AJSLP

Research Article

What Matters in Semantic Feature Analysis: Practice-Related Predictors of Treatment Response in Aphasia

Michelle L. Gravier,[a] Michael W. Dickey,[a,b] William D. Hula,[b,c] William S. Evans,[a] Rebecca L. Owens,[c] Ronda L. Winans-Mitrik,[c] and Patrick J. Doyle[a,b,c]

Purpose: This study investigated the predictive value of practice-related variables—number of treatment trials delivered, total treatment time, average number of trials per hour, and average number of participant-generated features per trial—in response to semantic feature analysis (SFA) treatment.
Method: SFA was administered to 17 participants with chronic aphasia daily for 4 weeks. Individualized treatment and semantically related probe lists were generated from items that participants were unable to name consistently during baseline testing. Treatment was administered to each list sequentially in a multiple-baseline design. Naming accuracy for treated and untreated items was obtained at study entry, exit, and 1-month follow-up.
Results: Item-level naming accuracy was analyzed using logistic mixed-effect regression models. The average number of features generated per trial positively predicted naming accuracy for both treated and untreated items, at exit and follow-up. In contrast, total treatment time and average trials per hour did not significantly predict treatment response. The predictive effect of number of treatment trials on naming accuracy trended toward significance at exit, although this relationship held for treated items only.
Conclusions: These results suggest that the number of patient-generated features may be more strongly associated with SFA-related naming outcomes, particularly generalization and maintenance, than other practice-related variables.
Supplemental Materials: https://doi.org/10.23641/asha.5734113

Word-finding difficulties are ubiquitous in aphasia (Goodglass, 1980; Goodglass & Wingfield, 1997; Schuell, Jenkins, & Jimenez-Pabon, 1964) and can contribute to conversation breakdowns and general communication difficulty (Lesser & Algar, 1995; Perkins, Crisp, & Walshaw, 1999). As a consequence, interventions targeting naming are a common component of aphasia therapy (e.g., Wisenburn & Mahoney, 2009). As word-finding deficits are often attributed to difficulty in mapping the intended concept to the spoken word (Dell, Schwartz, Martin, Saffran, & Gagnon, 1997; Foygel & Dell, 2000), naming therapy generally addresses breakdowns at one or more stages of this word retrieval process. Semantically based treatment approaches target the lexical-semantic processing stage (linking a conceptual representation with associated lexical-semantic representations), whereas phonologically based approaches target the phonological processing stage (mapping the lexical item to the appropriate word form, Dell et al., 1997; Levelt, Roelofs, & Meyer, 1999). Evidence supports the ability of both naming therapy approaches to improve naming of treated items and to foster maintenance of the treatment gains (Nickels, 2002). However, generalization to naming of untreated items above levels typically expected based purely on repeated exposure/probing is more limited (Nickels, 2002; Wisenburn & Mahoney, 2009). Generalization is important because the treated set is typically small, and improvement on treated items alone is unlikely to have a functional impact (Thompson, 1989).

Semantic feature analysis (SFA; Boyle & Coelho, 1995; Massaro & Tompkins, 1994; Ylvisaker & Szekeres, 1985) is a semantically based treatment approach for naming deficits, which has been somewhat successful at promoting

[a]Geriatric Research Education and Clinical Center, VA Pittsburgh Healthcare System, PA
[b]University of Pittsburgh, PA
[c]Audiology and Speech Pathology Service, VA Pittsburgh Healthcare System, PA
Correspondence to Michelle L. Gravier: michelle.gravier@va.gov

Editor: Margaret Blake
Associate Editor: Anastasia Raymer
Received October 30, 2016
Revision received April 17, 2017
Accepted May 26, 2017
https://doi.org/10.1044/2017_AJSLP-16-0196
Publisher Note: This article is part of the Special Issue: Select Papers From the 46th Clinical Aphasiology Conference.

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

generalization, especially when compared with more phonologically oriented treatment approaches. In a systematic review of published SFA treatment studies, Boyle (2010) reported that across seven studies, 16 of 17 individuals improved on naming of treated items, and of those, 13 showed some evidence of generalization to untreated items. In a separate meta-analysis, Oh, Eom, Park, and Sung (2016) reported that 18 of 23 participants for whom generalization was assessed in their sample of nine SFA treatment studies showed some improvements with naming of untreated items. In contrast, many phonologically oriented naming treatments show more limited evidence of generalization. For example, only three of 10 participants in Leonard, Rochon, and Laird (2008) showed generalization to untreated items in response to phonological components analysis, a treatment with a similar structure to SFA but focused on phonological word form features. Likewise, in her systematic review of naming treatment outcomes, Nickels (2002) reports that only 11 of 44 participants who received restorative naming treatment focused on phonological or orthographic representations showed evidence of generalization to untreated items (see findings by Best et al., 2013, for further evidence that generalization in response to phonological cuing treatment is relatively limited, appearing for only three of 16 participants in their study; although, note that this critique may not extend to treatments that focus on improving sublexical phonological processes rather than on facilitating word form retrieval, such as the phonomotor treatment by Kendall et al., 2008, which show better evidence of improved naming of untreated items).

The observed treatment and generalization effects resulting from SFA are hypothesized to originate from the treatment approach's focus on strengthening connections between semantically related concepts in the lexicon. According to automatic spreading activation models of semantic representation (e.g., Collins & Loftus, 1975), the activation of a concept automatically spreads to all semantically related items in the lexicon, depending on their degree of relatedness. For instance, activating the concept *lemon* would strongly activate *fruit, yellow,* and *sour,* and less so *pie* and *tree*. Furthermore, as these connections are claimed to be bidirectional and summative, activating *fruit* and *yellow* would strongly activate *lemon* and *banana,* whereas activating *fruit, yellow,* and *sour* would solely strongly activate *lemon* (Hutchison, 2003).

However, lexical-semantic processing is hypothesized to be impaired in some individuals with aphasia, such that either the lexical entries are underspecified (Butterworth, Howard, & McLoughlin, 1984; Nickels & Howard, 1994) or the process of automatic spreading activation is altered (Ferrill, Love, Walenski, & Shapiro, 2012; Silkes & Rogers, 2012). In either case, this impairment results in the activation of an insufficient number of distinguishing features to permit selection of the correct word form during phonological processing. For example, if an individual with aphasia had an underspecified lexical entry for *lemon* such that he or she was only able to activate the associated feature *fruit* during lexical-semantic processing, he or she might be just as likely to activate the phonological form for other items sharing that same feature, such as *banana* or *lime*. The resulting phenomenon of producing an unintended semantically related word in place of the target word, known as a semantic paraphasia, is commonly observed in individuals with aphasia (e.g., Foygel & Dell, 2000; Howard & Orchard-Lisle, 1984; Nickels, 2002).

In SFA, participants are provided with an item, generally for which naming has been established to be particularly problematic for that individual, and are guided by a clinician to generate item-specific features in different functional categories. In order to remediate the aforementioned underspecification, for instance, the participant might be guided within the context of SFA to generate a distinguishing feature for the category "Group" such as *citrus* and/or the features *yellow* and *sour* for the category "Description." These additional features would serve to distinguish the item from other yellow fruits, such as *banana,* or sour citrus fruits, such as *lime*. Thus, across multiple episodes of structured practice, SFA is intended both to strengthen the connections between related semantic concepts, encouraging automatic spreading activation, and to increase specification in the lexical-semantic system. Both of these changes should result in facilitated word retrieval. In addition, items that are semantically related to the treatment items are also indirectly treated via stimulation of shared semantic features, resulting in the observed generalization of treatment effects.

Although SFA treatment studies have generally reported positive outcomes, not all individuals benefit, and those who do benefit respond to varying degrees. One factor that likely contributes to variability in participant response is the wide variability in treatment intensity across SFA studies, as treatment intensity has been positively related to outcomes (Basso & Macis, 2011; Hinckley & Craig, 1998; Robey, 1998). However, studies investigating treatment intensity have often conflated the total amount of treatment and the amount of treatment per unit of time (Bhogal, Teasall, & Speechley, 2003; Cherney, Patterson, & Raymer, 2011). In more recent studies, it has been suggested that treatment amount may be best characterized by considering not only the total amount of time spent in treatment but also dose (number of treatment episodes or trials), dose frequency (the times that a dose or trial is provided in a given amount of time), and dose form (the task involved in a treatment episode; Baker, 2012; Cherney, 2012; Warren, Fey, & Yoder, 2007). Baker (2012) has suggested that within dose form, it is particularly important to consider all "active ingredients," including both therapeutic inputs (e.g., clinician questions/recasts) *and* client acts (e.g., quantity and quality of practice of a specific skill).

Characterizing the contribution of dose form may be especially salient for SFA. In SFA, the client act of generating features is the hypothesized source of the strengthened semantic network. Although clinicians initially guide participants through the feature chart, the goal is to have the clinician taper the amount of support over the duration of treatment until the participant is more or less

independent in generating relevant semantic features. The importance of dose form within SFA is supported by the finding that studies that have required participants to *generate* semantic feature exemplars have generally garnered better outcomes compared with those studies that have simply asked participants to *select* or *confirm* clinician-generated features (Boyle, 2010). Among the SFA studies Boyle (2010) reviews, the study with the clearest evidence of generalization independent of repeated exposure (Lowell, Beeson, & Holland, 1995) required feature generation. This finding suggests that although repeated exposure to paired features and lexical items may facilitate mapping between the lexical item and the word form resulting in improved naming of treated items, feature generation more likely strengthens connections between lexical items and their associated features, leading to improved naming of untreated items that share semantic features.

Aspects of these practice-related determinants of intensity, including dose, dose frequency, and treatment amount, are commonly but not consistently reported in published studies of SFA. For instance, in a meta-analysis, Quique, Evans, and Dickey (2017) found that 10 of the 14 studies reviewed reported approximate treatment session durations and total number of treatment sessions, permitting estimation of total treatment time by treatment list and participant. In addition, dose (the total number of trials completed—typically one trial per item per session) was reported in all of the reviewed studies, permitting averaged dose frequency (number of trials per unit of time) to be estimated. However, importantly, although the contribution of total treatment time, dose, and dose frequency on naming accuracy can be estimated for these studies given this information, it is provided only by list, not for each individual treated item. Averaging within lists obscures the contribution of each practice-related factor to improvements in naming accuracy. For instance, even in the absence of significant listwise naming improvements, it may be the case that a participant improved on some items for which they spent more time, for instance, but did not improve on those items within the list on which they spent less time.

In contrast to information on dose and dose frequency, quantitative information regarding dose form (the number of semantic features generated) is rarely, if ever, reported. Of the 14 studies reviewed by Quique et al. (2017), only one specified whether generation of one feature at a minimum was even required in each category. For instance, Boyle (2004) reported that in their study, "not every feature [category] was appropriate for a treatment stimulus" and "only those features deemed appropriate for a stimulus item were elicited" (p. 240). This statement implies that for some items, generation of features within certain categories was optional or at the discretion of the clinician, potentially resulting in many more opportunities for feature generation for some items compared with others. Without information on the number of opportunities for feature generation, or the number of features that were required for each category, it is impossible to determine how many features were generated per trial, item, or participant.

Furthermore, the number of features that are generated *by the participant* versus those that are generated by the clinician cannot be extrapolated even in cases where the number of trials or opportunities for feature generation is provided. For example, in an attempt to ensure that the features that were practiced during SFA trials were maximally shared by the treated and semantically related untreated items, and hence encourage generalization, Wallace and Kimelman (2013) limited feature generation responses to only those features (one per category) that they had predetermined on the basis of a semantic overlap between sets. In this study, each item was treated once per day for 12 sessions, allowing for a calculation of the total treatment dose (1 trial per item per session × 12 sessions = 12 trials per item) and the number of opportunities for feature generation (12 trials per item × 6 categories × 1 feature per category = 72 potential features generated). Although all three of the individuals with aphasia who participated in the study showed improved naming of treated items, only two of the three participants exhibited improvements on untreated items. Generalization did not appear to be related to aphasia severity: P1 (the participant who did not generalize) and P3 (one of the participants who did generalize) had almost identical Western Aphasia Battery scores (Kertesz, 2006). Because the number of trials and the total number of features generated for each item were held constant, one likely source of the response variability is the proportion of those features that were generated by the participant versus those that were supplied by the clinician over the course of treatment. If the act of providing features increases the strength and specificity of the lexical-semantic representation for treated items and facilitates lexical retrieval (Boyle, 2010), the number of features that a *participant* generates should predict treatment-related naming gains and generalization.

To inform best clinical practice and allow clinicians to prioritize treatment goals, it is important for research to address which aspects of treatment (total time in treatment, number of trials: dose), trials per hour (dose frequency), and/or number of client acts (features generated: dose form) are more likely to contribute not only to gains on treated items but also to the generalization of untreated items and maintenance of gains. As noted above, each of these practice-related dimensions may vary across individuals, items, or both, and each could plausibly contribute to treatment-related naming gains. This article presents an analysis of preliminary data for 17 participants from an ongoing clinical trial investigating predictors of treatment response to intensive SFA. It examines the effect of each of these practice-related factors on item-level naming performance. It is hypothesized that the number of trials completed for a particular item will positively predict naming improvement and maintenance for treated items, but not generalization, given that the number of trials is directly related to the number of naming opportunities for treated items. The number of participant-generated features, on the other hand, is hypothesized to predict response to treated and untreated items and maintenance of those

generalization effects, given that it is directly related to the strengthening of the semantic system hypothesized to underlie the efficacy of SFA. Given that the total amount of treatment as well as the number of hours of treatment per week have been found to be positively correlated with treatment response at the subject level (Bhogal et al., 2003; Robey, 1998), it is hypothesized that the total number of hours that an item is treated and the average number of trials per hour will both positively predict item-level treatment response. However, whether this effect will apply to both treated and untreated items is unclear.

Method

Participants

Nineteen adults with chronic aphasia greater than 6 months post-onset due to unilateral left-hemisphere stroke qualified for participation. Participants for this study were recruited from the Western Pennsylvania Research Registry, the Audiology and Speech Pathology Research Registry maintained by the VA Pittsburgh Healthcare System (VAPHS), VAPHS speech-pathology clinician referral, and the VA Pittsburgh's Program for Intensive Residential Aphasia Treatment and Education (PIRATE). Participants who were recruited from PIRATE were given the option to participate in this research study as an alternative prior to the initiation of treatment. As the screening assessment procedures differ somewhat between this study and PIRATE, participants recruited from the latter may have received additional speech-language and/or cognitive assessments during baseline testing. However, all participants enrolled in this study exclusively received SFA treatment (described in more detail in the Treatment section below) and did not receive any concurrent speech-language treatment outside of the study-related sessions for the duration of the study.

All participants obtained a modality mean T-score of 70 or below on the Comprehensive Aphasia Test (CAT; Swinburn, Porter, & Howard, 2004). The CAT modality mean T-score, the average standardized score across different levels and modalities of language performance, reflects overall aphasia severity. A T-score of 50 reflects mean performance for the CAT normative sample of individuals with aphasia, with a standard deviation of 10. Thus, a score of < 70 ensures that participants were within the impaired range (within 2 *SD*s of the mean). In addition, all participants obtained a T-score of 40 or above (1 *SD* below the mean) on the auditory comprehension and naming subtests of the CAT. These criteria were used to avoid enrolling participants who have profound naming and/or comprehension impairments and are therefore unlikely to benefit from the study treatment (e.g., patients with global aphasia and/or verbal output limited to recurrent stereotypy). Exclusion criteria included a history of progressive neurological disease or nervous system injury or disorder prior to the stroke and the presence of a severe motor speech disorder. Two participants voluntarily withdrew from treatment after initiation and were therefore excluded from analysis. One participant withdrew after 23 sessions due to reduced stamina resulting in an inability to tolerate the intensive treatment schedule, and another withdrew after 10 treatment sessions due to emotional distress related to self-perception of poor performance during treatment. Demographic information for the remaining 17 participants who completed the study is provided in Table 1. Additional language and cognitive testing was administered for the purposes of the ongoing clinical trial; see Supplemental Material S1 for a summary of performance on the language measures.

Stimuli

For participants S1–S4, four treatment lists were generated, each with 10 items from three semantic categories. To assess generalization, a list of 10 semantically related items and a list of 10 semantically unrelated items were generated for each treatment list. The items on the generalization lists were not directly treated but were assessed during naming probes, as described in the Naming Probes section below. During the course of treatment for these first four participants, it became apparent that the time burden of administering the daily naming probes (ranging from approximately 10–15 min for treated list probes consisting of 30 items to 30 min to 1 hr for all list probes consisting of 120 items) was limiting the amount of time available for treatment. The following changes were subsequently instituted for participants S5–S17: The number of items per list was reduced from 10 to five, three treatment lists were generated rather than four, and the list of semantically unrelated items was eliminated. Following these changes, probes for the treated list consisted of 10 items, and probes for all three lists consisted of 30 items. Due to this discrepancy between earlier and later participants, items from the semantically unrelated lists were not included in the analyses.

Items included in a given participant's treatment lists were determined on the basis of performance on a picture-naming task. The naming task consisted of 194 items across eight semantic categories: fruits and vegetables, transportation, four-legged animals, sports equipment, tools, musical instruments, occupations, and birds. The pictures were full-color photographs generated from Google images and were selected to have minimally competing backgrounds. To avoid effects of repeated exposure, items included on the naming task were constrained such that they did not occur in any of the other assessments. Pictures were presented one at a time, and participants were given 20 s to respond. Clinicians provided prompts for increased response specificity (e.g., "seagull" rather than "bird"), but no accuracy feedback was given. Self-corrections, sound distortions, and off-target responses due to dialectical differences or preexisting speech deficits (e.g., developmental articulation disorder such as a lisp) were scored as correct, whereas paraphasias or no-responses were scored as incorrect. Shortened names ("hippo" for "hippopotamus") and

Table 1. Participant demographics.

Subject	Age (years)	Race	Gender	Education (years)	Etiology	MPO	CAT mean modality T-score
S1	42	AA, NA	M	15	ischemic LH CVA w/ HC	55	53.7
S2	62	C	M	23	ischemic LH CVA	89	50.0
S3	51	C	M	13	ischemic LH CVA w/ HC	75	46.3
S4	68	C	F	12	ischemic LH CVA	199	47.5
S5	66	C	M	18	ischemic LH CVA	86	64.2
S6	52	C	M	12	ischemic LH CVA	10	56.0
S7	24	C	M	16	LH cerebral aneurysm	10	58.0
S8	78	H	M	25	ischemic LH CVA	16	52.5
S9	64	C	M	11	ischemic LH CVA w/ HC	84	47.5
S10	45	AA	M	12	ischemic LH CVA	120	49.8
S11	72	C	M	14	ischemic LH CVA	93	50.8
S12	51	AA	M	16	ischemic LH CVA	39	51.3
S13	75	C	M	13	ischemic LH CVA	14	47.8
S14	70	C	M	14	ischemic LH CVA	172	50.2
S15	48	C	M	14	ischemic LH CVA	8	50.2
S16	74	C	M	20	ischemic LH CVA	15	52.8
S17	67	C	F	12	ischemic LH CVA w/ HC	8	59.0
AVG (*SD*)	59 (15)		2F, 15M	15 (4)		64 (59)	52.2 (4.7)

Note. *SD* = standard deviation; AA = African American; NA = Native American; C = Caucasian; H = Hispanic; M = male; F = female; LH = left hemisphere; HC = hemorrhagic conversion; MPO = months post-onset; CAT = Comprehensive Aphasia Test.

exact synonyms ("swimsuit" for "bathing suit") were also considered acceptable correct responses. If a participant indicated that they could not name an item because they had never been exposed to it before, it was marked as incorrect but not selected as a treatment target.

The naming task was administered three times, twice in full. The third administration was a shortened version that consisted of items that the participant named correctly on only *one* of the first two administrations. To qualify for treatment, an item had to be named incorrectly twice. Thus, those items were named incorrectly on both initial administrations *or* on one of the initial administrations, and the third administration qualified for use in treatment (see Supplemental Material S2 for the full list of items on the picture-naming task and a sample naming task administration procedure/scoring). Treatment lists were generated from categories with a sufficient number of qualifying items. In the case that more than the minimum number of items or categories qualified for treatment, selection was based upon patient preference (see Supplemental Material S3 for a sample of list selection on the basis of eligible items by category). Treatment target items and untreated generalization items were balanced within treatment lists for a number of syllables.[1]

Treatment

Participants received individual SFA treatment (Boyle & Coelho, 1995; Coelho, McHugh, & Boyle, 2000) 4–5 days per week for 4 weeks, in two daily sessions of approximately 120 min each. Treatment was administered using a computer program (Winans-Mitrik et al., 2013) that permits the generation of individualized treatment lists, randomized item presentation in sets, and tracks performance including participant naming accuracy and responses (as input by the clinician) and time-keeping metrics.

For each treatment trial, participants were first shown a picture of the target and asked to name it aloud. Participants were given 20 s to respond, and accuracy was recorded. Accuracy criteria were identical to those used for the baseline naming task, described previously. Regardless of accuracy, participants were then asked to generate semantic features for the target in five categories: superordinate category ("group"), physical properties ("description"), use/action ("function"), location ("context"), and association ("personal association"). Treatment proceeded sequentially through these categories for each trial. If the participant correctly produced the name of the item at any point during the trial, it was typed by the clinician into the "free text" box. For an example of the feature chart used during SFA treatment, see Figure 1. The decision was made for this study to combine the "use" and "action" categories used by Boyle and Coelho (1995), as a majority of responses within these two categories tend to overlap.

For each category, the clinician began by providing a general prompt. For example, for the *description* category, the clinician might ask, "How would you describe this?" As features were verbally provided by the participant, they were typed by the clinician into the appropriate category box. Participants were encouraged to provide up to three features in each category, with the exception of "group" and "association" for which one feature was

[1]Lexical frequency was higher for treated compared with untreated items. However, upon evaluation, lexical frequency did not improve the model fit and, hence, was not included as a factor in reported analyses.

Figure 1. Sample semantic feature analysis (SFA) trial presented via a computer with associated program output. The SFA chart (top) is visible for the duration of the trial, whereas the trial information summary (bottom) is generated as a text document and output after each session. (PG) indicates that a feature was "patient generated." The star next to the sentence indicates that the participant was able to generate a correct sentence (minimally containing the subject and verb and two semantically related elements) without clinician assistance. Note that the item name (in the "free text" box) is available during the sentence generation task.


```
Rollerblade
Naming: Incorrect
Group: Transportation (PG)
Description: 4 wheels (PG), Brake (PG), Buckle
Function: Skating (PG), Lace them up (PG), Buy
Context: Skate Park (PG), Park (PG), Dick's (Sporting Goods) (PG)
Other/Personal: sore bum (PG)
Free Text: rollerblade

I laced up the rollerblades *
```

required, although more were accepted if offered. If the participant was unable to verbally provide a feature within 10 s, a more directed question was asked such as "What does this item feel like?" If the participant was still unable to provide a verbal response within 10 s or responded incorrectly, a binary forced choice was provided between two maximally dissimilar features, for example, "Is this item smooth or rough?" If the participant was unable to respond to the binary forced choice, or responded incorrectly, a feature would be provided by the clinician. To encourage faster progression through trials, if a participant was unable to generate a third feature for a category within a 5-s response window, the clinician would immediately provide a feature, omitting the cuing process. If a participant repeatedly generated the same features on consecutive trials, clinicians would accept that feature as correct, but encourage generation of a new feature. If a participant generated a "weak" feature (one that was only very indirectly related to the target item), it was not written on the chart, and the participant was prompted to provide an alternate response. Features were only scored as participant-generated if they were provided verbally in either of the first two cuing levels (with a general or a specific clinician-generated prompt).

At the end of each trial, participants were asked once again to name the target item. If an inaccurate response or no response was provided, the clinician first provided verbal feedback and then modeled the production, asking for the participant to repeat. As per methods of Wambaugh, Mauszycki, Cameron, Wright, and Nessler (2013), if the participant was unable to repeat the target, the clinician would prompt the participant to "watch me, listen to me, say what I say" to elicit a simultaneous production. Then, the clinician guided the participant in a review of the most salient feature from each of the five categories, pointing to the written word on the chart and reading it aloud. Following review, the clinician again requested that the participant name the item. If the participant was still unable to provide an accurate response, the clinician named the item and typed it into the chart.

After completing each trial, participants were asked to verbally generate a sentence using the target word. Although sentence generation is not a component of SFA as it is traditionally implemented, the sentence generation task was added to the protocol in order to increase the likelihood of generalization of the correct use of the treated items outside the context of confrontation naming. During sentence generation, the clinician provided up to two cues to assist with generation or modification of the response, for example, by directing the participant to the features written in the chart or by asking a "wh-" question to help the participant identify missing sentence elements if necessary. To be counted as correct, sentences had to contain minimally the target and a verb and have two semantically related elements. For example, if, for the target word *jeans,* a participant generated the sentence "I like jeans," they would have been cued by the clinician to generate a more closely semantically related verb by being directed to the feature chart category "function." However, the sentence "I *wear jeans*" or "*Jeans* are *blue*" would have been acceptable. If the participant was unable to verbally generate a complete sentence within a 10-s response window following prompts or generated an incorrect sentence, the clinician would provide a sentence, type it into the chart, and ask the participant to repeat it.

Stimuli were presented in random order within each list but were not repeated until the entire list had been presented. Only one list was targeted during each treatment session.

Treatment Fidelity

Thirty-minute daily treatment session samples were video-recorded to allow for assessment of treatment fidelity. A member of the study staff who was not providing treatment randomly selected 15 min for review and adherence to the SFA protocol using a Treatment Fidelity Checklist (see Supplemental Material S4). Any elements of the SFA protocol that were not checked as having been performed on the Checklist were identified as a deviation. All deviations were reviewed and clarified with the treating clinician to ensure that adherence improved in subsequent sessions. When session monitoring detected < 1 deviation across three consecutive samples, sessions were monitored weekly for the remainder of the 4-week treatment period. If subsequent weekly monitoring revealed > 1 deviation from the SFA protocol, daily monitoring was reinstituted until the criterion for weekly monitoring was once again reached.

Naming Probes

Lists being treated were probed daily, and all lists were probed every fourth day to assess naming maintenance and generalization. All probes were administered at the beginning of the morning session, prior to the initiation of treatment for the day. Administration and scoring of the naming probes followed the same procedure as the naming task, described in the Stimuli section above. It is notable that clinicians provided prompts for increased response specificity (e.g., "seagull" rather than "bird"), but no accuracy feedback was given.

Naming probes were audio-recorded and scored both online by the clinician and off-line by a second rater who was blinded to the order of probe administration. Interrater reliability (IRR) was calculated on the basis of percent agreement in the scoring of naming accuracy between the clinician and the second rater. If IRR < 90%, a third rater would also score the probe, and final accuracy would be determined on the basis of majority consensus. Otherwise, if IRR > 90%, the clinician scoring was used.

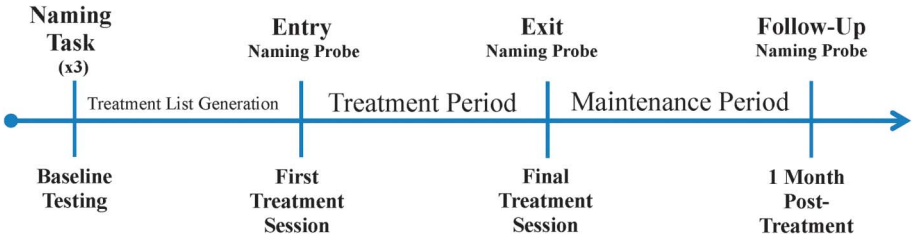
Experimental Design

Treatment was administered to each list sequentially in a multiple-baseline design. Treatment of a list was discontinued when participants named 90% (S1–S4) or 80% (S5–S17) of treated items accurately on three of four consecutive probes, or if the list was trained for a maximum of 8 days. To ensure sufficient exposure across lists, each list was also treated for a minimum of 4 days regardless of treatment probe accuracy. If a list met the criteria for discontinuation during a morning probe, the probe for the subsequent list was administered, and treatment on that list was initiated (for sample treatment list progression, see Supplemental Material S5).

Analysis

Naming probe data used for analysis were obtained at three time points: entry (immediately prior to the first treatment session), exit (the morning after the final treatment session), and follow-up (1 month after the final treatment session; see Figure 2 for the study timeline). See Tables 2 and 3 for a summary of the naming probe data by participant. Seven participants (S2–S4, S9–11, and S13) failed to reach the criterion to advance to list 3 during treatment, and therefore, naming probe data for items in that untreated list including matched semantically related items were excluded from analysis. All remaining item-level data were analyzed using multilevel generalized linear regression with a logistic link function in R 3.2.2 (R Core Team, 2015) using the lme4 package (Bates, Maechler, Bolker, & Walker, 2015). This modeling approach has several advantages over the more traditional repeated-measures techniques such as analysis of variance, including more appropriate handling of categorical response data via the logistic distribution (Jaeger, 2008) and the ability to handle unbalanced designs with varying numbers of observations per participant or condition (Baayen, Davidson, & Bates, 2008). Furthermore, multilevel models (also referred to as mixed-effect models) make it possible to account for variation in both participants and stimuli simultaneously, by incorporating crossed random effects structures, which improve the ability to make accurate inferences about the populations and effects of interest (Baayen et al., 2008). Using this approach, separate multilevel models were run for each practice-related fixed effects of interest (total number of hours of treatment,

Figure 2. Study timeline.



total number of trials, average number of trials per hour, and average number of patient-generated features per trial). In each case, this practice-related effect was tested in a three-way interaction crossed with the fixed effects of probe time (entry/exit/follow-up) and item type (treated/untreated). These models also include aphasia severity (CAT modality mean T-score) as a main effect covariate, as aphasia severity has been shown to affect treatment response (Conroy, Sage, & Lambon Ralph, 2009; Robey, 1998), and primary hypotheses were focused specifically on practice-related effects. Therefore, each of these models tested how a given practice-related factor moderated the effects of treatment over time, controlling for aphasia severity. In terms of random effects structures, all models crossed random effects intercepts for subjects and items.[2]

As it was expected that the slope of the predicted effect would differ between entry to exit and exit to follow-up, each model only included one pair of probe times as a fixed effects contrast.[3] These consist of a total of eight models overall (4 practice-related factors × 2 probe time contrasts; see Table 4 for a summary of the fixed effects structures evaluated in each model). All results were corrected for multiple comparisons using false discovery rate correction (Benjamini & Hochberg, 1995). All variables in these analyses were centered with the exception of probe time and item type. Probe time was reference-coded for "exit" in all entry-to-exit models and "follow-up" in all exit-to-follow-up models. Item type was reference-coded for "treated" items in all analyses.[4] Model fixed effects were plotted and interpreted using 95% confidence intervals estimated via bootstrapping using the "bootMer" function (Bates, Maechler, Bolker, & Walker, 2013), following the methods of Hertzog (2015). This approach simulates predicted values for a representative "average" participant by setting the combined random effects to 0.

Results

Entry to Exit (Models 1–4)

The main effects of probe time and item type were significant for all entry-to-exit models ($p < .001$), indicating a significant change in response accuracy from entry to exit across items and participants, with treated items named more accurately than untreated items across time points. The Probe Time × Probe Type interaction (treated vs. untreated) was also significant ($p < .01$) in all models, indicating that naming of treated items *improved* more than naming of untreated items during the treatment phase, controlling for aphasia severity. The main effect of aphasia severity was significant on its own in all entry-to-exit models ($p < .05$) with the exception of the model including the total number of trials, in which it approached significance ($\beta = .08$, $SE = 0.04$, $p = .06$), indicating that overall language performance also predicted better naming performance at exit.

Given this significant main effect of aphasia severity, a secondary analysis also looked at the two-way Aphasia Severity × Probe Time interaction as well as the three-way Severity × Probe Time × Probe Type interaction in the absence of practice-related factors. However, neither of these interactions were significant ($p > 0.2$), indicating that aphasia severity did not directly affect treatment response for either treated or untreated items in this study.

Total Hours of Treatment (Model 1, Figure 3A)

The main effect of total hours of treatment was not significant ($\beta = -.03$, $SE = 0.12$, $p = .78$), indicating that more time spent on an item, in and of itself, did not lead to improved naming accuracy for that item. The two-way Hours of Treatment × Probe Time ($\beta = .20$, $SE = 0.16$, $p = .23$) and Hours of Treatment × Item Type interactions ($\beta = .26$, $SE = 0.16$, $p = .13$) as well as the three-way Hours of Treatment × Probe Time × Item Type interaction ($\beta = .42$, $SE = 0.22$, $p = .09$) failed to reach significance,

[2]Although recommendations exist (Barr, Levy, Scheepers, & Tily, 2013) to include nested random slopes for fixed effects of interest (i.e., "maximal" random effects structures), the inclusion of random slopes in the presented models did not improve the model fit as assessed by Akaike information criterion and, in some cases, led to model convergence errors.

[3]Models including all three time points with an additional quadratic term to account for the change in slope were also evaluated, but this approach did not change the significance of any of the results. Therefore, the results of the two time-point models are reported here for clarity.

[4]Reference coding allows for the evaluation of a model at a particular level of the reference-coded variable. Thus, reference-coding the probe time variable for "exit," for example, allows for a direct assessment of whether the factors of interest predicted probe-naming accuracy at that point in time (treatment exit), in contrast to centering, which would result in evaluating the effect halfway between treatment entry and exit.

Table 2. Performance variables (summarized across all treated items).

Subject	Sessions	Trials	Minutes	Trials/Hour	Features[a]	Features[a]/Trial
S1	36	694	3884	10.7	4987	7.2
S2	37	411	2933	8.4	3388	8.2
S3	30	369	2943	7.5	1625	4.4
S4	28	274	2299	7.2	1656	6.0
S5	30	344	3330	6.2	2274	6.6
S6	32	347	4011	5.2	1985	5.7
S7	30	541	3698	8.8	4608	8.5
S8	32	522	3767	8.3	3361	6.4
S9	28	258	2927	5.3	834	3.2
S10	28	443	2990	8.9	4365	9.9
S11	29	320	2915	6.6	2346	7.3
S12	29	333	2837	7.0	2376	7.1
S13	28	511	3524	8.7	2851	5.1
S14	28	703	3045	13.9	6190	8.8
S15	26	670	3045	13.2	6609	9.9
S16	31	854	3585	14.3	8531	10.0
S17	27	446	3014	8.9	3104	7.0

[a]Participant-generated features.

suggesting that the total amount of time an item was treated did not significantly predict naming improvement from entry to exit for either treated or untreated items.

Total Number of Trials (Model 2, Figure 3B)

The main effect of the total number of trials was not significant (β = .33, *SE* = 0.13, *p* = .06), neither was the two-way Total Number of Trials × Probe Time interaction (β = .30, *SE* = 0.14, *p* = .44) or the two-way Total Number of Trials × Item Type interaction (β = .16, *SE* = 0.15, *p* = .44). The three-way Total Number of Trials × Probe Time × Item Type interaction also failed to reach significance (β = .19, *SE* = 0.20, *p* = .44).

Average Number of Trials per Hour (Model 3, Figure 3C)

The main effect of trials per hour was significant (β = .20, *SE* = 0.07, *p* < .01), although the two-way Trials per Hour × Probe Time interaction failed to reach significance (β = .13, *SE* = 0.08, *p* = .16). This suggests that the effect of trials per hour was not a *treatment-related* effect, but rather that the participants who were more likely to complete more trials per hour were also more likely to name items correctly at both study entry and exit. Likewise, items for which participants were more likely to complete more trials were also more likely to be named correctly at both entry and exit. The two-way Average Trials per Hour × Item Type interaction (β = .02, *SE* = 0.08, *p* = .84) as well as the three-way Average Trials per Hour × Item Type × Probe Time interaction (β = −.04, *SE* = 0.11, *p* = .81) were not significant.

Average Number of Participant-Generated Features per Trial (Model 4, Figure 3D)

The main effect of the average number of features generated per trial (β = .39, *SE* = 0.12, *p* = .014) was significant, as was the two-way Average Number of Features Generated per Trial × Probe Time interaction (β = .43, *SE* = 0.13, *p* = .015), indicating that, overall, generating more features for an item predicted an increase in naming accuracy from entry to exit. The two-way Average Number of Features Generated per Trial × Item Type interaction was not significant (β = .06, *SE* = 0.14, *p* = .64), nor was the three-way Average Number of Features Generated per Trial × Probe Time × Item Type interaction (β = .09, *SE* = 0.18, *p* = .64), indicating that the effect of generating more features was not significantly different for treated compared with untreated items.

Exit to Follow-up (Models 5–8)

Neither the main effects of probe time nor the two-way Probe Time × Item Type interaction were significant in any of the models (*p* > .05), indicating that naming performance did not decline significantly from exit to follow-up for either treated or untreated items. However, the main effect of item type remained significant (*p* < .001), indicating that naming accuracy for treated items remained more accurate overall than for untreated items. The main effect of aphasia severity was not significant in any models (*p* > .05).

Models from exit to follow-up including factors of total hours of treatment (Model 5, Figure 3E), total number of trials (Model 6, Figure 3F), and average trials per hour (Model 7, Figure 3G) all failed to find significant main effects or interactions (*p* > .2 for all effects in all models), indicating that these factors did not predict maintenance of treatment effects for either treated or untreated items.

The main effect of the average number of features generated per trial (β = .254, *SE* = 0.11, *p* = .02; Model 8, Figure 3H) was significant; although, the two-way Item Type × Average Number of Features Generated per Trial interaction (β = −.01, *SE* = 0.13, *p* = .94) failed to reach significance. This suggests that generating more features positively predicted maintenance for both treated and untreated items.

Table 3. Naming probe performance.

Subject	Item type	Entry	Exit	Follow-up
S1	Treated	8/30	27/30	28/30
	Untreated	9/30	16/30	15/30
S2	Treated	3/20	12/20	10/20
	Untreated	9/20	10/20	7/20
S3	Treated	5/20	5/20	7/20
	Untreated	5/20	5/20	4/20
S4	Treated	4/20	4/20	1/20
	Untreated	1/20	7/20	1/20
S5	Treated	8/15	11/15	12/15
	Untreated	7/15	4/15	5/15
S6	Treated	6/15	13/15	9/15
	Untreated	4/15	6/15	5/15
S7	Treated	6/15	15/15	15/15
	Untreated	3/15	9/15	6/15
S8	Treated	2/15	11/15	7/15
	Untreated	3/15	4/15	3/15
S9	Treated	1/10	2/10	4/10
	Untreated	1/10	0/10	1/10
S10	Treated	2/10	8/10	6/10
	Untreated	1/10	2/10	3/10
S11	Treated	2/10	4/10	5/10
	Untreated	4/10	4/10	7/10
S12	Treated	4/15	13/15	13/15
	Untreated	2/15	7/15	6/15
S13	Treated	3/10	5/10	5/10
	Untreated	4/10	4/10	3/10
S14	Treated	2/15	9/15	7/15
	Untreated	2/15	5/15	0/15
S15	Treated	3/15	12/15	13/15
	Untreated	5/15	8/15	8/15
S16[a]	Treated	7/15	14/15	8/15
	Untreated	8/15	13/15	10/15
S17[a]	Treated	6/15	7/15	8/15
	Untreated	5/13	3/13	4/13

[a]S16 and S17 did not have a sufficient number of items qualify for treatment, and a modified qualification criterion was adopted wherein items that were named incorrectly on either of the first two naming test administrations were considered candidate treatment items regardless of whether they were named correctly on the third naming test administration. S16 had two of 30 items (one treated and one untreated across one list), and S17 had nine of 30 items (four treated and five untreated across three lists) qualify via this criterion.

Discussion

This study investigated which practice-related aspects of SFA treatment (dose: total time in treatment, number of trials; dose frequency: number of trials per hour; and/or dose form: number of client acts/participant-generated features) predicted naming improvements from study entry to exit for both treated words and semantically related untreated words. It also examined the maintenance of those improvements at follow-up. On the basis of the results of previously published studies, it was hypothesized that the number of training trials that were performed for an item would predict the size of direct training effects for that item (Nickels, 2002; Wisenburn & Mahoney, 2009), whereas the average number of features generated per trial would predict generalization to the naming of untreated items (Boyle, 2010; Lowell et al., 1995). It was additionally hypothesized that the total hours of treatment and trials per hour would positively predict treatment response; although, it was unclear on the basis of previous findings whether that effect would apply to both treated and untreated items.

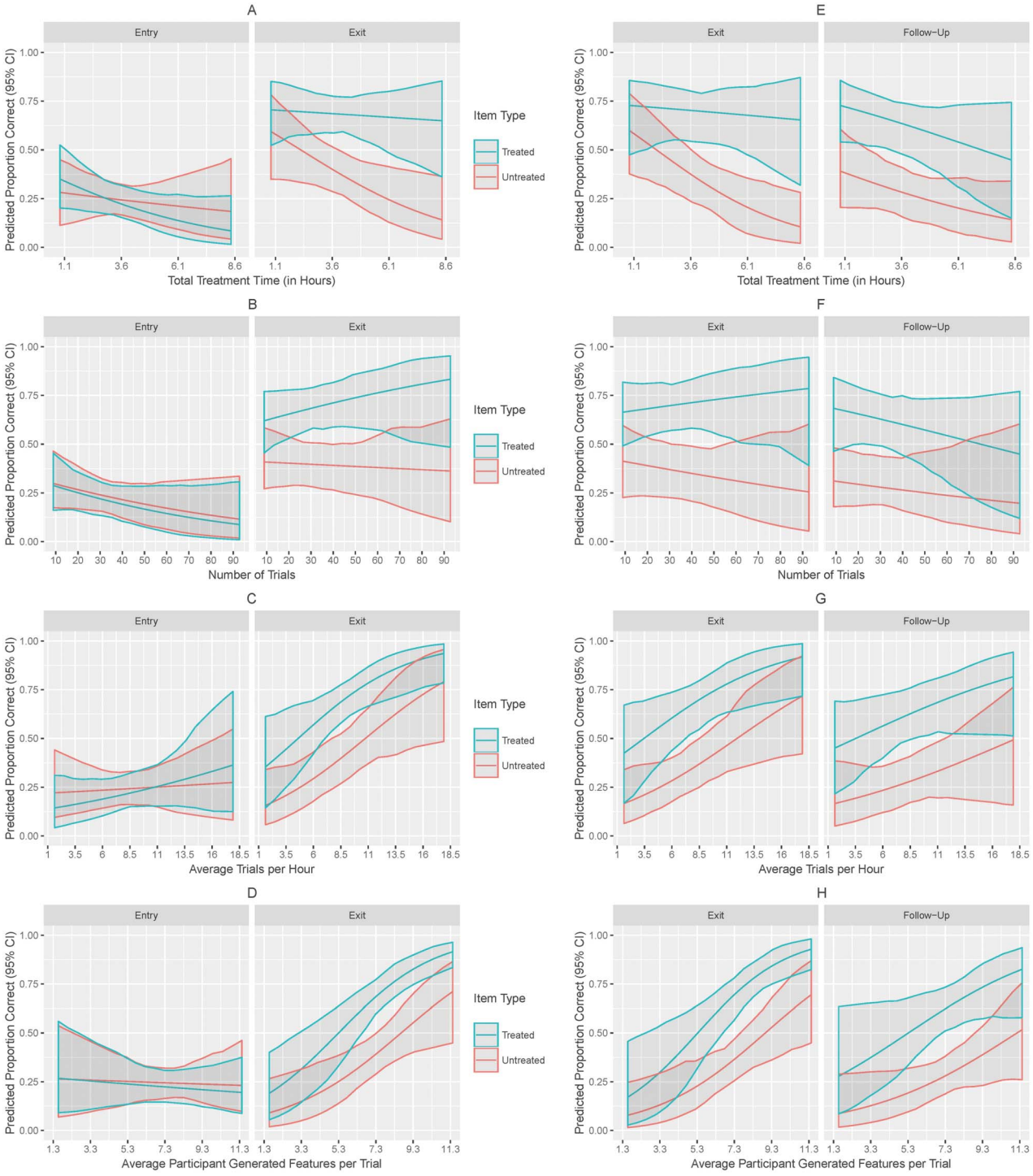
It was surprising that neither the total number of trials completed for an item, the average number of trials per hour, nor the total time spent treating an item significantly predicted treatment response for treated items or generalization to untreated items. In contrast and consistent with predictions, the average number of participant-generated features for an item predicted improvement on both treated items and generalization to naming of semantically related untreated items. The number of participant-generated features also predicted maintenance of improvements for both item types. In particular, generating just one additional feature per trial, on average, increased the odds of a correct naming response at exit by 1.47. This corresponds to an 8% increase in predicted naming accuracy at exit (an increase from 68% to 76% for treated items and from 39% to 47% for untreated items) and more than 5% at follow-up (an increase from 62% to 68% for treated items and from 28% to 33% for untreated items) at the average value of features generated and aphasia severity. Thus, participant-generated features (dose form) appear to be an especially robust practice-related predictor of treatment response for SFA.

Table 4. Fixed effects structure of evaluated models.

Model no.	Practice-related fixed effects	×	Probe times	×	Probe type	+	Fixed effects covariate
1	Total no. of trials						
2	Total hours of treatment						
3	Average no. of trials per hour		Entry/Exit				
4	Average no. of patient-generated features per trial						
5	Total no. of trials				Treated/Untreated		Comprehensive Aphasia Test modality mean T-score
6	Total hours of treatment						
7	Average no. of trials per hour		Exit/Follow-up				
8	Average no. of patient-generated features per trial						

Note. All models additionally included the crossed random effects intercepts for subjects and items.

Figure 3. Model results. Shaded regions in all graphs represent 95% confidence intervals (CI). (A) Model 1: Predicted proportion of correct naming responses by total hours of treatment at entry and exit probes. (B) Model 2: Predicted proportion of correct naming responses by total number of trials at entry and exit probes. (C) Model 3: Predicted proportion of correct naming responses by average number of trials per hour at entry and exit probes. (D) Model 4: Predicted proportion of correct naming responses by average number of participant-generated features per trial at entry and exit probes. (E) Model 5: Predicted proportion of correct naming responses by total hours of treatment at exit and follow-up probes. (F) Model 6: Predicted proportion of correct naming responses by total number of trials at exit and follow-up probes. (G) Model 7: Predicted proportion of correct naming responses by average number of trials per hour at exit and follow-up probes. (H) Model 8: Predicted proportion of correct naming responses by average number of participant-generated features per trial at exit and follow-up probes.



It may intuitively seem that the participants who generated more features would also have completed more trials because they were able to progress through each trial more quickly, using less of the provided response time window and requiring fewer clinician cues. However, this possibility is not strongly supported by item-level data (i.e., items for which more features were generated per trial on average were not necessarily targeted across more trials). The average number of features generated per trial was only moderately correlated with trials per hour (r = .58) and total number of trials (r = .54). In addition, the effects of these two practice-related predictors on treatment-related naming improvement follow discrete patterns, further supporting the claim that they are measuring different constructs. Although the interaction of probe time and number of trials approached significance in the entry-to-exit model (p = .06 after correction for multiple comparisons), this effect is driven primarily by a treatment effect for treated items only. For instance, completing the additional 10 trials for an item above the mean (an increase from 30 to 40 trials) increased the odds of a correct response at exit by 1.14, corresponding to a predicted increase of only 3% (from 68% to 71%) for treated items, whereas it predicted a *decrease* in accuracy by 0.5% for untreated items. Although this effect is nonsignificant, it suggests that there is something unique about feature generation, as compared with simply completing more trials, that contributes to generalization and maintenance in SFA.

The significant effect of the average number of participant-generated features per trial on treatment response may best be exemplified by S2 and S3. As shown in Table 3, S2 responded well to SFA, improving from three out of 20 correctly named treated items at entry to 12 correct at exit and 10 at follow-up. In contrast, S3 consistently named only five of 20 treated items at both entry and exit and only seven at follow-up. S2 and S3 completed almost the same number of trials (441 compared with 369) and minutes in therapy (2,933 compared with 2,943), but S2 generated almost twice as many features per trial on average (8.24 compared with 4.4). Furthermore, S2 and S3 had similar severity measures, receiving CAT mean modality T-scores of 50 and 46.33, respectively. Although these are overall measures of performance collapsed across items for only two participants, this example is reflective of the overall trends that can be seen in the models reported above.

The finding that the average number of participant-generated features per trial predicts treatment response, generalization, and maintenance, whereas other aspects of intensity do not, supports the claim of Boyle (2010) and others that the client act of feature generation is the hypothesized source of the strengthened semantic network engendered by SFA and is therefore the "active ingredient" in SFA. This finding also more generally supports the view of Baker (2012) and Cherney (2012) that the "active ingredient" of aphasia treatment intensity is dose form and client acts in particular, rather than the traditionally reported measures of intensity such as total treatment time or number of trials.

It may be the case that, of the practice-related aspects of intensity investigated here, the number of participant-generated features alone predicted response to SFA because these aspects are targeting different stages of the word retrieval process. For instance, completing more trials may be targeting the phonological processing stage (mapping the lexical item to the phonological representation) through providing more opportunities for repeated paired association. This paired associative learning likely improves retrieval of the phonological word form for treated items only. This claim is supported by the observation that repeated picture-naming attempts, even in the absence of accuracy feedback (akin to repeated paired association), results in improved naming for trained items but no generalization (Nickels, 2002). As discussed in the Introduction, phonologically oriented naming treatment studies also often report improvement on treated items, but less frequently report gains for untreated items (e.g., Best et al., 2013; Leonard et al., 2008; Nickels, 2002). Such treatments focus explicitly on the phonological processing stage of lexical retrieval, and they often involve repeated paired association of a picture stimulus and a phonological word form, with phonological cues as needed (e.g., Wambaugh, 2003; Wambaugh et al., 2001). This may help explain the somewhat weaker evidence of generalization for such treatments and, relatedly, the lack of a predictive effect of the number of trials on generalization found in this study.

In contrast, generating features may be targeting the lexical-semantic processing stage by strengthening the link between a conceptual representation and the associated lexical-semantic representations—in line with the hypothesized mechanism underlying SFA treatment (Boyle, 2010; Boyle & Coelho, 1995). This claim is supported by the finding that requiring participants to *actively* generate features requires deeper semantic processing than is required for associative learning via repeated exposure (Boyle, 2010). In addition, generating features strengthens the lexical-semantic network not only for the treated item but also for those items that share semantic features (as well as related features via automatic spreading activation). As noted in the Introduction, SFA has demonstrated somewhat stronger evidence in promoting improved naming of untreated items (Boyle, 2010; Oh et al., 2016) especially when compared with phonologically oriented naming treatments. The perspective that feature generation targets lexical-semantic processing then may help explain the predictive effect of generating more features on both improved naming of treated items and generalization to untreated items, as well as the relatively good generalization effects found for SFA (and other semantically oriented naming treatments that require active participation; Boyle, 2010; Oh et al., 2016). The hypothesis that active productive participation in learning results in better outcomes is by no means novel or unique to clinical intervention in aphasia. The generation effect (Slamecka & Graf, 1978), that is, the finding that active generation of a target results in better learning of that target, is a robust finding in the memory and learning literature (see Bertsch, Pesta, Wiscott, & McDaniel, 2007 for a meta-analysis).

The current findings suggest that this effect also holds for aphasia rehabilitation. This may also explain part of the reported advantage of feature generation compared with the feature verification variants of SFA (Boyle, 2010).

As a final note, it is unlikely that the generalization to untreated semantically related items noted in this study can be exclusively attributed to exposure resulting from repeated naming probes, as has been argued to occur (Howard, 2000). In the current study, untreated items were probed once daily only while the paired treated items were being targeted, and every 4 days otherwise, which is a schedule that is unlikely to lead to significant learning or different degrees of learning across individuals. Furthermore, the probing schedule was the same across participants and items and was not dependent on any practice-related predictors such as number of trials or features generated. Thus, the probing schedule for untreated items is unlikely to be behind the observed relationship between the practice-related predictors and treatment response. Consistent with this conclusion, Wallace and Mason-Baughman (2012) also found that participants exhibited equal gains on infrequently and frequently probed untreated generalization items during SFA. While exposure to the untreated items during probes may have affected participants' naming responses, it is thus unlikely to have been responsible for the observed patterns of generalization, or their moderation by practice-related factors.

Study Limitations

There were notable limitations to the current study. First, the design (as is standard in the aphasia treatment literature) stipulated that each list was treated for a minimum number of sessions or until an accuracy criterion was met. This means that total treatment time or number of sessions for a list was inversely related to naming accuracy on probes: Lists that were treated for more sessions were likely those with which participants had more difficulty. However, two caveats apply to this limitation. First, it does not apply to individual items within a list, that is, this inverse relationship may hold at the list level but not at the item level. For example, a participant could have spent 5 min per trial on an easier item and 15 min per trial on a more difficult item within the same list, for which they completed the same number of trials. The item-level analysis used here permits these differences to be directly examined, whereas traditional analysis using list-level accuracy would not. Second, both the average number of trials per hour and the total treatment time did not significantly predict treatment response. This suggests that spending more time on an item did not lead to better outcomes, *unless* that time was spent generating more features. Nevertheless, in future studies, manipulating total treatment time per item rather than making advancement dependent on accuracy would better isolate the effects of treatment time on item-level gains, maintenance, and generalization. In relation, although the average number of participant-generated features was shown to be predictive of the degree of positive treatment response, this does not imply causality. In order to establish a causal relationship, the number of participant-generated features would have to be experimentally manipulated as well.

A second limitation concerns the potential bias introduced by the exclusion from the present analyses of the two participants who did not complete the protocol. While the estimates of the overall treatment effects are likely upwardly biased as a result of the exclusion of these participants, it is less clear how their exclusion might have affected the findings with respect to the number of features generated and the other intensity-related predictor variables. At present, the current results and conclusions can only be generalized to persons with aphasia who can tolerate and complete an intensive treatment protocol.

Another aspect of the study that may limit the generalizability of its results is that it varied from other SFA studies in that it included a sentence generation component. As noted in the Method section, each treatment trial ended with a cued sentence generation task. Perhaps, generating sentences was the "active ingredient" that invoked deeper semantic processing for participants. However, sentence generation was consistent across trials, that is, the number of opportunities for sentence generation did not vary depending on the number of features generated. This means that sentence generation is unlikely to account for the observed effect of feature generation. However, data collection is ongoing, and the contribution of sentence generation to treatment response (for example, how many sentences were successfully generated) will be evaluated in future analyses.

Future Directions

As noted above in the Study Limitations section, the effects of different practice-related variables are not fully independent. For instance, there is generally an inverse relationship between number of trials and average trials per hour: During a given 90-min treatment session, completing more trials will mean that each trial took less time to complete, on average, across all items. However, to explore the unique contribution of each of these practice-related factors, they would need to be evaluated in the same model. There were insufficient data to support such a model in the current sample of 17 participants. As data collection continues and the sample size in this ongoing study increases, more detailed analyses regarding the relationships among practice-related predictors will be possible.

As noted in the Introduction, SFA is intended to improve naming by both strengthening the semantic network and increasing semantic specification for a particular targeted lexical item. Analyses thus far have only considered the *quantity* of semantic features independently generated by participants, rather than the semantic specificity of the generated features, or their ability to distinguish the treated item from other items (either treated or untreated). For example, the features *round, fruit,* and *pit* all adequately describe *peach,* but they also describe *nectarine* and *plum* equally well. However, generating the feature *fuzzy* would

uniquely identify the target word *peach*. Generation of such distinguishing features may predict improved naming of the particular target word, but they may not improve the retrieval of other related items that do not share these distinguishing features. Therefore, it may be the case that generation of more general features predicts generalization to untreated items, whereas generation of more specific or distinguishing features predicts improvement on treated items. Wallace and Mason-Baughman (2012) have previously reported relationships between performance on word retrieval tasks and knowledge of feature distinctiveness among individuals with aphasia. Future analyses will therefore explore the effects of not only the quantity of features generated but also the *quality* (such as whether the feature is general or specific/distinguishing).

Related to this question, future analyses will also explore the contribution of generating unique features versus repeated generation of the same features across trials. Participants were encouraged to generate new features if possible, but there was considerable variability across participants in this regard. Generating more unique features for an item across trials should lead to greater specification or elaboration of the lexical-semantic representation for that item and, therefore, greater naming improvement. It should also lead to a more richly elaborated shared semantic space, which should engender greater generalization.

Conclusions

The results of the analyses presented here, on the basis of preliminary data from an ongoing clinical trial, replicate the results of previous studies that have found that SFA is efficacious (e.g., Boyle, 2010). However, they go beyond those previous findings in helping to identify key practice-related factors that contribute to the positive effects of SFA. In particular, they suggest that the "active ingredient" underlying the benefits of SFA, particularly with regard to generalization, is feature generation, consistent with the mechanisms hypothesized to underlie SFA treatment (Boyle & Coelho, 1995).

Furthermore, the current findings suggest that dose form, which includes the number of discrete participant acts completed during a treatment session, is an important aspect of treatment intensity (Cherney, 2012). Therefore, the fact that the elements of dose form are often under-reported in treatment studies is a potential limitation to fully understanding the effect of intensity on treatment outcomes in order to better inform clinical practice guidelines. The findings reinforce the importance of quantifying participant acts, in addition to the standardly reported number of sessions, hours, or treatment trials completed when reporting on outcomes of treatment studies. In the future, as more studies include measures of dose form, meta-analyses of treatment intensity should also take this aspect into consideration.

In a clinical sense, these findings suggest that what individuals with aphasia do during treatment is perhaps just as important as, if not more important than, how many times or for how long they do it. This knowledge should help focus clinicians' efforts in administering SFA, as well as other restorative treatments.

Acknowledgments

This research was supported by VA Rehabilitation Research and Development Award I01RX000832 to Michael W. Dickey and Patrick J. Doyle and the VA Pittsburgh Healthcare System Geriatric Research Education and Clinical Center. The contents of this article do not represent the views of the Department of Veterans Affairs of the U.S. Government. The authors would also like to acknowledge the contribution of the research participants and the Program for Intensive Residential Aphasia Treatment and Education clinical and research staff, including early contributions to the project by Brooke Swoyer and Beth Friedman.

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.

Baker, E. (2012). Optimal intervention intensity. *International Journal of Speech-Language Pathology, 14*(5), 401–409.

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.

Basso, A., & Macis, M. (2011). Therapy efficacy in chronic aphasia. *Behavioural Neurology, 24*(4), 317–325.

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. (2013). *lme4: Linear mixed-effects models using S4 classes. R package version 1.1-12*. Retrieved from http://cran.rproject.org/web/packages/lme4/index.html

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289–300.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition, 35*(2), 201–210.

Best, W., Greenwood, A., Grassly, J., Herbert, R., Hickin, J., & Howard, D. (2013). Aphasia rehabilitation: Does generalisation from anomia therapy occur and is it predictable? A case series study. *Cortex, 49*(9), 2345–2357.

Bhogal, S. K., Teasell, R., & Speechley, M. (2003). Intensity of aphasia therapy, impact on recovery. *Stroke, 34*(4), 987–993.

Boyle, M. (2004). Semantic feature analysis treatment for anomia in two fluent aphasia syndromes. *American Journal of Speech-Language Pathology, 13*(3), 236–249.

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

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

Butterworth, B., Howard, D., & Mcloughlin, P. (1984). The semantic deficit in aphasia: The relationship between semantic errors in auditory comprehension and picture naming. *Neuropsychologia, 22*(4), 409–426.

Cherney, L. R. (2012). Aphasia treatment: Intensity, dose parameters, and script training. *International Journal of Speech-Language Pathology, 14*(5), 424–431.

Cherney, L. R., Patterson, J. P., & Raymer, A. M. (2011). Intensity of aphasia therapy: Evidence and efficacy. *Current Neurology and Neuroscience Reports, 11*(6), 560–569.

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

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*(6), 407–428.

Conroy, P., Sage, K., & Lambon Ralph, M. A. (2009). Errorless and errorful therapy for verb and noun naming in aphasia. *Aphasiology, 23*(11), 1311–1337.

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.

Ferrill, M., Love, T., Walenski, M., & Shapiro, L. P. (2012). The time-course of lexical activation during sentence comprehension in people with aphasia. *American Journal of Speech-Language Pathology, 21*(2), S179–S189.

Foygel, D., & Dell, G. S. (2000). Models of impaired lexical access in speech production. *Journal of Memory and Language, 43*(2), 182–216.

Goodglass, H. (1980). Disorders of naming following brain injury. *American Scientist, 68,* 647–655.

Goodglass, H., & Wingfield, A. (1997). Word-finding deficits in aphasia: Brain-behavior relations and clinical symptomatology. In H. Goodglass & A. Wingfield (Eds.), *Anomia: Neuroanatomical and cognitive correlates* (pp. 3–28). Boston, MA: Academic Press.

Hertzog, L. R. (2015). *Plotting regression curves with confidence intervals for LM, GLM and GLMM in R*. Retrieved from http://rpubs.com/hughes/116374 (Accessed 8/19/2016)

Hinckley, J. J., & Craig, H. K. (1998). Influence of rate of treatment on the naming abilities of adults with chronic aphasia. *Aphasiology, 12*(11), 989–1006.

Howard, D. (2000). Cognitive neuropsychology and aphasia therapy: The case of word retrieval. In I. Papathanasiou (Ed.), *Acquired neurogenic communication disorders: A clinical perspective* (pp. 76–99). London, United Kingdom: Whurr.

Howard, D., & Orchard-Lisle, V. (1984). On the origin of semantic errors in naming: Evidence from the case of a global aphasic. *Cognitive Neuropsychology, 1*(2), 163–190.

Hutchison, K. A. (2003). Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psychonomic Bulletin & Review, 10*(4), 785–813.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*(4), 434–446.

Kendall, D. L., Rosenbek, J. C., Heilman, K. M., Conway, T., Klenberg, K., Rothi, L. J. G., & Nadeau, S. E. (2008). Phoneme-based rehabilitation of anomia in aphasia. *Brain and Language, 105*(1), 1–17.

Kertesz, A. (2006). *Western Aphasia Battery–Revised.* San Antonio, TX: The Psychological Corporation.

Leonard, C., Rochon, E., & Laird, L. (2008). Treating naming impairments in aphasia: Findings from a phonological components analysis treatment. *Aphasiology, 22*(9), 923–947. https://doi.org/10.1080/02687030701831474

Lesser, R., & Algar, L. (1995). Towards combining the cognitive neuropsychological and the pragmatic in aphasia therapy. *Neuropsychological Rehabilitation, 5*(1–2), 67–92.

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

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

Massaro, M., & Tompkins, C. A. (1994). Feature analysis for treatment of communication disorders in traumatically brain-injured patients: An efficacy study. *Clinical Aphasiology, 22,* 245–256.

Nickels, L. (2002). Therapy for naming disorders: Revisiting, revising, and reviewing. *Aphasiology, 16*(10–11), 935–979.

Nickels, L., & Howard, D. (1994). A frequent occurrence? Factors affecting the production of semantic errors in aphasic naming. *Cognitive Neuropsychology, 11*(3), 289–320.

Oh, S. J., Eom, B., Park, C., & Sung, J. E. (2016). Treatment efficacy of semantic feature analyses for persons with aphasia: Evidence from meta-analyses. *Communication Sciences & Disorders, 21*(2), 310–323.

Perkins, L., Crisp, J., & Walshaw, D. (1999). Exploring conversation analysis as an assessment tool for aphasia: The issue of reliability. *Aphasiology, 13*(4–5), 259–281.

Quique, Y., Evans, W. S., & Dickey, M. W. (2017). *Acquisition and generalization responses in aphasia naming treatment: a meta-analysis of Semantic Feature Analysis outcomes.* Manuscript submitted for publication.

R Core Team. (2015). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Robey, R. R. (1998). A meta-analysis of clinical outcomes in the treatment of aphasia. *Journal of Speech, Language, and Hearing Research, 41*(1), 172–187.

Schuell, H., Jenkins, J. J., & Jimenez-Pabon, E. (1964). *Aphasia in adults*. New York, NY: Harper & Row.

Silkes, J. P., & Rogers, M. A. (2012). Masked priming effects in aphasia: Evidence of altered automatic spreading activation. *Journal of Speech, Language, and Hearing Research, 55*(6), 1613–1625.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*(6), 592–604.

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

Thompson, C. (1989). Generalization research in aphasia. *Clinical Aphasiology, 18,* 195–222.

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

Wallace, S. E., & Mason-Baughman, M. B. (2012). Relationship between distinctive feature knowledge and word retrieval abilities in people with aphasia. *Aphasiology, 26*(10), 1278–1297.

Wambaugh, J. (2003). A comparison of the relative effects of phonologic and semantic cueing treatments. *Aphasiology, 17*(5), 433–441.

Wambaugh, J. L., Linebaugh, C. W., Doyle, P. J., Martinez, A. L., Kalinyak-Fliszar, M., & Spencer, K. A. (2001). Effects of two cueing treatments on lexical retrieval in aphasic speakers with different levels of deficit. *Aphasiology, 15*(10–11), 933–950.

Wambaugh, J. L., Mauszycki, S., Cameron, R., Wright, S., & Nessler, C. (2013). Semantic feature analysis: Incorporating typicality treatment and mediating strategy. *American Journal of Speech-Language Pathology, 22*(2), S334–S369.

Warren, S. F., Fey, M. E., & Yoder, P. J. (2007). Differential treatment intensity research: A missing link to creating optimally effective communication interventions. *Mental Retardation and Developmental Disabilities Research Reviews, 13*(1), 70–77.

Winans-Mitrik, R., Chen, S., Owens, R., Hula, W., Eichhorn, K., Ebrahimi, M., ... Doyle, P. (2013). *Tele-Rehabilitation*

Solutions for Patients with Aphasia. AVASLP National Conference, San Francisco, CA.

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

Ylvisaker, M., & Szekeres, S. (1985, November). Cognitive-language intervention with brain injured adolescents and adults. In *Annual Convention of the Illinois Speech-Language-Hearing Association*.