Practice-related predictors of semantic feature verification treatment for aphasia [IN PRESS]

Robert Cavanaugh[*1, 2], Yina Quique[3], Michael Walsh Dickey[1, 2], William D. Hula[1, 2], Emily Boss[4], & William S. Evans[1]

[1] Communication Sciences and Disorders, University of Pittsburgh, PA
[2] Geriatric Research Education and Clinical Center, VA Healthcare System, Pittsburgh, PA
[3] Center for Education in Health Sciences, Northwestern University, Chicago, IL, USA
[4] Integrative Reconnective Aphasia Therapy, Pittsburgh, PA, USA

Abstract

Purpose: Specifying the active ingredients in aphasia interventions can inform treatment theory and improve clinical implementation. This secondary analysis examined three practice-related predictors of treatment response in semantic feature verification (SFV) treatment. We hypothesized that: (1) successful feature verification practice would be associated with naming outcomes if SFV operates similarly to standard feature generation semantic feature analysis; (2) successful retrieval practice would be associated with naming outcomes for treated, but not semantically related, untreated words if SFV operates via a retrieval-practice oriented lexical activation mechanism.

Methods: Item-level data from 9 participants with post-stroke aphasia who received SFV treatment reported in Evans, Cavanaugh, Quique et al. (2021) were analyzed using Bayesian generalized linear mixed effect models. Models evaluated whether performance on three treatment components (facilitated retrieval, feature verification, and effortful retrieval) moderated treatment response for treated and semantically related, untreated words.

Results: There was no evidence for or against a relationship between successful feature verification practice and treatment response. In contrast, there was a robust relationship between the two retrieval practice components and treatment response for treated words only.

Conclusions: Findings were consistent with the second hypothesis: retrieval practice, but not feature verification practice, appears to be a practice-related predictor of treatment response in SFV. However, treatment components are likely interdependent, and feature verification may still be an active ingredient in SFV. Further research is needed to evaluate the causal role of treatment components on treatment outcomes in aphasia.

Introduction

Clinical rehabilitation interventions have been criticized as being underspecified in terms of the key treatment components which comprise interventions (Turkstra et al., 2016). Understanding the active ingredients and mechanisms of action in aphasia interventions is vital for increasing treatment efficacy in clinical research, determining treatment candidacy, and implementing treatments successfully in clinical practice (Hart et al., 2019; Van Stan et al., 2019). Understanding the relationship between specific treatment components, hypothesized mechanisms of action, and treatment outcomes can also make important contributions to theoretically-motivated treatments and psycholinguistic theory, by examining which treatment components (and allied psycholinguistic mechanisms) contribute most strongly to observed treatment outcomes (Nickels et al., 2015; Zanca et al., 2019).

One method of investigating active ingredients and mechanisms of action is to examine the relationships between different aspects of performance during treatment and treatment outcomes, such as accuracy on a specific treatment component, the amount of time spent on treatment, or the number of practice attempts. Understanding which practice-related aspects of performance are predictive of treatment outcomes can inform our understanding of treatment theory (i.e., why a treatment works) and help clinicians optimize treatments when implemented in clinical practice (i.e., which ingredients should be prioritized). The present study aims to extend recent work on practice-related predictors of treatment response in semantic feature analysis (SFA) treatment for aphasia (Boyle & Coelho, 1995) by examining a common modification: semantic feature verification (SFV).

Practice related predictors of semantic feature analysis

The classic version of SFA is a semantically oriented treatment focused on the lexical-semantic stage of word retrieval (Boyle, 2010). Individuals with aphasia are asked to generate semantically related concepts (features) in response to a set of feature prompts (such as context: *where is it found?* or use: *what is it used for?*) for a set of target words. Generation of target words and semantically related features is thought to strengthen the connections between conceptual and lexical representations through repeated practice, consistent with the bi-directional excitatory connections that support the first, conceptual-to-lexical step in two-step interactive activation models of lexical access (Dell, 1986; Foygel & Dell, 2000). Successful activation of lexical representations is a prerequisite for the second step in such models, mapping from lexical to phonological representations.

Reinforcement of these connections between conceptual and lexical representations might also increase the spread of activation between semantically related items in the

*Corresponding Author: Robert Cavanaugh: rob.cavanaugh@pitt.edu

Conflicts of Interest: There are no conflicts of interest.

Funding Statement: This research was funded by the VA Pittsburgh Healthcare System Geriatric Research Education and Clinical Center, the VA Healthcare Network- VISN 4 Competitive Career Development Fund, and the VA RR&D service (IK1 RX002475), with funds awarded to William S. Evans.

lexical-semantic network, via feedback from activated (target) lexical representations to associated conceptual representations. This in turn engenders feedforward activation from those conceptual representations to other semantically related lexical items. Repeated feature generation and target naming may further increase the resting activation for items within a semantic category, on the assumption that these items have similar conceptual representations to the target, thereby improving the ability to retrieve words within a treated semantic category. As shared semantic features become more accessible through repeated activation and greater resting activation, lexical access to both trained and semantically related, untrained words should improve. This account predicts that by training a limited subset of words, SFA can induce response generalization and improve lexical access to a broader range of semantically related words, resulting in more widespread improvements in word finding.

In an interim analysis (n = 18) of a recently completed, intensive SFA clinical trial, Gravier et al. (2018) examined four practice-related predictors of treatment response: the total number of hours of treatment, the total number of trials, the average number of trials per hour, and the average number of patient-generated features per trial. Gravier and colleagues found that only the average number of patient-generated features per trial was predictive of treatment response for both treated and semantically related, untrained words at treatment exit and one month follow-up. Gravier et al. (2018) interpreted this finding as evidence that successfully generating features was likely to be an active ingredient in SFA, consistent with the hypothesized treatment theory of SFA. In Evans, Cavanaugh, Gravier, et al. (2021), we updated this analysis with the full study sample (n = 44), and reported that the findings held for treated words, but were no longer statistically significant for semantically related, untreated words at the group level. Although the relationship between feature generation and outcomes (particularly for treated words) was strong, the study design did not permit us to evaluate whether it was causal. This question is being tested empirically in an ongoing comparative-effectiveness clinical trial, in which the number of features generated by participants is directly manipulated (NCT04215952).

Modifications to semantic feature analysis

The popularity of the original SFA studies by Boyle and colleagues has spurred many variants and modifications (Boyle, 2010). For example, the target stimuli and intervention components of SFA have been altered to focus on verb and action retrieval (Kristensson et al., 2014; Wambaugh & Ferguson, 2007), word typicality (Gilmore et al., 2020; Kiran, 2008), and concrete vs. abstract concepts (Sandberg & Gray, 2020). SFA has also been implemented within group settings (Antonucci, 2009; Peach & Reuter, 2010).

SFV (sometimes called semantic feature review) is another modification of the classic SFA paradigm. SFV does not require individuals with aphasia to generate semantic features; instead, individuals listen to and/or read semantic feature descriptions of target words and verify whether they are correct or incorrect. Verbal productions of the target word are interleaved around feature verification trials. Feature verification may be advantageous compared to feature-generation SFA in that it lessens production demand, which can be frustrating and time-consuming (thereby limiting treatment dose) for individuals with more severe aphasia. Previous work reported that an individual with aphasia's ability to generate

5/11 features per trial on average throughout feature generation SFA appeared to be a minimum criterion for positive treatment response (Evans, Cavanaugh, Gravier, et al., 2021). Thus, SFV may broaden candidacy requirements in SFA by circumventing the need to generate features while still providing an opportunity to engage mechanisms of spreading activation and potentially strengthen connections between conceptual and lexical representations.

Compared to standard SFA, studies of SFV have generally reported the same pattern of effects with regard to both direct training and response generalization to semantically related words. Edmonds and Kiran (2006) and Kiran and Roberts (2010) successfully implemented the feature verification paradigm in bilingual individuals with aphasia and reported positive treatment effects on both trained and related, untrained words. Gilmore and colleagues (2020) utilized a typicality-based SFV treatment that used written and auditory feature verification, finding generalization from trained atypical words to untrained, typical words within a semantic category. Recently, we deployed SFV in a pilot study with nine (monolingual) individuals with aphasia in a computer game-based treatment, which sought to improve accuracy and response speed on naming and verification tasks (Evans, Cavanaugh, Quique, et al., 2021). In this study, we replicated the expected findings on probe accuracy: participants demonstrated clear improvements in trained words, and small, though reliable generalization effects to semantically related, untreated words (a change that we were not able to distinguish statistically from rising baseline performance or from untreated, semantically unrelated words).

As Boyle (2010) stated in her review, altering treatment components (and potentially the active ingredients) may also change the underlying mechanisms of action which engender treatment response. Although SFV and SFA operate under the same assumption that processing semantic features will strengthen the connections between conceptual and lexical representations, there are a number of differences in how these treatments might do so, which have theoretical and clinical implications. For feature generation, input processing consists of a number of semantic category prompts (e.g., *what is it used for?*) presented with a target word, and output demands consist of verbally producing a relevant semantic feature that matches the target. For feature verification, the input consists of potential semantic features for the target word presented in auditory and written format (e.g., *knife; used for slicing onions?*), while output consists of providing a binary yes/no verification response as to whether or not the provided feature matches the target. Essentially, both task components activate semantic information about the target word but differ in whether features are elicited by a general category cue or directly provided and in their output demands. Despite the theoretical and practical implications of these differences, relatively little work has evaluated the practice-related predictors of SFV, which is the focus of the present study.

The present study

A first step in understanding potential mechanistic differences between SFA and SFV is to evaluate whether the practice-related predictors of treatment response in SFV are similar to SFA (i.e., is successful feature verification predictive of treatment response, similar to successful feature generation). Where mechanisms of action are consistent across treatments, we propose that we should see similar patterns of association between treatment components and treatment response. Therefore, this study examines the relationships between three practice-related predictors of treatment response in Evans, Cavanaugh, Quique et al. (2021), which piloted a SFV treatment called *BEARS+SFV* (Balancing Effort, Accuracy, and Response Speed). In addition to the reported results, Evans, Cavanaugh, Quique et al. (2021) found a great deal of variability in treatment outcomes, providing an optimal environment to evaluate predictors of treatment response in the present study (Hedge et al., 2018).

The treatment paradigm in BEARS+SFV consisted of four core components that may contribute to treatment response, three of which are evaluated here: effortful retrieval practice, feature verification, and a second facilitated retrieval attempt, primed by feature verification (the fourth component, meta-cognitive strategy training and feedback focused on optimizing speed-accuracy tradeoffs during naming response, is not analyzed here; see discussion). As discussed, processing and verifying semantic features is thought to promote improvements on treated words and generalization to semantically related, untreated words in this type of semantic anomia treatment. However, recent work on learning in aphasia suggests that treatment response in our study may also result from successful retrieval practice in the two naming components. Effortful retrieval practice (i.e., confrontation naming without cues or other supports) has been shown to produce robust improvements in naming ability for trained words (Middleton et al., 2016). Retrieval practice has also been studied in the context of cueing hierarchies (e.g., Abel et al., 2005), and priming studies (e.g., Tabrizi et al., 2021), where treatment paradigms facilitate successful naming by reducing retrieval effort. These studies of facilitated retrieval practice also found improvements in naming, with the added advantage that facilitated retrieval practice may be more suitable for individuals with more severe anomia who perform poorly on effortful naming tasks.

In general, retrieval practice, with varying degrees of effort and cueing, is thought to strengthen both lexical-semantic and lexical-phonological connections, thereby improving word retrieval specifically for trained words. While retrieval practice has not been the theoretical focus of SFA or SFV, it is almost always included as a treatment component. Because feature verification in SFV consumes much less time than feature generation in SFA, retrieval practice dose may in fact be greater in SFV than SFA and thus play a more prominent role in facilitating treatment response. This theoretical foundation led to the following hypotheses:

(1) *Hypothesis 1*: If SFA and SFV engender treatment response under the same spreading activation mechanism, successful feature verification practice should be positively associated with naming outcomes for treated and semantically related, untreated items in SFV.

(2) *Hypothesis 2*: If SFV engenders treatment response through retrieval-practice oriented lexical activation, successful retrieval practice (both effortful and facilitated) should be positively associated with naming outcomes for treated words, but not untreated words.

Methods

Participants

Nine people with chronic aphasia participated in a computer-administered SFV treatment called BEARS ("Balancing Effort, Accuracy, and Response Speed") in a single-subject design with multiple baselines across participants. As this is a secondary analysis of a larger study, additional participant characteristics and study details are reported in the article reporting the full clinical trial (Evans, Cavanaugh, Quique, et al., 2021). Participants were recruited from the Western Pennsylvania Research Registry, the Audiology and Speech Pathology Research Registry maintained by the VA Pittsburgh Healthcare System (VAPHS), and clinician referrals. Inclusion criteria were as follows: participants were required to be at least six months post-onset of their stroke, have a diagnosis of aphasia by a certified speech-language pathologist, be community-dwelling, at least wheel-chair ambulatory, have English as their native language, and be older than 18 years of age. Participants were also required to demonstrate less than or equal to 50% correct performance on at least 80 treatment item probes during the stimuli selection phase (discussed below). Potential participants were excluded if they had a history of neurodegenerative disease, active, unmanaged psychopathology or alcohol/substance abuse, severe motor speech disorder (i.e., apraxia of speech or dysarthria), or were participating in any other speech/language therapy during the time of the study. Participants were also excluded if they presented with severe anomia, as measured by a Comprehensive Aphasia Test (Swinburn et al., 2004) naming modality T-score of less than 40. A total of 13 participants were recruited for the study, with 9 meeting enrollment criteria. Participants included 8 males and 1 female with an average age of 67.9 years and an average of 125.3 months post-stroke onset. All participants had a diagnosis of aphasia following left-hemisphere cerebrovascular accident and an average CAT mean modality T-score of 51.6. Individual participant demographics are summarized in Table 1.

Participant	Sex	Age	MPO	Years Edu	Race/ ethnicity	Hemiparesis	CAT mean modality T-score
p1	M	63	259	14	African American	R UE	49
p2	M	73	194	14	Caucasian	None	50.5
p3	M	68	21	12	Caucasian	None	57.33
p4	M	70	522	13	Caucasian	R UE	52.23
p5	M	70	39	16	Caucasian	None	52
p6	F	71	8	14+ years	African American	None	54.5
p7	M	70	9	18+ years	Caucasian	R UE and LE	46
p8	M	54	18	16	African American	None	51.33
p9	M	72	58	14	Caucasian	None	52

Table 1. Participant descriptive characteristics (reprinted from Evans, Cavanaugh, Quique et al., 2021).

Stimuli

Treatment and probe stimuli for study participants consisted of picturable nouns from two photographic databases (Brodeur et al., 2010, 2014; Moreno-Martínez & Montoro, 2012). Stimuli were balanced within semantic categories based on an item complexity algorithm from Fergadiotis et al. (2015). In this approach, item complexity was estimated using the following equation: B = -1.22 - 0.36 (log word frequency) + 0.21 (age of acquisition) + 0.15 (number of phonemes), which they reported to account for 63% of variance in naming difficulty. Categories were chosen for functional relevance, resulting in 233 items across 15 semantic categories eligible for treatment.

For each participant, treatment lists were generated based on performance on a confrontation picture-naming task. Pictures were presented one at a time, and participants were given 15 seconds to name the picture. The first complete response given was judged as correct or incorrect based on the Philadelphia Naming Test rules (Roach et al., 1996). If a participant indicated that they did not know what the picture was, the item was marked as incorrect but not selected as a treatment target. Each participant was administered the naming task on two separate occasions. Words that were named at or below 50% accuracy across both administrations were included as potential treatment items. Semantic categories had to have at least eight qualifying items to be included for treatment. A total of 5 semantic categories with eight qualifying items in each category were selected for treatment for a total of 40 treatment targets. Selection was based on clinician knowledge of patient's interests, patient preference, and quality of stimuli pictures (some images consisted of dated depictions of the target or were of poor image quality). Following stimuli selection, 20 of the treatment stimuli were paired with their semantically related word, and 20 treatment stimuli were paired with an unrelated word of similar complexity (using the item complexity algorithm noted above) to generate a list of 40 total generalization items. Therefore, each participant had a unique set of 40 treatment items and 40 untreated items (20 semantically related).

Probe Administration

Treatment was administered in a single case-series design with multiple baselines across participants. Baseline probe performance was established via multiple probe assessments, with the number of baseline probes pseudorandomly assigned (3, 4, or 5 sessions) to better localize the onset of treatment-related changes. Following baseline, treated and untreated words were assessed at the beginning of each treatment session before initiating treatment, within one week of finishing treatment, and at one-month follow-up.

##Treatment paradigm

Treatment consisted of the computer-based implementation of SFV, augmented by meta-cognitive strategy training designed to help participants improve their self-monitoring of speed-accuracy tradeoffs. Meta-cognitive training consisted of education and feedback from the clinician and points-feedback from the computer software. Participants received this treatment from a certified speech-language pathologist 3-4 days per week over 3-4 weeks, for a total of 25 hours.

The SFV treatment consisted of three major components in each trial. First, participants were presented with the target picture and asked to name the item with a single verbal response without cueing. Second, the target picture was shown again along with an audio recording of the correct response, followed by the auditory and written presentation of four semantic feature verification questions (e.g., "*greyhound, has long fur?*"), randomly selected from a list of eight questions for each item (four correct and four incorrect). In this component, participants were asked to indicate whether the feature was correct via computer button press. After the four questions were completed, participants were presented with the item again and asked to name it. One trial was completed for all eight items within a semantic category until all 5 semantic categories for the participant were completed, after which the paradigm started again. Performance accuracy was logged online by the clinician during treatment, and feedback was provided to the participant for every response.

Because the first component was comprised of a confrontation naming attempt without cueing or other facilitation, we considered it to align closely with existing research on effortful retrieval practice (Middleton et al., 2016). In contrast, the second naming attempt followed an initial naming attempt with feedback and four feature verification questions in which the target word-form was provided in auditory and written format. Therefore, the second naming attempt approximated a naming task with both semantic and phonological priming and had higher accuracy in comparison to the first naming attempt. Thus, we considered these components to compose three distinct practice-related variables: effortful retrieval practice (practice step 1), feature-verification practice (practice step 2), and facilitated retrieval practice (practice step 3).

Analysis

In the primary analysis of this pilot clinical trial (Evans, Cavanaugh, Quique, et al., 2021), we found rising baseline performance at the group level, which was driven by a subset of participants and items and which we attributed to regression to the mean after initial stimuli selection. In order to examine practice-related predictors for items that clearly improved during treatment, we limited the data to only include items with a mean performance of less than or equal to 50% accuracy, corresponding to our pre-baseline item selection criteria. This resulted in a subset of 75% of treated (80%) and related, untreated (65%) probe items for analysis. Consequently, the current secondary analysis results should shed light on associations between feature verification, retrieval practice, and improvement specifically for words that remained challenging for participants throughout the baseline phase. For overall estimates of treatment effects on treated, untreated semantically related, and untreated semantically unrelated words from this pilot trial, we refer readers to the main report (Evans, Cavanaugh, Quique, et al., 2021).

Data were analyzed using item-level Bayesian generalized linear mixed-effects models via R statistical software (R Core Team, 2020) with the BRMS package (Bürkner, 2018). The deidentified dataset and analysis code used in the current study are publically available at https://osf.io/dxvyj/. The mixed-effects modeling approach is well-suited to this study, as it is able to handle the complex dependency structure within the data (repeated measurements of items across subjects) and leverage partial pooling (i.e., sharing information across items and participants) to make more generalizable estimates (McElreath, 2020). The Bayesian implementation of mixed-effects models is advantageous because it aligns with how researchers typically understand statistical results (i.e., interpreting the probability of the hypothesis, given the data), characterizes uncertainty in the effects of interest (e.g., a 90% credible interval is interpreted as a 90% probability that the effect falls within the interval, given the data), and typically permits inclusion of more complex random-effect structures that characterize the dependencies in the data (see Nalborczyk et al., 2019 for a tutorial on Bayesian mixed-effects modeling in communication sciences and disorders).

Models were run separately for each practice component and item condition (treated and semantically related, untreated). The model structure was intentionally mapped to the previous analysis focused on SFA (Gravier et al., 2018) for comparison. Model fixed effects included timepoint, the total number of successful practice attempts for a given treatment component, and their interaction. We defined timepoint as a three-level variable consisting of treatment baseline, exit, and follow-up (reference: baseline). *Treatment baseline* included the probe sessions within the baseline phase, treatment *exit* included the probe within 1 week of finishing the treatment, and treatment *follow-up* was conducted 1 month after treatment ended. The practice component fixed effect consisted of the total number of correct trials and was centered and z-scored. As in Gravier et al. (2018), aphasia severity (z-scored and centered Comprehensive Aphasia Test mean modality T-score) was included as a covariate.

Based on this model structure, support for hypothesis 1 would consist of a reliable interaction effect between timepoint and successful feature verification practice indicating that successful feature verification practice is positively associated with naming outcomes for treated and semantically related, untreated items in SFV. Alternatively, support for hypothesis 2, would consist of a reliable interaction between timepoint and either successful retrieval practice component, indicating that successful retrieval practice is positively associated with naming outcomes for treated words but not untreated, related words.

Models were fit systematically with a Bernoulli probability distribution (0 or 1), beginning with the maximal model and subsequently reducing the random effects structure to ameliorate non-convergence based on existing recommendations (Barr et al., 2013; Matuschek et al., 2017). Final models included crossed random effect structures: random intercepts for participants with a random slope for timepoint, practice accuracy, and their interaction and random intercepts for words. These random effects structures allow both participants and items to vary in their intercept, while each participants' performance is permitted to vary with regards to time, successful practice, and their interaction.

Dispersed starting values were assigned to four Markov chains. Each chain was run with 1000 iterations as a warm-up, which were not included in the estimation of each parameter. Parameter estimation was based on 2000 samples per chain post warm-up. Model convergence was assessed by ensuring the potential scale reduction factor was less than 1.05 for each model (Gelman et al., 2013), and the bulk and tail effective sample size was >400. Trace plots were also examined to evaluate chain convergence. Posterior predictive checks were conducted by plotting the observed responses and 500 simulated distributions from the posterior distribution of the Bayesian model. After plotting the simulated distributions, we found that they followed the response patterns of the observed distributions of correct responses. Given these model checks, we concluded that all models adequately fit our data.

Results

Treated items

Results from the three models for treated items are reported in Table 2. For all models, coefficients are reported in logits (the log odds of a correct response). For the feature verification model, there were robust main effects of timepoint, indicating that words that were difficult during the baseline phase improved substantially from treatment baseline to treatment exit (B = 3.18, 90% CI: [2.33, 3.91]) and from treatment baseline to treatment follow-up (B = 2.52, 90% CI: [1.73, 3.19]). However, the interaction between the number of correct features verified and timepoint was not reliably different from zero for baseline to exit (B = 0.11, 90% CI: [-0.67, 0.79]) or baseline to follow-up (B = 0.41, 90% CI: [-0.27, 1.05]), providing little evidence that additional successful feature verification attempts translated to greater improvements in naming accuracy.

For the facilitated retrieval model, there were similar main effects of timepoint. However, the 90% credible interval for the interaction between the number of correct facilitated retrievals and timepoint nearly excluded zero from baseline to treatment exit (B = 0.75, 90% CI: [-0.03, 1.65]) and did exclude zero from baseline to treatment follow-up (B = 0.85, 90% CI: [0.14, 1.58]). The posterior probability that these interaction estimates were greater than zero were .94 and .98, respectively, providing moderate evidence that more correct facilitated retrieval attempts were associated with greater improvements in naming accuracy. For the effortful retrieval model, these interaction estimates were relatively similar in size, but with narrower credible intervals (baseline to exit: B = 0.87, 90% CI: [0.28, 1.43]; baseline to follow-up: B = 0.89, 90% CI: [0.32, 1.46]), indicating greater certainty around the finding that the number of successful effortful retrieval attempts during treatment was associated with improvement in naming accuracy from baseline to exit and follow-up.

Untreated items

Results from the three models for semantically related, untreated items are reported in Table 3. For all models, coefficients are again reported in logits. For the feature verification model, there was a main effect of timepoint from baseline to treatment exit (B = 1.57, 90% CI: [0.85, 2.24]), indicating that related, untreated items which were difficult to name at baseline improved from baseline to treatment exit, but this effect was less certain from baseline to follow-up (B = 0.85, 90% CI: [-0.13, 1.71]). This main effect was similar across the facilitated and effortful retrieval models. However, none of the interactions between successful practice attempts across components and timepoint were reliably different from zero, providing uncertain evidence for a relationship between practice-related performance and response generalization.

Discussion

The overarching purpose of the current study was to extend recent work on practice-related predictors of treatment response in semantic feature analysis treatment by examining practice-related predictors of treatment response in SFV treatment. We first hypothesized that successful feature verification practice would be associated with naming outcomes for both treated and semantically related, untreated items if SFA and SFV

Model	Parameter	Estimate	Est. Error	90% CI
Feature Verification	Population level effects			
	Intercept	-2.77	0.32	-3.30, -2.27
	Num. correct (centered, z)	0.12	0.319	-0.40, 0.63
	Timepoint baseline v exit	3.18	0.489	2.33, 3.91
	Timepoint baseline v followup	2.52	0.455	1.73, 3.19
	CAT mean T-score (centered, z)	0.14	0.358	-0.43, 0.73
	Num. correct x timepoint baseline v exit	0.11	0.451	-0.67, 0.79
	Num. correct x timepoint baseline v followup	0.41	0.408	-0.27, 1.05
Facilitated Retrieval	Population level effects			
	Intercept	-2.67	0.361	-3.27, -2.10
	Num. correct (centered, z)	0.8	0.348	0.22, 1.35
	Timepoint baseline v exit	3.18	0.443	2.40, 3.85
	Timepoint baseline v followup	2.54	0.545	1.59, 3.37
	CAT mean T-score (centered, z)	-0.12	0.405	-0.81, 0.51
	Num. correct x timepoint baseline v exit	0.75	0.509	-0.03, 1.65
	Num. correct x timepoint baseline v followup	0.85	0.442	0.14, 1.58
Effortful Retrieval	Population level effects			
	Intercept	-2.56	0.338	-3.10, -2.00
	Num. correct (centered, z)	1.36	0.25	0.97, 1.78
	Timepoint baseline v exit	3.43	0.343	2.85, 3.94
	Timepoint baseline v followup	2.67	0.354	2.09, 3.21
	CAT mean T-score (centered, z)	-0.52	0.341	-1.08, 0.03
	Num. correct x timepoint baseline v exit	0.87	0.357	0.28, 1.43
	Num. correct x timepoint baseline v followup	0.89	0.356	0.32, 1.46

Table 2. Bayesian generalized linear mixed effects model results for feature verification, facilitated retrieval, and effortful retrieval for treated words.

operate under the same mechanism. However, the results of this analysis do not support this claim. There was no evidence for an association between successful feature verification practice and treatment outcomes. This finding contrasts with our previous work using the feature-generation variant of SFA, where successful feature generation was associated with naming outcomes (Evans, Cavanaugh, Gravier, et al., 2021; Gravier et al., 2018).

Instead, we found evidence supporting the second hypothesis: there was a relationship between facilitated and effortful retrieval practice and naming improvement from treatment baseline to exit and 1-month follow-up for treated words. While this relationship was more certain for effortful retrieval (i.e., a narrower posterior distribution), the effect sizes for both conditions were similar, indicating that a 1 standard deviation increase in the number of successful retrieval attempts was associated with at least a two-fold increase in the odds of a correct response from baseline to treatment exit and 1-month follow-up. The relationship between either retrieval practice component and change in semantically related, untreated words was not reliably different than zero, also consistent with hypothesis 2. However, the

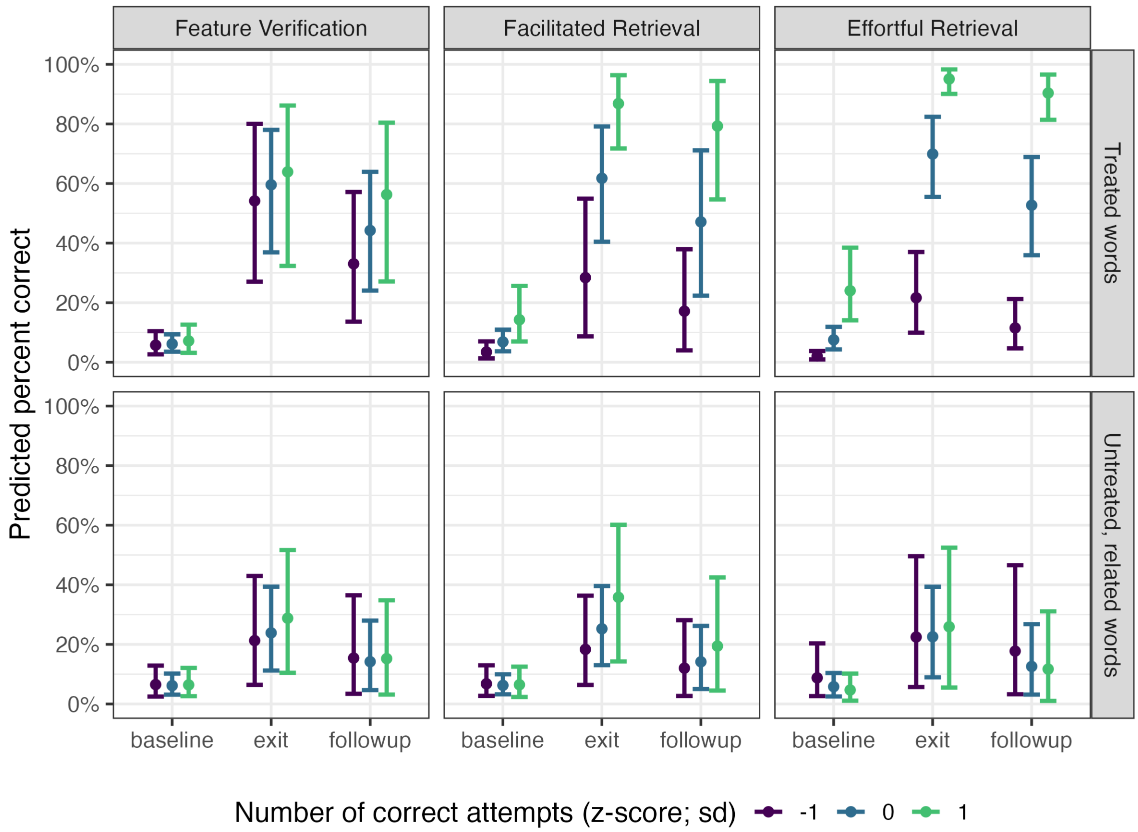

Figure 1. Relationship between successful feature verification and retrieval practice and naming performance at treatment baseline, exit, and 1-month follow-up. Error bars reflect 90% credible intervals.

substantial width of the posterior distributions for these interaction coefficients does not support any strong conclusions about the lack of an effect.

No evidence for a relationship between feature verification and treatment response

There are a number of possibilities for the lack of a statistically robust relationship between successful feature verification performance and naming outcomes (treated and semantically related, untreated). First, the task requirements of the feature-verification component of SFA in this study (i.e., answering verification questions via button-press in a speed-accuracy focused anomia treatment game) may be too far removed from the probe task requirements (i.e., verbal confrontation picture naming) to increase the ability of participants to produce trained words effectively. Evidence from learning theory indicates that performance can transfer between different types of tasks, in part as a function of their similarity (Gick & Holyoak, 1989). In this context, it may be the case that feature

Model	Parameter	Estimate	Est. Error	90% CI
Feature Verification	Population level effects			
	Intercept	-2.79	0.389	-3.43, -2.18
	Num. correct (centered, z)	0	0.349	-0.56, 0.56
	Timepoint baseline v exit	1.57	0.431	0.85, 2.24
	Timepoint baseline v followup	0.85	0.57	-0.13, 1.71
	CAT mean T-score (centered, z)	0.16	0.37	-0.43, 0.77
	Num. correct x timepoint baseline v exit	0.23	0.438	-0.48, 0.93
	Num. correct x timepoint baseline v followup	-0.01	0.51	-0.88, 0.77
Facilitated Retrieval	Population level effects			
	Intercept	-2.77	0.369	-3.40, -2.20
	Num. correct (centered, z)	-0.03	0.386	-0.67, 0.58
	Timepoint baseline v exit	1.63	0.385	1.01, 2.24
	Timepoint baseline v followup	0.85	0.521	-0.06, 1.61
	CAT mean T-score (centered, z)	0.19	0.366	-0.40, 0.78
	Num. correct x timepoint baseline v exit	0.52	0.426	-0.19, 1.20
	Num. correct x timepoint baseline v followup	0.33	0.522	-0.51, 1.17
Effortful Retrieval	Population level effects			
	Intercept	-2.88	0.463	-3.66, -2.15
	Num. correct (centered, z)	-0.34	0.535	-1.30, 0.42
	Timepoint baseline v exit	1.56	0.455	0.78, 2.25
	Timepoint baseline v followup	0.75	0.625	-0.34, 1.65
	CAT mean T-score (centered, z)	0.42	0.556	-0.39, 1.40
	Num. correct x timepoint baseline v exit	0.43	0.499	-0.39, 1.21
	Num. correct x timepoint baseline v followup	0.02	0.652	-1.17, 0.94

Table 3. Bayesian generalized linear mixed effects model results for feature verification, facilitated retrieval, and effortful retrieval for semantically related, untreated words.

verification is not sufficiently similar to confrontation naming to engender an effect, even on trained items. Instead, these findings suggest that it may be more effective to elicit lexical-semantic processing in the context of speech production directly, as in feature generation treatment activities. This account naturally predicts that transfer may be more likely for feature-generation variants of SFA, where there is more similarity between treatment and naming-probe tasks. Future comparative effectiveness studies could test this task-similarity effect in anomia treatments by varying treatment components (e.g., feature generation, reading, or repetition) within an intervention.

Second, it is possible that the feature-verification component in *BEARS+SFV* was not sufficiently difficult to elicit activation of the underlying semantic networks. Accuracy for the feature verification component was relatively high across participants (mean: 92.1% correct; range: 86.5% - 97.6%), and participants completed more than 50,000 feature-verification attempts in total throughout treatment. Although participants responded to a random 4 out of 8 possible feature verification questions for each trained word in each trial, clinical observations and discussions with the treating clinician (author E.B.) raised the possibility that some participants may have learned the questions over time and increasingly answered the feature-verification questions by rote stimulus-response mappings rather than fully processing the semantic feature. This contrasts with SFA as employed by Gravier et al.,(2018), which included the generation of 11 features per trial, though it was evident that some participants repeatedly generated the same features throughout treatment, which raises similar questions about rote stimulus-response mappings in SFA (Evans, Cavanaugh, Gravier, et al., 2021).

Similarly, the emphasis on speed and accuracy throughout all treatment components in *BEARS+SFV* may have reduced the depth of semantic processing, as evidenced by decreasing response latency on the feature-verification task throughout treatment (Evans, Cavanaugh, Quique, et al., 2021). However, increasing automaticity is a natural consequence of task learning in any study. Increasing the difficulty or variety of feature verification questions in future studies which employ SFV may increase the depth of processing by reducing speed and automaticity, but it is unclear whether these changes would be sufficient to improve generalization outcomes in SFV. Finally, although there was variability across items in terms of the number of correct feature verification practice trials per item, the relatively strong performance on feature verification does raise the possibility that task performance was too similar across items to obtain robust covariance with treatment outcomes.

No evidence for a relationship between feature verification or retrieval practice and generalization to related, untreated items

While there was a statistically reliable improvement in performance for semantically related, untreated words that were difficult at baseline, there was no evidence that any of the three components moderated change for semantically related items. Credible intervals for the interaction between practice components and timepoint were relatively wide, indicating a large degree of uncertainty around the potential moderating effects of practice and generalization to semantically related, untreated words. The degree of uncertainty may indicate substantial variation in the moderating effect or simply reflect the smaller sample size for semantically related, untreated words.

These findings neither provide evidence for nor rule out the possibility that the spreading-activation mechanisms that motivated SFA in combination with lexical-retrieval mechanisms play a role in these response generalization effects. It remains possible that the combination of retrieval practice and feature activation through verification could have boosted activation of treated words independent of feature verification accuracy, in turn improving retrieval of related, untreated items. Consistent with this possibility, individuals with aphasia who are better able to activate lexical-semantic representations for word retrieval (i.e., individuals with relatively spared semantic abilities) have demonstrated stronger response generalization in SFA (Dickey et al., 2016). Additional research is required to tease out practice-related predictors of response generalization in SFV and other semantic treatment studies.

Practice related predictors of semantic feature verification outcomes

The robust association between retrieval practice and naming outcomes suggests that performance on either retrieval practice component may be a practice-related predictor of treatment response in SFV. This finding is not surprising but is nevertheless important to confirm empirically. Even though the facilitated practice component likely required less effort as a consequence of the treatment paradigm, both retrieval practice treatment components largely resembled the confrontation naming probe task. Thus, near transfer of performance from retrieval practice to probe naming requires minimal changes in task demands, especially for the effortful retrieval component. This relationship supports existing findings in the retrieval practice literature that successful retrieval practice can engender improvements in lexical retrieval (Middleton et al., 2016) and further suggests that it may play a key role in treatments like SFA and SFV, even if it is not the focus of the treatment.

Successful retrieval practice, both effortful and facilitated, also predicted retention of treatment effects at one month after treatment, suggesting that including retrieval practice treatment components may be one potential solution for modest maintenance of therapy gains across the aphasia literature (Menahemi-Falkov et al., 2021). Similarly, these findings suggest that ongoing assessment of individuals' ability to produce treatment targets in a retrieval (i.e., uncued) context might predict post-treatment retention and inform decisions about when to discontinue treatment in clinical settings. In other words, individuals who are not able to successfully produce treatment targets under uncued conditions throughout treatment appear less likely to improve as a result of treatment or retain any improvements. In the BEARS+SFV study, participants who demonstrated the most difficulty with the retrieval practice component also had the most severe anomia, apraxia of speech, and/or impaired semantic control (see Evans, Cavanaugh, Quique et. al., 2021 for a more detailed explanation of candidacy in BEARS + SFV) and the smallest improvements in naming accuracy at treatment exit and follow-up. However, this secondary analysis is not designed to confirm these causal relationships between retrieval practice and naming outcomes.

If the results of this analysis had been consistent with hypothesis 1, they would have lent support to a shared mechanism between SFA and SFV, substantiating consistent treatment theory between these related but meaningfully different semantic treatment approaches. However, the lack of a robust relationship between feature verification practice and treatment response doesn't invalidate the hypothesized mechanism of action in SFV. Instead, it suggests that feature verification may not be a particularly strong practice-related predictor of treatment response. Behavioral treatments for aphasia are complex and holistic, and often the combination of multiple treatment components results in therapy gains that exceed the sum of the parts. This fact underlines the importance of carefully considering the interactions between different treatment components. For example, discussions with the treating clinician (author E.B.) raised the possibility that the facilitated retrieval practice treatment component may have provided an intermediate level of practice difficulty, subsequently facilitating independent retrieval of challenging words. In other words, the feature verification component provided priming that supported less effortful, facilitated retrieval practice, which may have helped participants with more impaired naming abilities maintain confidence and engage with the treatment despite the difficulty of the effortful treatment component.

More generally, this set of observations points to an important distinction between active ingredients and practice-related predictors. Although practice-related predictors of treatment response are likely to be active ingredients (or at least strongly correlated with active ingredients), the absence of a strong association between a given predictor and treatment outcomes does not preclude that component's role as an active ingredient and importance to engendering positive treatment response. Finally, we suggest that the secondary analyses employed in Gravier et al. (2018), Evans, Cavanaugh, Gravier et al. (2021), and the present study provide a fruitful approach to beginning to tease out the complex relationships between treatment components and outcomes. Together, these findings suggest that accurate, successful verbal naming as a component of semantically based anomia treatment is a robust predictor of treatment response. More importantly, they lay the empirical groundwork needed to support comparative effectiveness research to examine causal relationships between treatment components and outcomes.

Limitations

None of the condition manipulations in this secondary analysis are experimental, and as such, we cannot infer any strong causal relationships. Although we evaluated each component separately, the effects of each component are not truly dissociable, and the results reported herein are merely correlational. Experimental manipulation of treatment components in SFA is the focus of ongoing work (NCT04215952). Furthermore, we are not aware of any empirical comparisons of SFA and SFV treatment components that could control for other differences between the two treatments (e.g., the number of features verified or generated per trial) and provide a more controlled comparison of the potential practice-related predictors and active ingredients in SFA versus SFV. SFV studies in particular often have variations, such as focusing on bilingualism in aphasia or the effects of typicality (Edmonds & Kiran, 2006; Gilmore et al., 2020; Kiran & Roberts, 2010). In the current case, it is unclear how the meta-cognitive training strategy employed in BEARS might have affected treatment response in this study relative to other feature verification paradigms.

Conclusions

The purpose of this study was to evaluate practice-related predictors of treatment response in SFV treatment for aphasia. Although successful feature verification practice was not associated with treatment gains, success on both facilitated and effortful retrieval practice components appeared to moderate treatment response for trained words. The method employed here can provide preliminary evidence for potential ingredients and mechanisms in aphasia treatments and support future work focusing on a comparative evaluation of different treatment components necessary for causal inference. Such careful study of treatments in aphasiology will inform treatment theory and effective implementation of research to clinical practice where pragmatic constraints require a clear understanding of essential treatment components (Zanca et al., 2019).

References

Abel, S., Schultz, A., Radermacher, I., Willmes, K., & Huber, W. (2005). Decreasing and increasing cues in naming therapy for aphasia. *Aphasiology*. https://doi.org/10.1080/02687030500268902

Antonucci, S. M. (2009). Use of semantic feature analysis in group aphasia treatment. *Aphasiology*, *23*(7–8), 854–866. https://doi.org/10.1080/02687030802634405

Barr, D. J., Levy, R.,Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Boyle, M. (2010). Semantic feature analysis treatment for aphasic word retrieval impairments: What's in a name? *Topics in Stroke Rehabilitation*, *17*(6), 411–422. https://doi.org/10.1310/tsr1706-411

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

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). 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*. https://doi.org/10.1371/journal.pone.0106953

Bürkner, P. C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *R Journal*. https://doi.org/10.32614/rj-2018-017

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283.

Dickey, M. W., Doyle, P. J., Gravier, M., & Hula, W. D. (Eds.). (2016). *Psycholinguistic predictors of treatment response in Semantic Feature Analysis.*

Edmonds, L. A., & Kiran, S. (2006). Effect of semantic naming treatment on crosslinguistic generalization in bilingual aphasia. *Journal of Speech, Language, and Hearing Research*, *49*(4), 729. https://doi.org/10.1044/1092-4388(2006/053)

Evans, W. S., Cavanaugh, R., Gravier, M. L., Autenreith, A. M., Doyle, P. J., Hula, W. D., & Dickey, M. W. (2021). Effects of Semantic Feature Type, Diversity, and Quantity on Semantic Feature Analysis Treatment Outcomes in Aphasia. *American Journal of Speech-Language Pathology*, *30*(1S), 344–358. https://doi.org/10.1044/2020_AJSLP-19-00112

Evans, W. S., Cavanaugh, R., Quique, Y., Boss, E., Starns, J. J., & Hula, W. D. (2021). Playing With BEARS: Balancing Effort, Accuracy, and Response Speed in a Semantic Feature Verification Anomia Treatment Game. *Journal of Speech, Language, and Hearing Research*, *64*(8), 3100–3126. https://doi.org/10.1044/2021_JSLHR-20-00543

Foygel, D., & Dell, G. S. (2000). Models of impaired lexical access in speech production. *Journal of Memory and Language*, *43*(2), 182–216. https://doi.org/10.1006/jmla.2000.2716

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Gilmore, N., Meier, E. L., Johnson, J. P., & Kiran, S. (2020). Typicality-based semantic treatment for anomia results in multiple levels of generalisation. *Neuropsychological Rehabilitation*, *30*(5), 802–828. https://doi.org/10.1080/09602011.2018.1499533

Gravier, M. L., Dickey, M. W., Hula, W. D., Evans, W. S., Owens, R. L., Winans-Mitrik, R. L., & Doyle, P. J. (2018). What Matters in Semantic Feature Analysis: Practice-Related Predictors of Treatment Response inAphasia. *American Journal of Speech-Language Pathology / American Speech-Language-Hearing Association*, *27*(1S), 438–453. https://doi.org/10.1044/2017_AJSLP-16-0196

Hart, T., Dijkers, M. P., Whyte, J., Turkstra, L. S., Zanca, J. M., Packel, A., Van Stan, J. H., Ferraro, M., & Chen, C. (2019). A theory-driven system for the specification of rehabilitation treatments. *Archives of Physical Medicine and Rehabilitation*, *100*(1), 172–180.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Kiran, S. (2008). Typicality of inanimate category exemplars in aphasia treatment: Further evidence for semantic complexity. *Journal of Speech, Language, and Hearing Research*, *51*(6), 1550–1568. https://doi.org/10.1044/1092-4388(2008/07-0038)

Kiran, S., & Roberts, P. M. (2010). Semantic feature analysis treatment in Spanish–English and French–English bilingual aphasia. *Aphasiology*, *24*(2), 231–261. https://doi.org/10.1080/02687030902958365

Kristensson, J., Behrns, I., & Saldert, C. (2014). Effects on communication from intensive treatment with semantic feature analysis in aphasia. *Aphasiology*, 1–22. https://doi.org/10.1080/02687038.2014.973359

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press.

Menahemi-Falkov, M., Breitenstein, C., Pierce, J. E., Hill, A. J., O'Halloran, R., & Rose, M. L. (2021). A systematic review of maintenance following intensive therapy programs in chronic post-stroke aphasia: Importance of individual response analysis. *Disability and Rehabilitation*, 1–16.

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*. https://doi.org/10.1044/2016_JSLHR-L-15-0303

Moreno-Martínez, F. J., & Montoro, P. R. (2012). An ecological alternative to Snodgrass & Vanderwart: 360 high quality colour images with norms for seven psycholinguistic variables. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0037527

Nalborczyk, L., Batailler, C., Loevenbruck, H., Vilain, A., & Bürkner, P. C. (2019). An introduction to bayesian multilevel models using brms: A case study of gender effects on vowel variability in standard Indonesian. *Journal of Speech, Language, and Hearing Research*, *62*(5), 1225–1242. https://doi.org/10.1044/2018_JSLHR-S-18-0006

Nickels, L., Rapp, B., & Kohnen, S. (2015). Challenges in the use of treatment to investigate cognition. *Cognitive Neuropsychology*, *32*(3–4), 91–103. https://doi.org/10.1080/02643294.2015.1056652

Peach, R. K., & Reuter, K. A. (2010). A discourse-based approach to semantic feature analysis for the treatment of aphasic word retrieval failures. *Aphasiology*, *24*(9), 971–990. https://doi.org/10.1080/02687030903058629

R Core Team. (2020). *R: A language and environment for statistical computing* (4.0.3). 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.

Sandberg, C. W., & Gray, T. (2020). Abstract Semantic Associative Network Training: A Replication and Update of an Abstract Word Retrieval Therapy Program. *American Journal of Speech-Language Pathology*, *29*(3), 1574–1595.

Swinburn, K., Porter, G., & Howard, D. (2004). *The Comprehensive Aphasia Test*. Hove: Psychology Press.

Tabrizi, R., Walton, L., Simon, E., & Silkes, J. P. (2021). Repetition priming in treatment of anomia. *American Journal of Speech-Language Pathology*, 1–19.

Turkstra, L. S., Norman, R., Whyte, J., Dijkers, M. P., & Hart, T. (2016). Knowing What We're Doing: Why Specification of Treatment Methods Is Critical for Evidence-Based Practice in Speech-Language Pathology. *American Journal of Speech-Language Pathology*, *25*(2), 164–171. https://doi.org/10.1044/2015_AJSLP-15-0060

Van Stan, J. H., Dijkers, M. P., Whyte, J., Hart, T., Turkstra, L. S., Zanca, J. M., & Chen, C. (2019). The rehabilitation treatment specification system: Implications for improvements in research design, reporting, replication, and synthesis. *Archives of Physical Medicine and Rehabilitation*, *100*(1), 146–155. https://doi.org/10.1016/j.apmr.2018.09.112

Wambaugh, J. L., & Ferguson, M. (2007). Application of semantic feature analysis to retrieval of action names in aphasia. *Journal of Rehabilitation Research & Development*, *44*(3).

Zanca, J. M., Turkstra, L. S., Chen, C., Packel, A., Ferraro, M., Hart, T., Van Stan, J. H., Whyte, J., & Dijkers, M. P. (2019). Advancing rehabilitation practice through improved specification of interventions. *Archives of Physical Medicine and Rehabilitation*, *100*(1), 164–171.