, 15 or more neurotypical adults provided WPV responses to the targets and foils in pseudorandom order. Following the first wave, items that were associated with an accuracy rate below 95% were eliminated and a second set of pairs was added to the task for the second wave of norming. All images in the final 408-item set were associated with 95% or greater accuracy in the normative sample ($M = 0.99$; $SD = 0.02$).[2]

### Target Name Properties

All target names were nouns comprising a range of word length in phonemes (2–15), syllables (1–5), and letters (2–16). The number of syllables, letters, and phonemes were collected from the Merriam-Webster Online Dictionary (www.merriam-webster.com), with trained transcribers reaching consensus regarding regional dialect with respect to alternative pronunciations as necessary (e.g., /braa-kuh-lee/ or /braa-klee/, for the target broccoli). Frequency values for all target names were collected from the SUBTLEX$_{us}$ project (Brysbaert & New, 2009). Name frequency, number of syllables, phoneme length, and orthographic length were only collected for the target names, the rationale being that the participants were never exposed to the name of the foil in any phase of the experiment.

## Procedure and Design

### Overview

Participants began the study by completing the WPV item selection task, with the full set of 408 pairs (target and foils) administered twice for WPV (Figure 1). After a minimum of 2 weeks, the participant engaged in one of the two modules. Each module required three sessions (training session, 1-day retention test, 1-week retention test; see Figure 4). Whether a PWA participated only in the comprehension module or in both the production and comprehension modules depended on the number of errorful items during item selection.[3] For those who only had enough errorful items for one module, completion of the comprehension module was prioritized because of the novelty of examining a receptive version of retrieval practice (vs. restudy) in PWA. All 12 PWA completed the comprehension training module; eight participants additionally completed the production module. For the participants who completed both modules, the order of the two modules was counterbalanced across participants, with a minimum of 2 weeks between the modules. All stimuli were presented using the E-Prime software on a PC desktop or laptop. All sessions were audio-taped, and all sessions except production training were video-recorded.

[2] Due to experimenter error, a small number of images (9, or 1.1% of 816) in the final 408-item set were not included in the WPV task norming.

[3] In the item sets for three participants, a small number of correct items from item selection were required to populate the design to achieve an even number of observations per condition. These correct items were assigned into the conditions in a matched fashion.

## Item Selection and Item Assignment

In the WPV item selection task, the 408 pairs (targets and foils) were administered twice for WPV during item selection to increase the set of candidate errorful items per participant. Each of the two administrations of the 408 pairs required one or two sessions per participant. A pair was considered an errorful item if an error or response omission was made on either of the target trials or either of the foil trials; the pair was then a candidate for use in the module(s).

In each administration of the 408 pairs at item selection for a participant, items were divided into two blocks. The first block included targets for a randomly selected half of the 408 pairs and foils for the other half of items, presented in random order. The second block included the remaining targets and foils, presented in random order. This randomization procedure ensured that a target and its foil appeared in different blocks and thus did not appear in contiguous trials; this resulted in an average separation of 407 trials between a target and its foil (range 1–811; median: 408). Each item selection session began with six practice trials of items that were not in the 408-pair corpus. On each WPV trial, the (target or foil) image was accompanied by the target name in written and auditory form (see Figure 1, for trial structure). Participants were instructed to judge whether the picture matched the name presented. Anticipating the possibility of studying PWA on the extreme end of impairment (e.g., impulsivity; hemiparesis), the mode of response during item selection involved asking the PWA to indicate their response by pointing to a Yes card (with a thumbs-up graphic) versus a No card (with a thumbs-down graphic) on the desk in front of them (“yes” indicated the image and the word matched; “no” indicated the image and word did not match). The participant was given 20 s to respond. If the participant responded, the experimenter registered the response and advanced the trial; if the participant did not respond, the trial advanced on its own after 20 s.

Errorful pairs for each participant were pseudorandomly assigned to the modules and the conditions while matching for target word frequency, number of letters, number of phonemes, number of syllables, target and foil image complexity, target and foil visual similarity, and target and foil semantic similarity. In the comprehension module, the number of observations per the retrieval practice, restudy, and untrained conditions ranged from 10 to 50 across participants ($M$ = 25.83). All participants completed the comprehension module, and eight additionally completed the production module; one participant (P1) produced enough errors during item selection to populate two cycles for each of the comprehension and production modules (see Figure 4). In the production module, the number of observations per the retrieval practice, errorless learning, and untrained conditions ranged from 20 to 50 across participants ($M$ = 30). Averages of the variables across the participants’ item sets in the different conditions and modules are displayed in Table 2.

## Training Sessions

During the training session in a module, a participant completed between 1 and 4 blocks of items, depending on the number of errorful items identified at item selection. In a training block, the two types of training trials in that module were intermixed in the block. A single training block in either module consisted of 90 trials, with

**Table 2**

*Mean Item Characteristics per Condition and Module Across Participants’ Item Sets*

| | | Log-word frequency[a] (target name) | No. of letters (target name) | No. of phonemes (target name) | No. of syllables (target name) | Visual complex (target image)[b] | Visual complex (foil image)[b] | Visual similarity[c] | Semantic related[d] |
|---|---|---|---|---|---|---|---|---|---|
| Module | Condition | *M* (*SD*) | *M* (*SD*) | *M* (*SD*) | *M* (*SD*) | *M* (*SD*) | *M* (*SD*) | *M* (*SD*) | *M* (*SD*) |
| Comprehension | Retrieval practice | 0.69 (0.09) | 6.52 (0.35) | 5.65 (0.27) | 2.03 (0.15) | 2.93 (0.17) | 2.91 (0.13) | 3.93 (0.23) | 0.25 (0.06) |
| | Restudy | 0.74 (0.13) | 6.51 (0.25) | 5.75 (0.19) | 2.04 (0.10) | 2.96 (0.13) | 2.94 (0.11) | 4.09 (0.28) | 0.28 (0.04) |
| | Untrained | 0.69 (0.11) | 6.62 (0.31) | 5.69 (0.26) | 2.05 (0.10) | 2.99 (0.10) | 3.01 (0.13) | 3.99 (0.27) | 0.26 (0.04) |
| Production | Retrieval practice | 0.77 (0.12) | 6.55 (0.28) | 5.67 (0.24) | 2.05 (0.17) | 2.98 (0.11) | 2.97 (0.16) | 3.84 (0.33) | 0.25 (0.02) |
| | Errorless learning | 0.74 (0.12) | 6.48 (0.14) | 5.70 (0.13) | 2.09 (0.06) | 2.91 (0.24) | 2.94 (0.11) | 3.85 (0.28) | 0.28 (0.03) |
| | Untrained | 0.73 (0.11) | 6.72 (0.37) | 5.81 (0.29) | 2.12 (0.10) | 2.95 (0.13) | 3.00 (0.12) | 3.89 (0.24) | 0.3 (0.06) |

[a] Log-word frequency per million. [b] Visual complexity of the target and foil image. Values range from 1 to 5, with 1 = *very simple* and 5 = *very complex*. [c] Visual similarity between the target image and the foil image. Values range from 0 to 10, with 0 = *no similarity at all* and 10 = *very similar*. [d] Correlation values from latent semantic analysis as an index of semantic relatedness of the target and foil pairs (LSA; Landauer et al., 1998).

80 experimental trials and 10 filler trials. The first training block of a session was preceded by 10 practice trials. To avoid privileging memory for any experimental items due to list primacy or recency effects, the first and last five trials of all training blocks were filler trials, which involved items that were not assigned to any experimental condition. If a participant completed both modules, the number of blocks matched across the modules. Breaks were taken between blocks or as needed. Each block lasted approximately 30 min. Each experimental item was presented for four training trials. The lag between a given item's trials was not fixed, but there were at least 11 intervening trials between an item's trials (avg. = 19; min = 11; max = 27). The decision to incorporate a large number of intervening items was based on the learning literature and findings in aphasia suggesting that longer lag is correlated with better retention (e.g., Middleton et al., 2016; Pyc & Rawson, 2009).[4] Furthermore, the varying number of intervening items allowed the item order to be less predictable (e.g., the item "diamonds" was not always followed by the item "hamster"). The average ordinal position within a block was equated for the items in the various training conditions. All trials ended in correct-answer feedback.

**Comprehension Training Trials.** See Figure 2 for trial structure. During each trial of comprehension training (regardless of training type), the target and foil image for a pair were shown, and the target name was presented in auditory and written form. On a retrieval practice trial in the comprehension module, the participant was instructed to choose the object that matched the target name by pointing to one of two cards on the desk that showed colored arrows pointing to the left or right side of the screen. The experimenter registered the participant's choice by clicking the chosen image on the screen, which was then outlined with a gray border to provide visual confirmation of the participant's choice. The trial advanced after 8 s. On a restudy trial, the participant saw the target and foil image for a pair, and the target name was presented after 1 s. The 1-s delay was included to promote processing of both images prior to target name presentation. When the target name was presented, the target image was identified with a yellow highlighted border. The participant was instructed to quietly study the target and its name, and the trial advanced after 8 s. Feedback followed each retrieval practice and restudy trial—the software identified the target image with a flashing yellow border, and an opaque prohibition symbol was overlayed on the foil image (see Figure 2).

For items assigned to the retrieval practice condition, the first trial was a restudy trial (following the standard practice of initial familiarization prior to retrieval practice in the memory literature) and the subsequent three trials involved retrieval practice. Items assigned to the restudy condition were presented for four restudy trials.

**Production Training Trials.** During each trial of production training (regardless of training type), only the target image was ever shown. On a retrieval practice trial in the production module, the target image was displayed and the participant had 8 s to attempt to name the picture (Figure 3). During errorless learning, the target image was presented along with the target name presented visually and auditorily, and the participant was instructed to repeat the name once; the image and written name were displayed for the full 8 s. Each trial ended in correct-answer feedback, in which the target image was presented with the target name, and the participant was instructed to repeat the name. For items assigned to the retrieval practice condition, the first trial involved errorless learning to serve as initial familiarization with the item, and a subsequent three trials involved retrieval practice for that item. For items assigned to the errorless learning condition, each item was presented for four errorless learning trials.

### Retention Test Sessions

Each training module concluded with a 1-day retention test and a 1-week retention test involving WPV for the items assigned to that module (trained and untrained items) following procedures described in the Item Selection and Item Assignment section (see Figure 1). Each target and foil image for a pair were probed once per test. Just like during item selection, items were divided into blocks such that the first block included targets for a randomly selected half of items and foils for the other half of items, presented in random order. The second block included the remaining targets and foils, presented in random order. For all participants across all modules, the 1-day retention test was administered the next day. In some cases (due to inclement weather, problems with transport services, illness, etc.), the 1-week retention test was administered before or after 7 days following the training session (range 6–9 days, $M$ = 7.14, $SD$ = .77).

## Response Coding

### WPV Accuracy

During item selection and at each retention test, WPV accuracy for a given item was coded as correct if the participant responded "yes" when the target picture (e.g., lobster) was shown with the target name (e.g., "lobster") and they also responded "no" when the foil picture (e.g., crab) was shown with the target name. All other response combinations were coded as incorrect.

### Choice Accuracy

During comprehension training, an accurate response on a retrieval practice trial corresponded to correctly choosing the target picture for the target word (hereafter, *choice accuracy*).

### Production Accuracy

During production training, naming attempts were coded with a binary variable of *production accuracy* (correct vs. error), which required two stages of coding. Following transcription into the International Phonetic Alphabet, the first complete response per trial was first coded for phonological overlap (Lecours & Lhermitte, 1969), a continuous measure of phonological similarity between the response and the target. A python script was used to automatically calculate the percentage of similar phonemes between the transcribed response and target, with manual recalculation by trained coders as needed. For example, descriptions of any kind, including nonnoun responses (e.g., "planting" for the item *garden*)

[4] Selection of lag in learning studies can be challenging. Lags that are too short may fail to capitalize on effortful processing. On the other hand, in the case of lags that are too long, an item may not be consistently referenced in memory across its trials, undermining learning. We chose the present lag range based on prior work on naming in aphasia (Middleton et al., 2016) showing similar retention test performance for items trained at lag-15 and lag-30.

as well as part-of-picture errors (e.g., "bandages" for *mummy*) received a phonological overlap score of zero to avoid credit for coincidental phonological similarity to the target. Second, phonological overlap scores were converted into a binary variable of production accuracy, in which ≥.75 = correct, <.75 = error. The .75 phonological overlap cutoff is designed to give credit for successful word retrieval while being lenient for minor phonological-phonetic encoding disturbances that commonly occur following word retrieval in PWA.

## Analyses

WPV accuracy per each item at each test timepoint was modeled with mixed logistic regression using the lme4 package in R (R Core Team, 2021). Items and participants were treated as random effects in each model except in one case in which it was necessary to drop items as a random effect for model convergence (generalization analysis—production module, going from item selection to the 1-day test, reported in Table 3). By-participant random slopes for the design factors were included if they improved the fit of the model by a chi-square test of deviance in model log likelihoods (α = .05) and their inclusion did not lead to model nonconvergence. All design factors were dummy coded. Within each module, at each test timepoint, planned comparisons were conducted to assess the effects of training type. Additional models examining training performance, durability of learning, and generalization to foils in the production module in a pre-to-post analysis, are described in more detail below. Last, an exploratory analysis of individual participant response to the retrieval practice learning factors is described in the Exploratory Individual Differences Analyses section.

### *Transparency and Openness*

All analyses were conducted in R Version 4.0.3 (R Core Team, 2021). There were no data exclusions, and all manipulations are reported. Determination of the sample size is described in the Overview of Current Research section. The study's design and analysis were not preregistered. The data and analysis code can be retrieved from https://osf.io/wbs6c/?view_only=42b0db10f2f946c4b3bac0f14fafb91e.

# Results

## Training Performance

Table 4 displays average production accuracy during production retrieval practice and errorless learning, as well as choice accuracy during comprehension retrieval practice. Training accuracy was generally high, although production accuracy during production retrieval practice was considerably lower than during errorless learning training ($p$ = .007; model reported in Table 5).

## Retention Test Performance

Figure 5 displays mean increase in WPV accuracy for each training condition, calculated relative to untrained items in a given module for simplicity of presentation (for all WPV accuracy condition means, see Appendix Table A2; for WPV retention test performance per condition per participant, see Appendix Table A3). In Figure 5, performance in the production module is shown in the left panel, and performance in the comprehension module is shown in the right panel.

Table 6 reports regression output for models of WPV accuracy at the 1-day and 1-week retention tests in the production module. The errorless learning condition was superior to the untrained items at the 1-day test ($p$ = .004; Table 6) and the 1-week test ($p$ = .025; Table 6). The retrieval practice condition showed marginal enhanced performance relative to untrained items at the 1-day test ($p$ = .063) but a robust increase in WPV accuracy relative to untrained items by the 1-week test ($p$ < .001). The retrieval practice and errorless learning conditions did not differ from one another at either test (both $ps$ > .12; Table 6). However, a model testing the interaction of condition (retrieval practice vs. errorless learning) and time of test was significant (coefficient = 0.62, $SE$ = 0.31,

**Table 3**
*Interaction Coefficients in Mixed Logistic Models Examining Change in Foil Accuracy From Item Selection to Retention Test*

| Module and condition | Interaction coefficient | *SE* | *Z* | *p* |
|---|---|---|---|---|
| Comprehension module: item selection to 1-day test | | | | |
| Restudy (reference level: untrained) | 1.66 | 0.25 | 6.60 | <.001 |
| Retrieval practice (reference level: untrained) | 1.88 | 0.25 | 7.40 | <.001 |
| Retrieval practice (reference level: restudy) | 0.22 | 0.27 | 0.81 | .42 |
| Comprehension module: item selection to 1-week test | | | | |
| Restudy (reference level: untrained) | 0.95 | 0.24 | 4.02 | <.001 |
| Retrieval practice (reference level: untrained) | 1.14 | 0.24 | 4.80 | <.001 |
| Retrieval practice (reference level: restudy) | 0.19 | 0.25 | 0.79 | .43 |
| Production module: item selection to 1-day test | | | | |
| Errorless learning (reference level: untrained) | 0.32 | 0.25 | 1.31 | .19 |
| Retrieval practice (reference level: untrained) | 0.31 | 0.25 | 1.23 | .22 |
| Retrieval practice (reference level: errorless learning) | −0.02 | 0.25 | −0.07 | .94 |
| Production module: item selection to 1-week test | | | | |
| Errorless learning (reference level: untrained) | 0.31 | 0.27 | 1.15 | .249 |
| Retrieval practice (reference level: untrained) | 0.76 | 0.27 | 2.81 | .005 |
| Retrieval practice (reference level: errorless learning) | 0.45 | 0.27 | 1.68 | .092 |

*Note*. Interaction coefficient = model estimation in log odds of the change in foil accuracy as a function of the interaction of phase (item selection to retention test) and condition (each training condition relative to the specified reference level); $SE$ = standard error of the estimate; $Z$ = Wald $Z$-test statistic.

**Table 4**

*Average Training Performance and Standard Errors Across Participants by Training Type and Module*

| Training type | Production accuracy (*SE*) | Choice accuracy (*SE*) |
|---|---|---|
| Production retrieval practice | .79 (.06) | — |
| Errorless learning | .95 (.03) | — |
| Comprehension retrieval practice | — | .98 (.01) |

*Note.* *SE* = standard error of the estimate.

$Z = 2.00$, $p = .045$), suggesting greater relative benefit from retrieval practice after a longer retention interval.

Table 7 reports regression output for models of WPV accuracy in the comprehension module. Restudy and retrieval practice conferred superior WPV accuracy benefits over untrained items at the 1-day and 1-week retention tests (all *p*s < .001; Table 7). Retrieval practice and restudy did not differ at either of the retention tests (both *p*s > .59).[5]

### *Exploratory Individual Differences Analyses*

An important component of an evidence base for clinical decision-making involves an understanding of how an individual's profile of deficits relates to differential response to different treatment approaches. In our participant sample, individuals varied in degree of semantic comprehension deficit and verbal comprehension deficit (Camel and Cactus Test and synonymy matching, respectively; Table 1). Impairment on both is consistent with a multimodal semantic deficit. To provide initial observations regarding individual differences in comprehension abilities and response to retrieval practice learning factors, we conducted two regression analyses, one per module (see Appendix Table A5, for model output; WPV retention test performance per participant is provided in Appendix Table A3). The dependent variable in each model was a difference score corresponding to the relative advantage for retrieval practice over the comparison treatment (restudy or errorless learning) at the retention tests within a module. Camel and Cactus score, synonymy matching, and time of test were entered as covariates. In the comprehension module, synonym matching score showed a strong negative relationship with the dependent variable (coefficient = −0.004, *SE* = 0.002, *t* = −2.37), potentially indicating that retrieval practice is particularly beneficial for those with more severe verbal comprehension impairment. In the production module, the model required simplification to a linear regression to achieve convergence. In that model, nonverbal semantic comprehension was strongly and positively related to the dependent variable, that is, higher Camel and Cactus scores related to greater relative benefit from retrieval practice (coefficient = 0.016, *SE* = 0.005, *t* = 3.15). We return to these findings in the discussion.

**Table 5**

*Mixed Logistic Model Coefficients and Associated Test Statistics: Analysis on Training Performance in the Production Module*

| Model terms | Coefficient | *SE* | *Z* | *p* |
|---|---|---|---|---|
| | Production accuracy at training | | | |
| Fixed effects | | | | |
| Intercept | 5.85 | 1.40 | | |
| Training type effect | | | | |
| Retrieval practice[a] | −3.68 | 1.36 | −2.70 | .007 |
| Random effects | $s^2$ | | | |
| Participants: training type | 6.66 | | | |
| Participants | 7.34 | | | |
| Items | 2.55 | | | |

*Note.* Excluding the intercepts, coefficient = model estimation of the change in production accuracy (in log odds) from the reference level for the fixed effect; *SE* = standard error of the estimate; *Z* = Wald *Z*-test statistic, two-tailed; $s^2$ = random effect variance.
[a] Reference level is errorless learning.

## Durability of Learning

In each module, durability of learning as a function of condition was examined by identifying correct WPV responses at the 1-day test; and, among those, modeling the proportion of items that were still correct at the 1-week test (e.g., Mettler et al., 2016). Figure 6 plots durability per condition in the production module (left panel) and comprehension module (right panel). In the comprehension module, durability was similar across the restudy, retrieval practice, and untrained items (all *p*s > .40; model results reported in Table 8). However, in the production module, retrieval practice was associated with superior durability compared to both the untrained items ($p = .011$) and errorless learning ($p = .013$; Table 8). The untrained items and errorless learning did not differ in terms of durability in the production module ($p = .809$).

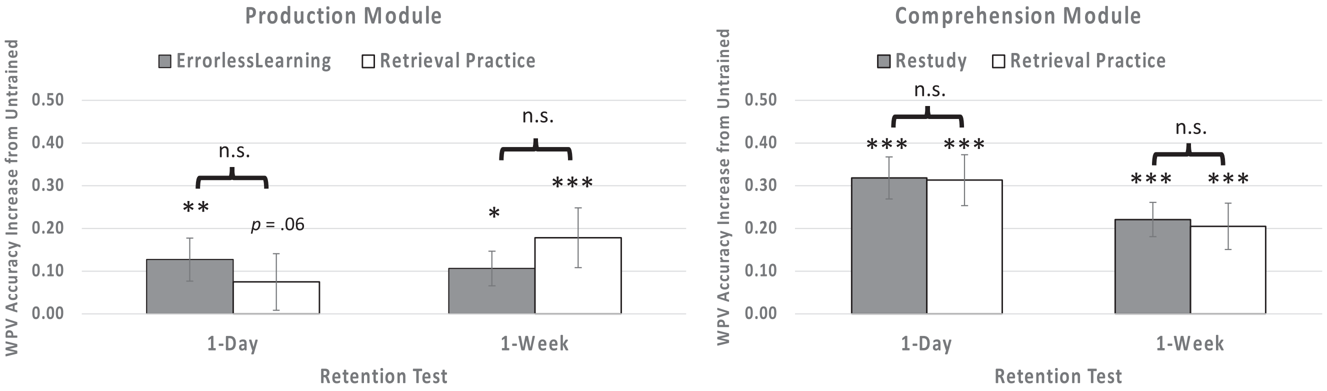
## Generalization Analysis: Improvement on Foils

In this section, we provide a more complete understanding of the basis for the improvements in WPV accuracy. Examination of errorful behavior during item selection revealed our participants were considerably more likely to err on foils than target trials. Following item selection, of the errorful items assigned into the various conditions, item selection performance on target trials was 92% accurate (*SD* = 6%) across participants but only 33% accurate (*SD* = 14%) on foil trials. This difference can be understood to reflect our population's ability to successfully reference a concept's general semantic space but not reliably make refined semantic distinctions within a semantic domain. As the targets were very near ceiling prior to treatment, in a final set of analyses we modeled improvement on the foils from item selection to each of the retention tests as a function of the conditions. These analyses were motivated by questions concerning generalization. The typical form of confrontation naming and errorless learning in clinical practice involves practice naming individual items. In contrast, receptive forms of treatment typically involve some form of

[5] Our designated main dependent variable of WPV accuracy is of primary focus because it was designed specifically to measure success at making refined semantic distinctions. However, retention test performance in the present study is also amenable to signal-detection analysis. For interested readers, we have examined potential differences in d prime (sensitivity) and β (response bias) between conditions within each module at each retention test. Model output is provided in Appendix Table A4, which shows no significant differences in bias between conditions. Significant differences in d prime largely track patterns of results obtained in the WPV accuracy analyses.

**Figure 5**

*Mean Increase in WPV Accuracy for Each Training Condition Relative to Their Untrained Control in a Given Module (Production Module, Left Panel; Comprehension Module, Right Panel)*



*Note.* Error bars reflect standard error of the mean difference between conditions across participants. Significance levels estimated with mixed-effects regression reported in Tables 6 and 7. WPV = word–picture verification.
* $p < .05$. ** $p < .01$. *** $p < .001$.

contrastive encoding such as choosing between a target and foil(s), as in the present study. A consequence of this is that foils are not presented as a comparator during production training, as they are during comprehension training. Thus, improvements on foils in the present study from production-based practice would constitute generalization in the form of refinement of semantics from treatment.

**Table 6**

*Mixed Logistic Model Coefficients and Associated Test Statistics: Analyses of Retention Test Performance in the Production Training Module*

| Model terms | Coefficient | *SE* | *Z* | *p* |
|---|---|---|---|---|
| WPV accuracy at the 1-day test | | | | |
| Fixed effects | | | | |
| Intercept | 0.05 | 0.24 | | |
| Training type effect | | | | |
| Errorless learning[a] | 0.62 | 0.22 | 2.88 | .004 |
| Retrieval practice[a] | 0.39 | 0.21 | 1.86 | .063 |
| Retrieval practice[b] | −0.23 | 0.21 | −1.10 | .272 |
| Random effects | $s^2$ | | | |
| Participants | 0.27 | | | |
| Items | 0.55 | | | |
| WPV accuracy at the 1-week test | | | | |
| Fixed effects | | | | |
| Intercept | −0.58 | 0.29 | | |
| Training type effect | | | | |
| Errorless learning[a] | 0.48 | 0.22 | 2.25 | .025 |
| Retrieval practice[a] | 0.80 | 0.22 | 3.68 | <.001 |
| Retrieval practice[b] | 0.32 | 0.21 | 1.53 | .127 |
| Random effects | $s^2$ | | | |
| Participants | 0.49 | | | |
| Items | 0.49 | | | |

*Note.* Excluding the intercepts, coefficient = model estimation of the change in WPV accuracy (in log odds) from the reference level for each fixed effect; *SE* = standard error of the estimate; *Z* = Wald *Z*-test statistic; $s^2$ = random effect variance; WPV = word–picture verification.
[a] Reference level is untrained condition in the production module. [b] Reference level is errorless learning condition.

Table 3 reports model results corresponding to estimated interaction coefficients for phase (item selection vs. retention test) and condition (within a module, each condition compared to a specified reference condition; note, for mean foil accuracy per condition, module, and phase, see Appendix Table A6). First, going from item selection to the 1-day test and the 1-week test, the comprehension module showed a robust enhancement in performance on foils from restudy and the receptive form of retrieval practice compared to untrained (all *p*s < .001). These results may not be particularly surprising given the role of foils as comparators during comprehension training. Of greater interest are improvements on performance on foils in the production module. Figure 7 displays improvement in foils from item selection (pretraining) to each of the retention tests (for simplicity, relative to untrained items) for errorless learning and retrieval practice in the production module. Going from item selection to the 1-day test, the benefit to foils was not different for errorless learning or production retrieval practice compared to untrained items (all *p*s > .18). However, going from item selection to the 1-week test, the production retrieval practice condition showed an advantage over untrained items in terms of improvement on the foils ($p = .005$), whereas the benefit in the errorless learning condition was not reliable ($p = .249$). The more robust effects of retrieval practice at longer retention intervals align with the durability findings reported in the Durability of Learning section. It is of considerable theoretical interest and clinical relevance that in the production module, retrieval practice (but not errorless learning) led to robust improvements in foils (relative to the untrained items), despite the foils never having been presented during training. This constitutes a definitive demonstration of generalization and points to a refinement of the semantic space from repeated semantically driven retrieval in the course of production.

## Discussion

Twelve PWA with lexical–semantic deficits completed the present study, which examined how retrieval practice improved ability

**Table 7**

*Mixed Logistic Model Coefficients and Associated Test Statistics: Analyses of Retention Test Performance in the Comprehension Training Module*

| Model terms | Coefficient | *SE* | *Z* | *p* |
|---|---|---|---|---|
| | WPV accuracy at the 1-day test | | | |
| Fixed effects | | | | |
| Intercept | 0.16 | 0.21 | | |
| Training type effect | | | | |
| Restudy[a] | 1.72 | 0.23 | 7.60 | <.001 |
| Retrieval practice[a] | 1.80 | 0.22 | 8.01 | <.001 |
| Retrieval practice[b] | 0.07 | 0.23 | 0.32 | .749 |
| Random effects | $s^2$ | | | |
| Participants | 0.30 | | | |
| Items | 0.50 | | | |
| | WPV accuracy at the 1-week test | | | |
| Fixed effects | | | | |
| Intercept | 0.38 | 0.31 | | |
| Training type effect | | | | |
| Restudy[a] | 1.23 | 0.21 | 5.89 | <.001 |
| Retrieval practice[a] | 1.12 | 0.20 | 5.56 | <.001 |
| Retrieval practice[b] | −0.11 | 0.20 | −0.53 | .596 |
| Random effects | $s^2$ | | | |
| Participants | 0.91 | | | |
| Items | 0.50 | | | |

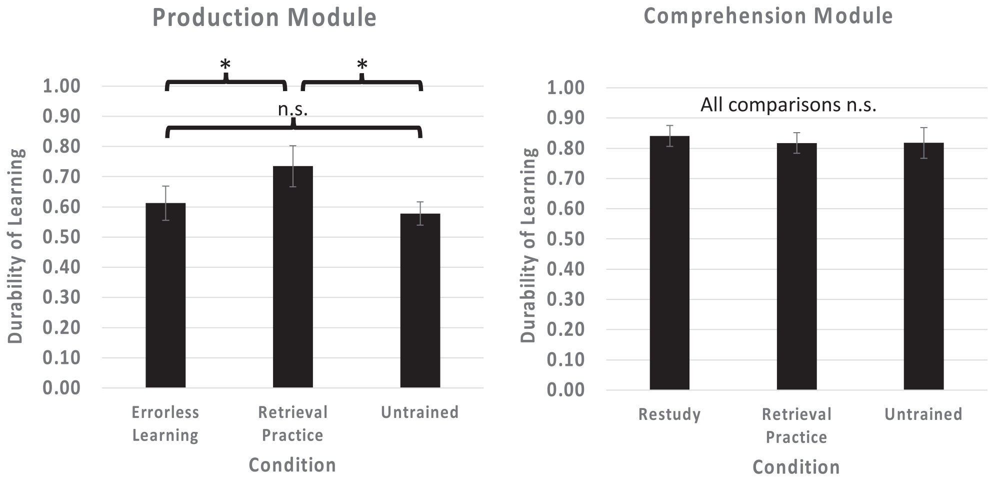
*Note.* Excluding the intercepts, coefficient = model estimation of the change in WPV accuracy (in log odds) from the reference level for each fixed effect; *SE* = standard error of the estimate; *Z* = Wald *Z*-test statistic; $s^2$ = random effect variance; WPV = word–picture verification.
[a] Reference level is untrained condition. [b] Reference level is restudy condition.

to discriminate closely related but distinct concepts. At the post-training comprehension tests, the main dependent variable was WPV accuracy. An accurate WPV response required a participant to both correctly accept the word (“backpack”) for its target picture (*backpack*) and reject the word for a semantically related foil picture (*lunchbox*) on nonconsecutive trials (Hillis et al., 1990; Rapp & Caramazza, 2002). All participants completed the comprehension training module, which compared receptive retrieval practice (i.e., participant chooses between the target and foil object as the match to the word) to restudy (i.e., the software identifies the target object for the word from among the target and foil objects). Eight participants completed both the comprehension module and a production module. The production module involved retrieval practice (i.e., target object is presented for naming practice) versus errorless learning (i.e., target object and word are presented simultaneously; participant repeats the word). Feedback followed all training trials. In each module, trained and untrained target–foil pairs were probed with WPV at tests administered 1 day and 1 week following that module’s training session.

In the comprehension module, both training conditions conferred robust benefits to WPV accuracy compared to untrained items at both retention tests. However, receptive retrieval practice and restudy did not differ at the retention tests in terms of WPV accuracy, and they conferred similar durability of learning (see Durability of Learning section). As described in the Retrieval Practice and Aphasia Treatment section, we did not have strong expectations for the relative benefits of receptive retrieval practice versus restudy. As reported in a meta-analysis, retrieval practice that involves receptive tests (e.g., multiple choice; recognition judgments) confers a benefit relative to restudy but it is weaker and less consistent than when retrieval practice requires participants to generate target information such as during cued or free recall (Rowland, 2014). However, other work has found that receptive retrieval practice can confer potent learning when the foils are competitive with the target, which arguably was the case in the present study (Little & Bjork, 2015). Overall, this experimental contrast applied to this domain of cognitive rehabilitation is just a first step, and worthy of follow-up. First, more generally there is a paucity of research examining the efficacy of receptive forms of

**Figure 6**

*Durability of Word–Picture Verification Accuracy per Condition in the Production Module (Left Panel) and Comprehension Module (Right Panel)*



*Note.* Error bars reflect standard error of the mean per condition across participants. Significance levels estimated with mixed-effects regression reported in Table 8.
* *p* < .05.

**Table 8**

*Mixed Logistic Model Coefficients and Associated Test Statistics: Durability of Learning Analyses*

| Model terms | Coefficient | *SE* | *Z* | *p* |
|---|---|---|---|---|
| Production module | | | | |
| Fixed effects | | | | |
| Intercept | 0.40 | 0.26 | | |
| Training type effect | | | | |
| Errorless learning[a] | 0.06 | 0.27 | 0.24 | .809 |
| Retrieval practice[a] | 0.76 | 0.30 | 2.55 | .011 |
| Retrieval practice[b] | 0.69 | 0.28 | 2.48 | .013 |
| Random effects | $s^2$ | | | |
| Participants | 0.24 | | | |
| Items | 0.13 | | | |
| Comprehension module | | | | |
| Fixed effects | | | | |
| Intercept | 1.83 | 0.34 | | |
| Training type effect | | | | |
| Restudy[c] | 0.13 | 0.29 | 0.45 | .652 |
| Retrieval practice[c] | −0.07 | 0.28 | −0.24 | .812 |
| Retrieval practice[d] | −0.20 | 0.24 | −0.81 | .42 |
| Random effects | $s^2$ | | | |
| Participants | 0.59 | | | |
| Items | 0.34 | | | |

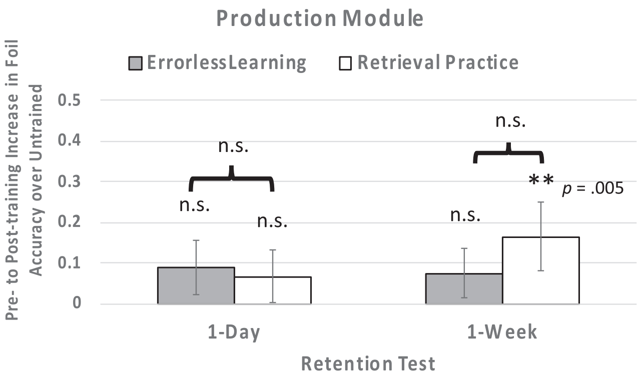
*Note.* Excluding the intercepts, coefficient = model estimation of the change in rate of WPV performance retention from 1-day to 1-week test (in log odds) from the reference level for each fixed effect; *SE* = standard error of the estimate; *Z* = Wald *Z*-test statistic; $s^2$ = random effect variance; WPV = word–picture verification.

[a] Reference level is untrained condition in the production module. [b] Reference level is errorless learning condition. [c] Reference level is untrained condition in the comprehension module. [d] Reference level is restudy condition.

treatment for addressing semantic-based word-comprehension deficits in aphasia, and further development of the evidence base is important. Second, this study is the first to examine the impact of a potentially potent learning factor on semantic-based word-comprehension deficits in aphasia. It provides a launch point for future work seeking to characterize the full clinical applicability of receptive retrieval practice for semantic-based word-comprehension deficits in aphasia. Such an endeavor is likely to be worthwhile, considering retrieval practice research developed alongside a vast literature addressing optimal dosing and scheduling of learning experiences for maximizing the benefits from retrieval practice and other types of learning. Systematic translation from these literature to aphasia has shown promise for optimizing naming treatment efficacy for lexical access deficits, a primary contributor to naming disorders in aphasia (for reviews, see de Lima et al., 2020; Middleton et al., 2020).

**Figure 7**

*Generalization Measured by Mean Improvement in Foil Accuracy in the Production Module*



*Note.* Improvement in foils from pretraining (i.e., item selection; see Item Selection and Item Assignment section) to each of the retention tests relative to untrained items for errorless learning and retrieval practice in the production module. Error bars correspond to standard error of the mean interaction estimate across participants. Significance levels estimated with mixed-effects regression reported in Table 3.

In the production module, both methods of training (retrieval practice and errorless learning) were associated with *transfer* at one or both test timepoints; that is, items that underwent production practice were associated with higher WPV accuracy at the retention tests relative to the untrained items. This benefit was observed at both tests for errorless learning and at the 1-week test for retrieval practice, with a marginal benefit of retrieval practice over untrained items at the 1-day test for retrieval practice. Additionally, the retrieval practice and errorless learning conditions did not differ in terms of WPV accuracy at either test; however, an interaction indicated greater relative benefit from retrieval practice with a longer retention interval. Also, the durability of learning analysis (see Durability of Learning section) revealed greater retention of correct responding in the retrieval practice condition relative to the errorless learning condition and to untrained items, with no difference between untrained and errorless learning items. This finding aligns with the learning and memory literature, in which the benefits from retrieval practice over nonretrieval forms of learning are reflective of the more durable learning conferred from retrieval (for discussion, Kornell et al., 2011). Another important result was the generalization pattern observed in the production module, in that improvement on foils from pre- to post-training was greater for retrieval practice compared to errorless learning (see Generalization Analysis: Improvement on Foils section). Because the foils were never shown during training in the production module, we take this as evidence of greater semantic refinement following retrieval practice naming treatment.

The durability and generalization advantages for retrieval practice over errorless learning may be understood by assuming that retrieval practice (i.e., naming from a depicted object) more strongly engages the first stage of lexical access (mapping from semantics to words) than errorless learning (i.e., repeating the word) does. This difference in engagement or ‘use’ confers greater strengthening to that mapping compared to errorless learning, in which along with some semantic activation from the picture, the word is activated by auditory input, reducing reliance on the semantics-word mapping for production (Schuchard & Middleton, 2018a, 2018b). More durable changes to comprehension could arise because greater strengthening of the semantic-to-lexical mapping in word production also benefits the mapping from the target to semantics (i.e., word comprehension). Greater generalization to foils could result if using the mapping from semantics to words (required for retrieval practice) sharpens the mapping both ways or retrieval sharpens conceptual distinctions within semantics itself.

When weighing the relative merits of the training conditions in each module, it is worth considering that the different training conditions provide key information in different dosages. That is, necessarily there is lack of control in exposure to target information in each retrieval practice condition with respect to its control (restudy, or errorless learning). During a retrieval practice trial, processing of the target is limited to what the participant themself can generate (production module) or successfully identify (comprehension module), versus errorless learning/restudy, where target information is provided on every trial. Despite these disadvantages, WPV posttest performance revealed largely similar benefits for each retrieval practice condition with respect to its control. In the production module, even though production accuracy during training was lower for retrieval practice versus errorless learning, retrieval practice was associated with the superior durability of learning and generalization to foils relative to errorless learning. Future work could revisit the clinical applicability of retrieval practice for treating word-comprehension deficits by equating the rate of correct responding per item during retrieval practice versus non-retrieval-based learning. This involves presenting each item for retrieval practice in a distributed fashion until it elicits a set number of correct responses matched to the number of trials for the control learning method. This form of learning, termed "criterion learning," can be an efficient, potent schedule for administering retrieval practice because, compared to administering each item for a fixed number of trials despite the rate of success, criterion-learning schedules more retrieval practice for the items that need it.

Because of the clinical orientation of the work, our main analysis strategy did not include a direct comparison between the two training modules. However, if a clinician were pressed to choose a comprehension-based or production-based approach to addressing a lexical–semantic deficit, there is evidence the receptive forms of training were sometimes more potent in the present study. Among the eight participants who completed both modules, an interaction revealed that relative to untrained items, the benefit for retrieval practice was greater in the comprehension versus production module (coefficient = −1.54, $SE = 0.31$, $Z = -4.92$, $p < .001$) at the 1-day test; this pattern did not hold at the 1-week test ($p = .37$). Interactions also revealed the advantage for restudy versus untrained (comprehension module) was greater than the advantage for errorless learning over untrained (production module) at both the 1-day test (coefficient = −1.09, $SE = 0.31$, $Z = -3.50$, $p < .001$) and 1-week test (coefficient = −0.70, $SE = 0.30$, $Z = -2.32$, $p = .02$). Note however, there are many key differences between the comprehension and production training approaches, reflective of the clinical literature. Semantic-based treatments generally provide a semantic or phonological contrast set on each training trial, with such presentations permitting the direct comparison for encoding of distinguishing features. Second, in production training, only the target is ever experienced. Third, the format of the retention test more closely resembles the training procedure in the comprehension versus production module; greater similarity in processing at training and test can promote performance (e.g., Blaxton, 1989). Any of these differences may have contributed to a greater training benefit in the comprehension module. Nevertheless, we consider either the expressive or receptive training approaches in the present work appropriate depending on a patient's ability to engage in either type of treatment and the identified goals of treatment.

## Constraints on Generality

This work is novel in its efforts to apply retrieval practice principles to semantics-based word-comprehension disorders, which are clinically significant and prevalent in many neurological populations. However, in order to study lexical–semantic disorder in aphasia, our participants were selected from a larger sample because they demonstrated a particular neuropsychological profile. Given these selection constraints, the resulting sample size was small. We anticipated this outcome and built power into the design by maximizing the number of observations per condition per participant. However, it will be important to conduct similar examinations of learning principles with additional samples resembling the present one to examine the generality of our findings. In such studies, it would be of value to engage all participants in the same types of training with a similar number of items across participants to enhance statistical confidence and generalizability. It will also be important to examine other patient populations for whom semantic deficits can be more profound, such as Wernicke's or global aphasia, or those suffering from semantic variants of dementia. Likewise, it will be important in future work to examine deficits in comprehending the other major word classes, such as verbs, whose training might be expected to affect more complex comprehension processes such as sentence or discourse comprehension. Last, we note that individual differences will be important to consider in future work. Exploratory analyses in the present study suggested that in the production module, retrieval practice is more beneficial relative to errorless learning for individuals with better nonverbal semantic comprehension abilities. Conversely, in the comprehension module, poorer verbal comprehension (synonymy matching) ability related to a greater advantage for retrieval practice over restudy. Any interpretation of these results at this point would be purely speculative. However, the study design and effect size estimates provide a starting point for future studies to advance an understanding of individual response to retrieval practice learning factors.

## Conclusion

This study contributes to a growing body of work (Friedman et al., 2017; Middleton et al., 2015, 2016, 2019, 2020; Rapp & Wiley, 2019; Schuchard et al., 2020) seeking to translate from a vast literature on fundamental principles of human learning to help guide selection and scheduling of commonly used clinical tools for maximizing the efficiency and efficacy of interventions for aphasia. We have provided an experimental framework and original observations, to provide a foundation for future work seeking to systematically translate from fundamental principles of human learning to improve cognitive rehabilitation.

## References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Basso, A., Casati, G., & Vignolo, L. A. (1977). Phonemic identification defect in aphasia. *Cortex*, *13*(1), 85–95. https://doi.org/10.1016/S0010-9452(77)80057-9

Blaxton, T. (1989). Investigating dissociations among memory measures: Support for a transfer-appropriate processing framework. *Journal of*

*Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 657–668. https://doi.org/10.1037/0278-7393.15.4.657

Blumstein, S. E., Baker, E., & Goodglass, H. (1977). Phonological factors in auditory comprehension in aphasia. *Neuropsychologia*, *15*(1), 19–30. https://doi.org/10.1016/0028-3932(77)90111-7

Blumstein, S. E., Cooper, W. E., Zurif, E. G., & Caramazza, A. (1977). The perception and production of voice-onset time in aphasia. *Neuropsychologia*, *15*(3), 371–383. https://doi.org/10.1016/0028-3932(77)90089-6

Bozeat, S., Lambon Ralph, M. A., Patterson, K., Garrard, P., & Hodges, J. R. (2000). Non-verbal semantic impairment in semantic dementia. *Neuropsychologia*, *38*(9), 1207–1215. https://doi.org/10.1016/S0028-3932(00)00034-8

Breese, E. L., & Hillis, A. E. (2004). Auditory comprehension: Is multiple choice really good enough? *Brain and Language*, *89*(1), 3–8. https://doi.org/10.1016/S0093-934X(03)00412-7

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

Brodeur, M. B., Guérard, K., & Bouras, M. (2014). Bank of Standardized Stimuli (BOSS) phase II: 930 new normative photos. *PLOS ONE*, *9*(9), Article e106953. https://doi.org/10.1371/journal.pone.0106953

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

Casarin, F. S., Branco, L., Pereira, N., Kochhann, R., Gindri, G., & Fonseca, R. P. (2014). Rehabilitation of lexical and semantic communicative impairments: An overview of available approaches. *Dementia & Neuropsychologia*, *8*(3), 266–277. https://doi.org/10.1590/S1980-57642014DN83000011

Chen, L., Lambon Ralph, M. A., & Rogers, T. T. (2017). A unified model of human semantic knowledge and its disorders. *Nature Human Behaviour*, *1*(3), 1–10. https://doi.org/10.1038/s41562-016-0039

Clare, L., & Jones, R. S. (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

De Groot, F., Koelewijn, T., Huettig, F., & Olivers, C. N. (2016). A stimulus set of words and pictures matched for visual and semantic similarity. *Journal of Cognitive Psychology*, *28*(1), 1–15. https://doi.org/10.1080/20445911.2015.1101119

de Lima, M. F. R., Cavendish, B. A., de Deus, J. S., & Buratto, L. G. (2020). Retrieval practice in memory-and language-impaired populations: A systematic review. *Archives of Clinical Neuropsychology*, *35*(7), 1078–1093. https://doi.org/10.1093/arclin/acaa035

Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, *369*(1634), Article 20120394. https://doi.org/10.1098/rstb.2012.0394

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. https://doi.org/10.1037/0033-295X.104.4.801

Dial, H., & Martin, R. (2017). Evaluating the relationship between sublexical and lexical processing in speech perception: Evidence from aphasia. *Neuropsychologia*, *96*, 192–212. https://doi.org/10.1016/j.neuropsychologia.2017.01.009

Dignam, J. K., Rodriguez, A. D., & Copland, D. A. (2016). Evidence for intensive aphasia therapy: Consideration of theories from neuroscience and cognitive psychology. *PM & R*, *8*(3), 254–267. https://doi.org/10.1016/j.pmrj.2015.06.010

Duñabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2018). MultiPic: A standardized set of 750 drawings with norms for six European languages. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *71*(4), 808–816. https://doi.org/10.1080/17470218.2017.1310261

Dunn, L. M., & Dunn, D. M. (2007). *Examiner's manual for the PPVT-IV: Peabody Picture Vocabulary Test* (4th ed.). Pearson Publications.

Fillingham, J. K., Hodgson, C., Sage, K., & Lambon Ralph, M. A. (2003). The application of errorless learning to aphasic disorders: A review of theory and practice. *Neuropsychological Rehabilitation*, *13*(3), 337–363. https://doi.org/10.1080/09602010343000020

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

Gambi, C., & Pickering, M. J. (2017). Models linking production and comprehension. In E. M. Fernández & H. S. Cairns (Eds.), *The handbook of psycholinguistics* (pp. 157–181). Wiley-Blackwell. https://doi.org/10.1002/9781118829516.ch7

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

Howard, D., Best, W., & Nickels, L. (2015). Optimising the design of intervention studies: Critiques and ways forward. *Aphasiology*, *29*(5), 526–562.

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

Kay, J., Lesser, R., & Coltheart, M. (1992). *PALPA: Psycholinguistic assessments of language processing in aphasia*. Lawrence Erlbaum.

Kertesz, A. (2007). *Western Aphasia Battery—Revised (WAB-R)*. Pearson.

Knollman-Porter, K., Dietz, A., & Dahlem, K. (2018). Intensive auditory comprehension treatment for severe aphasia: A feasibility study. *American Journal of Speech-Language Pathology*, *27*(3), 936–949. https://doi.org/10.1044/2018_AJSLP-17-0117

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*(2), 85–97. https://doi.org/10.1016/j.jml.2011.04.002

Kornell, N., & Vaughn, K. E. (2016). How retrieval attempts affect learning: A review and synthesis. *Psychology of Learning and Motivation*, *65*, 183–215. https://doi.org/10.1016/bs.plm.2016.03.003

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

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

Little, J. L., & Bjork, E. L. (2015). Optimizing multiple-choice tests as tools for learning. *Memory & Cognition*, *43*(1), 14–26. https://doi.org/10.3758/s13421-014-0452-8

Martin, N., Dell, G. S., Saffran, E. M., & Schwartz, M. F. (1994). Origins of paraphasias in deep dysphasia: Testing the consequences of a decay impairment to an interactive spreading activation model of lexical retrieval. *Brain and Language*, *47*(4), 609–660. https://doi.org/10.1006/brln.1994.1061

Mettler, E., Massey, C. M., & Kellman, P. J. (2016). A comparison of adaptive and fixed schedules of practice. *Journal of Experimental Psychology: General*, *145*(7), 897–917. https://doi.org/10.1037/xge0000170

Miceli, G., Gainotti, G., Caltagirone, C., & Masullo, C. (1980). Some aspects of phonological impairment in aphasia. *Brain and Language*, *11*(1), 159–169. https://doi.org/10.1016/0093-934X(80)90117-0

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

Middleton, E. L., Schuchard, J., & Rawson, K. A. (2020). A review of the application of distributed practice principles to naming treatment in aphasia. *Topics in Language Disorders*, *40*(1), 36–53. https://doi.org/10.1097/TLD.0000000000000202

Middleton, E. L., & Schwartz, M. F. (2012). Errorless learning in cognitive rehabilitation: A critical review. *Neuropsychological Rehabilitation*, *22*(2), 138–168. https://doi.org/10.1080/09602011.2011.639619

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

Middleton, E. L., Schwartz, M. F., Rawson, K. A., Traut, H., & Verkuilen, J. (2016). Towards a theory of learning for naming rehabilitation: Retrieval practice and spacing effects. *Journal of Speech, Language, and Hearing Research*, *59*(5), 1111–1122. https://doi.org/10.1044/2016_JSLHR-L-15-0303

Mirman, D., & Britt, A. E. (2014). What we talk about when we talk about access deficits. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, *369*(1634), Article 20120388. https://doi.org/10.1098/rstb.2012.0388

Morris, J., & Franklin, S. (2012). Investigating the effect of a semantic therapy on comprehension in aphasia. *Aphasiology*, *26*(12), 1461–1480. https://doi.org/10.1080/02687038.2012.702885

Morris, J., & Franklin, S. (2017). Disorders of auditory comprehension. In I. Papathanasiou & P. Coppens (Eds.), *Aphasia and related neurogenic communication disorders* (2nd ed., pp. 151–168). Jones & Bartlett Learning.

Nickels, L. (2000). Semantics and therapy in aphasia. In W. Best, K. Bryan, & J. Maxim (Eds.), *Semantic processing: Theory and practice* (pp. 108–124). Whurr.

Oren, S., Willerton, C., & Small, J. (2014). Effects of spaced retrieval training on semantic memory in Alzheimer's disease: A systematic review. *Journal of Speech, Language, and Hearing Research*, *57*(1), 247–270. https://doi.org/10.1044/1092-4388(2013/12-0352)

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

Pylkkänen, L. (2019). The neural basis of combinatory syntax and semantics. *Science*, *366*(6461), 62–66. https://doi.org/10.1126/science.aax0050

Rapp, B., & Caramazza, A. (2002). Selective difficulties with spoken nouns and written verbs: A single case study. *Journal of Neurolinguistics*, *15*(3–5), 373–402. https://doi.org/10.1016/S0911-6044(01)00040-9

Rapp, B., & Wiley, R. W. (2019). Re-learning and remembering in the lesioned brain. *Neuropsychologia*, *132*, Article 107126. https://doi.org/10.1016/j.neuropsychologia.2019.107126

Rawson, K. A., & 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. https://doi.org/10.1037/a0023956

Rawson, K. A., & Dunlosky, J. (2013). Relearning attenuates the benefits and costs of spacing. *Journal of Experimental Psychology: General*, *142*(4), 1113–1129. https://doi.org/10.1037/a0030498

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

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

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

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

Roediger, H. L., III, Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. *Psychology of Learning and Motivation*, *55*, 1–36. https://doi.org/10.1016/B978-0-12-387691-1.00001-6

Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception*, *33*(2), 217–236. https://doi.org/10.1068/p5117

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

Saffran, E. M., Schwartz, M. F., Linebarger, M., Martin, N., & Bochetto, P. (1988). *The Philadelphia comprehension battery* [Unpublished test].

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

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

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

Semenza, C. (2020). Lexical-semantic disorders in aphasia. In G. Denes & L. Pizzamiglio (Eds.), *Handbook of clinical and experimental neuropsychology* (pp. 215–244). Psychology Press. https://doi.org/10.4324/9781315791272-13

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 174–215. https://doi.org/10.1037/0278-7393.6.2.174

Szekely, A., Jacobsen, T., D'Amico, S., Devescovi, A., Andonova, E., Herron, D., Lu, C. C., Pechmann, T., Pléh, C., Wicha, N., Federmeier, K., Gerdjikova, I., Gutierrez, G., Hung, D., Hsu, J., Iyer, G., Kohnert, K., Mehotcheva, T., Orozco-Figueroa, A., . . . Bates, E. (2004). A new online resource for psycholinguistic studies. *Journal of Memory and Language*, *51*(2), 247–250. https://doi.org/10.1016/j.jml.2004.03.002

Webster, J., Whitworth, A., & Morris, J. (2015). Is it time to stop "fishing"? A review of generalisation following aphasia intervention. *Aphasiology*, *29*(11), 1240–1264. https://doi.org/10.1080/02687038.2015.1027169

(*Appendix follows*)

# Appendix

## Stimulus List and Supplemental Data Tables

**Table A1**
*List of 408 Target–Foil Semantic Minimal Pairs*

| Target | Foil | Target | Foil | Target | Foil |
|---|---|---|---|---|---|
| accordion | keyboard | exclamation point | semicolon | pizza | quesadilla |
| Africa | Australia | eyelashes | eyebrows | pliers | tweezers |
| airplane | rocket | farm | zoo | plug | USB |
| aisle | hallway | farmer | fisherman | plumber | mechanic |
| alligator | lizard | faucet | shower | pocket | collar |
| ambulance | fire truck | fingerprint | footprint | popcorn | potato chips |
| anchor | anvil | fireman | policeman | potato | spaghetti squash |
| ant | cricket | fireworks | explosion | puddle | stream |
| apple | mango | fist | peace sign | purse | briefcase |
| astronaut | scuba diver | flute | clarinet | push-up | sit-up |
| avalanche | flood | flyswatter | net | Q-tips | cotton balls |
| avocado | pear | football | bowling ball | queen | king |
| backpack | lunch box | fox | cat | rabbit | kangaroo |
| bagel | danish | freckles | wrinkles | radio | television |
| baker | butcher | frisbee | boomerang | rainbow | Aurora Borealis |
| balcony | cellar | frog | salamander | raincoat | overcoat |
| ballerina | gymnast | funnel | Erlenmeyer flask (V2) | rake | pitchfork |
| Band-Aid | bandage | futon | recliner | rat | ferret |
| bar | restaurant | garage | hangar | ravioli | spring rolls |
| barber | optometrist | garden | forest | receipt | coupon |
| barn | cabin | gate | fence | refrigerator | vending machine |
| barrel | chest | genie | ghost | rice | oatmeal |
| basket | tray | globe | map | ring | earrings |
| basketball | volleyball | goggles | sunglasses | rock | gem |
| battery | charger | gold | silver | rooster | hen |
| batting cage | driving range | goldfish | clown fish | safe | locker |
| beak | snout | golf cart | four-wheeler | sailboat | ship |
| beard | mustache | gorilla | baboon | salute | wave |
| beaver | weasel | grapes | blueberries | sandal | high heel |
| bed | couch | grass | moss | satellite dish | antenna |
| bedroom | living room | grenade | bomb | saw | box cutter |
| beehive | nest | guitar | violin | saxophone | trombone |
| bench | rocking chair | hairnet | shower cap | scar | scab |
| bib | diaper | ham | turkey | scarecrow | snowman |
| bicep | forearm | hamster | chipmunk | school | church |
| bicycle | motorcycle | hand | foot | scientist | doctor |
| billboard | banner | handstand | backbend | scissors | tongs |
| binder | folder | hay | leaves | shaving cream | mouthwash |
| birdcage | carrier | helicopter | drone | shawl | sweatshirt |
| blimp | submarine | high five | punch | shield | armor |
| blow-dryer | hair straightener | highway | street | shot | IV |
| bongos | bass drum | hose | watering can | shovel | hoe |
| bonnet | beanie | hot-air balloon | parachute | shrimp | scallops |
| bookshelf | filing cabinet | hug | kiss | silo | tower |
| bottle | vase | hut | igloo | silverware | china |
| bowling alley | shuffleboard | ice skates | Rollerblades | singer | drummer |
| boxing | wrestling | iceberg | island | sink | bathtub |
| boy | girl | icicles | crystals | skis | snowshoes |
| braces | fillings | ink | watercolors | skunk | porcupine |
| braid | ponytail | iPod | Walkman | slinky | jack-in-the-box |
| brain | skull | iron | sewing machine | smile | frown |
| branch | stump | ironing board | table | smoke detector | fire alarm |
| bread | crackers | Italy | China | snake | eel |
| bride | nun | jeans | shorts | sneakers | flip-flops |
| broccoli | parsley | jeep | pickup | snow | rain |
| broom | duster | jet ski | snowmobile | sock | stocking |
| bubble | balloon | judge | politician | sombrero | cowboy hat |

(*table continues*)

(*Appendix continues*)

**Table A1** (*continued*)

| Target | Foil | Target | Foil | Target | Foil |
|---|---|---|---|---|---|
| burrito | calzone | jukebox | slot machine | spatula | ladle |
| butterfly | moth | jump rope | Hula-Hoop | spider | scorpion |
| cabinets | shelves | ketchup | jam | sponge | scrub brush |
| cake | muffin | knight | samurai | spot | stripe |
| calves | thighs | ladder | staircase | squat | lunge |
| camel | llama | lamppost | spotlight | stable | coop |
| can | jar | lantern | lamp | stadium | theater |
| candle | flashlight | lemon | lime | staples | paper clips |
| cannon | pistol | leprechaun | witch | stem | trunk |
| cannonball | dive | lettuce | celery | strawberry | cranberries |
| canoe | paddleboat | librarian | teacher | stretcher | hospital bed |
| car seat | booster chair | life jacket | life preserver | sun | comet |
| cardinal | blue jay | lighter | matches | sunflower | tulips |
| cash register | adding machine | limbo | tug-of-war | swan | goose |
| casino | arcade | limo | station wagon | swimming | surfing |
| castle | lighthouse | lollipop | candy cane | swimming pool | hot tub |
| caterpillar | centipede | lumberjack | miner | sword | spear |
| cave | tunnel | lungs | kidneys | syrup | honey |
| CD | floppy disk | magazines | brochures | tablecloth | place mat |
| ceiling | roof | magnifying glass | binoculars | tater tots | French fries |
| cell phone | walkie-talkie | mannequin | dummy | taxi | police car |
| chain saw | circular saw | marbles | jacks | teapot | coffeepot |
| champagne | beer | maze | puzzle | tear | sweat |
| cheese | butter | meal | snack | tepee | lean-to |
| chess | backgammon | measuring tape | ruler | telescope | microscope |
| chest | back | medal | trophy | tennis | volleyball |
| chick | bunny | mermaid | centaur | Texas | California |
| cigar | cigarette | microphone | megaphone | thermometer | scale |
| cinnamon | sugar | milkshake | parfait | thermos | flask |
| class | team | mime | jester | thimble | pincushion |
| coconut | kiwi | monkey | sloth | thumb | pinky |
| coffee | tea | moon | planet | ticket | key |
| colosseum | arch | moose | wildebeest | tiger | cheetah |
| comb | brush | mosquito | fly | tile | brick |
| compass | stopwatch | mother | father | tinfoil | Saran Wrap |
| corsage | bouquet | muffin pan | Bundt | tire | wheel |
| cowboy | bullfighter | mummy | zombie | toast | English muffin |
| cowboy boots | rain boots | muzzle | harness | toaster | microwave |
| crane | excavator | nachos | tacos | toilet paper | paper towels |
| crayon | marker | nail | pushpin | tomato | red pepper |
| crib | highchair | nail polish | mascara | tongue | lips |
| crosswalk | sidewalk | napkins | tissues | tooth | bone |
| crow | robin | night | day | tornado | lightning |
| cube | cylinder | notebook | paper | tractor | bulldozer |
| cucumber | squash | nurse | maid | traffic cone | sawhorse |
| cuff link | button | octopus | jellyfish | trail mix | almonds |
| cup | bowl | orchestra | band | tree | cactus |
| cupcake | cookie | organ | piano | trumpet | tuba |
| cupid | fairy | ostrich | flamingo | tutu | kilt |
| cymbals | tambourine | outhouse | shed | twine | thread |
| dandelion | daffodil | owl | eagle | typewriter | keyboard |
| dart | birdie | pacifier | baby bottle | vacuum | lawn mower |
| dentist | masseuse | package | envelope | van | bus |
| desert | beach | paintbrush | toothbrush | vein | spine |
| diamonds | pearls | painter | sculptor | velvet | plaid |
| dice | dominoes | paints | colored pencils | vest | leotard |
| dinosaur | dragon | pantry | liquor cabinet | video camera | tape recorder |
| diploma | report card | party | meeting | visor | hat |
| dishwasher | washing machine | paw | hoof | volcano | butte |
| doghouse | birdhouse | peach | pomegranate | vulture | bald eagle |
| dolphin | killer whale | peacock | turkey | waiter | chef |
| door | window | peas | lima beans | walker | crutches |
| doorbell | knocker | pedicure | manicure | watch | belt |
| doorknob | lever | pen | pencil | waterfall | river |
| dragonfly | praying mantis | penguin | toucan | watermelon | honeydew melon |
| dream catcher | mobile | penny | dime | whale | shark |
| dress | apron | perfume | lotion | wheat | corn |

(*table continues*)

(*Appendix continues*)

**Table A1** (*continued*)

| Target | Foil | Target | Foil | Target | Foil |
|---|---|---|---|---|---|
| dresser | desk | pharmacy | grocery store | whip | fishing pole |
| drill | glue gun | photograph | frame | whiskers | antennae |
| driver's license | credit cards | pianist | DJ | windmill | water mill |
| duffel bag | suitcase | pickles | zucchini | wing | tail |
| dumpster | trash can | picnic | dinner | wolf | hyena |
| dust | cobwebs | pig | hippopotamus | woodpecker | hummingbird |
| dustpan | scoop | pillow | blanket | wrap | sandwich |
| earmuffs | headphones | pilot | captain | wrist | ankle |
| egg rolls | corn dogs | pimple | mole | X-ray | ultrasound |
| elbow | knee | pinecone | walnut | xylophone | harmonica |
| elephant | rhinoceros | pirate | sailor | zebra | pony |
| elevator | escalator | pitcher | Erlenmeyer flask (V1) | zipper | Velcro |

**Table A2**

*Word–Picture Verification Accuracy Mean (Standard Error) per Condition per Module at the 1-Day and 1-Week Retention Test*

| Module | Condition | 1-day test | 1-week test |
|---|---|---|---|
| Comprehension | Retrieval practice | 0.84 (0.03) | 0.76 (0.04) |
| | Restudy | 0.85 (0.04) | 0.78 (0.05) |
| | Untrained | 0.53 (0.05) | 0.56 (0.06) |
| Production | Retrieval practice | 0.58 (0.06) | 0.54 (0.08) |
| | Errorless learning | 0.63 (0.07) | 0.47 (0.06) |
| | Untrained | 0.50 (0.03) | 0.36 (0.04) |

**Table A3**

*Word–Picture Verification Accuracy Mean per Participant per Condition per Module at the 1-Day and 1-Week Retention Test*

| | Comprehension module | | | Production module | | |
|---|---|---|---|---|---|---|
| Condition | Restudy | Retrieval practice | Untrained | Errorless learning | Retrieval practice | Untrained |
| 1-day test | | | | | | |
| Participant | | | | | | |
| P1 | 0.74 | 0.70 | 0.70 | 0.66 | 0.58 | 0.50 |
| P2 | 1.00 | 0.95 | 0.35 | 0.60 | 0.45 | 0.55 |
| P3 | 0.85 | 0.80 | 0.45 | | | |
| P4 | 1.00 | 0.90 | 0.80 | | | |
| P5 | 0.70 | 0.73 | 0.30 | 0.40 | 0.50 | 0.50 |
| P6 | 0.90 | 0.80 | 0.60 | | | |
| P7 | 0.80 | 0.93 | 0.37 | 0.63 | 0.63 | 0.53 |
| P8 | 0.80 | 0.93 | 0.30 | 0.38 | 0.55 | 0.30 |
| P9 | 0.60 | 0.73 | 0.40 | 0.67 | 0.30 | 0.53 |
| P10 | 0.95 | 0.95 | 0.80 | | | |
| P11 | 0.85 | 0.85 | 0.65 | 0.75 | 0.70 | 0.55 |
| P12 | 0.95 | 0.80 | 0.60 | 0.95 | 0.90 | 0.55 |
| 1-week test | | | | | | |
| Participant | | | | | | |
| P1 | 0.70 | 0.56 | 0.66 | 0.34 | 0.30 | 0.24 |
| P2 | 0.80 | 0.75 | 0.45 | 0.45 | 0.55 | 0.30 |
| P3 | 0.75 | 0.85 | 0.50 | | | |
| P4 | 1.00 | 0.90 | 0.90 | | | |
| P5 | 0.53 | 0.60 | 0.20 | 0.43 | 0.37 | 0.47 |
| P6 | 0.95 | 0.90 | 0.75 | | | |
| P7 | 0.67 | 0.67 | 0.47 | 0.50 | 0.67 | 0.40 |
| P8 | 0.78 | 0.88 | 0.30 | 0.33 | 0.53 | 0.23 |
| P9 | 0.47 | 0.50 | 0.47 | 0.27 | 0.23 | 0.33 |
| P10 | 1.00 | 0.95 | 0.85 | | | |
| P11 | 0.80 | 0.70 | 0.45 | 0.70 | 0.85 | 0.45 |
| P12 | 0.90 | 0.90 | 0.70 | 0.75 | 0.85 | 0.50 |

(*Appendix continues*)

**Table A4**

*T Values Estimated With Mixed Linear Models for Pairwise Differences Between Conditions in d (Sensitivity) and β (Bias) as a Function of Module and Retention Test*

| Module and condition | *t* val. (d′)[a] | *t* val. (β) |
|---|---|---|
| Comprehension module: 1-day test | | |
| Restudy (reference level: untrained) | 5.46 | 1.01 |
| Retrieval practice (reference level: untrained) | 4.75 | 1.56 |
| Retrieval practice (reference level: restudy) | −0.72 | 0.55 |
| | *t* val. (d′) | *t* val. (β) |
| Comprehension module: 1-week test | | |
| Restudy (reference level: untrained) | 6.60 | 0.40 |
| Retrieval practice (reference level: untrained) | 4.93 | 0.94 |
| Retrieval practice (reference level: restudy) | −1.66 | 0.54 |
| | *t* val. (d′)[a] | *t* val. (β) |
| Production module: 1-day test | | |
| Errorless learning (reference level: untrained) | 2.32 | −1.79 |
| Retrieval practice (reference level: untrained) | 1.78 | −1.92 |
| Retrieval practice (reference level: errorless learning) | −0.55 | −0.13 |
| | *t* val. (d′) | *t* val. (β) |
| Production module: 1-week test | | |
| Errorless learning (reference level: untrained) | 1.81 | −1.22 |
| Retrieval practice (reference level: untrained) | 2.98 | −1.02 |
| Retrieval practice (reference level: errorless learning) | 1.18 | 0.20 |

*Note*. *t* val. (d′) = *t* value for pairwise contrast in d′ (sensitivity) using linear mixed-model estimation per condition (each training condition relative to the specified reference condition); *t* val. (β) = *t* value for pairwise contrast in β (bias) using linear mixed-model estimation per condition (each training condition relative to the specified reference level). *T* values 2 or higher correspond to significant at $p \leq .05$ by a two-tailed test (Baayen et al., 2008; see Footnote 5, for details).

[a] Estimated with simple linear regression due to nonconvergence of mixed linear model.

**Table A5**

*Individual Differences Analyses: Mixed Linear Model Results in the Comprehension Module and Production Module*

| Model terms | coefficient | *SE* | *t* |
|---|---|---|---|
| Comprehension module: difference (retrieval practice—restudy) in WPV accuracy (collapsed across retention test) | | | |
| Fixed effects | | | |
| Intercept | −0.11 | 0.19 | |
| 1-week test[a] | −0.01 | 0.03 | −0.37 |
| Camel & Cactus Test | 0.006 | 0.003 | 1.66 |
| Synonymy Matching Test | −0.004 | 0.002 | −2.37 |
| Random effects | $s^2$ | | |
| Participants | 0.002 | | |
| Production module: difference (retrieval practice—errorless learning) in WPV accuracy (collapsed across retention test)[b] | | | |
| Fixed effects | | | |
| Intercept | −1.53 | 0.40 | |
| 1-week test[a] | 0.12 | 0.05 | 2.50 |
| Camel & Cactus Test | 0.016 | 0.005 | 3.15 |
| Synonymy Matching Test | 0.004 | 0.005 | 0.873 |

*Note*. WPV = word–picture verification; coefficient = model estimation of the change in difference score per fixed effect; *SE* = standard error of the estimate; *t* = *t*-test statistic; $s^2$ = random effect variance. *T* values 2 or higher correspond to significant at $p \leq .05$ by a two-tailed test (Baayen et al., 2008).

[a] Reference level is a 1-day retention test. [b] Estimated with simple linear regression due to nonconvergence of mixed linear model.

(*Appendix continues*)

**Table A6**

*Foil Accuracy Mean (Standard Error) at Item Selection, 1-Day Test, and 1-Week Test, per Condition per Module*

| Module | Condition | Item selection | 1-day test | 1-week test |
|---|---|---|---|---|
| Comprehension | Retrieval practice | 0.33 (0.05) | 0.85 (0.03) | 0.78 (0.04) |
| | Restudy | 0.35 (0.04) | 0.86 (0.03) | 0.78 (0.05) |
| | Untrained | 0.32 (0.04) | 0.55 (0.06) | 0.58 (0.06) |
| Production | Retrieval practice | 0.24 (0.04) | 0.59 (0.06) | 0.55 (0.09) |
| | Errorless learning | 0.26 (0.04) | 0.63 (0.06) | 0.49 (0.06) |
| | Untrained | 0.25 (0.05) | 0.53 (0.04) | 0.40 (0.05) |

Received March 2, 2022
Revision received May 19, 2022
Accepted June 13, 2022 ■

## Correction to Cockcroft (2022)

In the article “Are Working Memory Models WEIRD? Testing Models of Working Memory in a Non-WEIRD Sample,” by Kate Cockcroft (*Neuropsychology*, 2022, Vol. 36, No. 5, pp. 456–467, https://doi.org/10.1037/neu0000811), in Table 2, for Verbal STM and Verbal WM, the means and standard deviations should have been set in bold but were not, and the median, 25th *p*, and 75th *p* values were set in bold but should not have been. In this table, bold values were those most appropriate for the data distribution. In Table 3, the correlations between the LR subtest and the OOO, MRX, and SR subtests were incorrectly listed as .13, .05, and .18*, respectively. They should have been .26***, .30****, and .27***, respectively. In Table 4, the *df* for the four-factor model was listed as 54 but should have been 48. The online version of this article has been corrected.

https://doi.org/10.1037/neu0000863