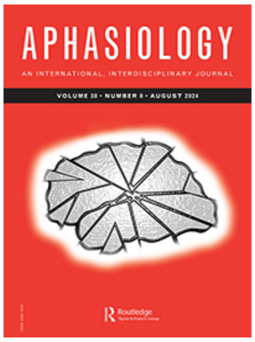

Aphasiology

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/paph20

Evolution of word production errors after typicality-based semantic naming treatment in individuals with aphasia

Ran Li, Natalie Gilmore, Mia O’Connell & Swathi Kiran

To cite this article: Ran Li, Natalie Gilmore, Mia O’Connell & Swathi Kiran (18 Jul 2024): Evolution of word production errors after typicality-based semantic naming treatment in individuals with aphasia, Aphasiology, DOI: 10.1080/02687038.2024.2377325

To link to this article: https://doi.org/10.1080/02687038.2024.2377325

Published online: 18 Jul 2024.

APHASIOLOGY
https://doi.org/10.1080/02687038.2024.2377325

Evolution of word production errors after typicality-based semantic naming treatment in individuals with aphasia

Ran Li [a], Natalie Gilmore [b,c], Mia O'Connell [d] and Swathi Kiran [e]

[a]Academy of Language and Culture, Hong Kong Baptist University, Hong Kong SAR, China; [b]Research and Development Service, James A. Haley Veterans' Hospital, Tampa, FL, USA; [c]Tampa VA Research and Education Foundation, Tampa, FL, USA; [d]Department of Neuroscience, Boston College, Boston, MA, USA; [e]Department of Speech, Language and Hearing Sciences, Boston University, Boston, MA, USA

ABSTRACT

Background: The Complexity Account for Treatment Efficacy (CATE) has been applied to semantic typicality in aphasia naming therapy, i.e. training atypical items of a category would improve naming of typical untrained-related items. However, most aphasia treatment studies have implemented a binary scoring system to measure response accuracy, which may not thoroughly reveal linguistic mechanisms underlying aphasia recovery.

Aims: The current study investigated the evolution of error patterns following typicality-based Semantic Features Analysis (SFA) treatment in individuals with post-stroke aphasia.

Methods & Procedures: Thirty individuals with chronic aphasia participated in a typicality-based SFA treatment, and ten individuals with chronic aphasia served as controls. The treatment participants and controls completed a naming screener before and after either a treatment period or a no-treatment period, respectively. Responses were coded using an error coding scale and analyzed with mixed-effects models.

Outcomes & Results: Treatment participants demonstrated significant treatment and generalization effects, as captured by significant improvements on the error coding system for both trained and untrained items. However, the group-level analysis did not reveal significant generalization from training atypical items to untrained typical items. Subgroup analyses based on participants' performance in treatment showed significant gains in naming untrained typical items from training atypical items in responders, but improved naming of untrained atypical items from training typical items in nonresponders.

Conclusions: These findings suggest different linguistic mechanisms underlying aphasia recovery and highlight the importance of investigating treatment and generalization effects using a fine-grained error coding system as a complement to a binary scoring system.

ARTICLE HISTORY
Received 21 January 2024
Accepted 3 July 2024

KEYWORDS
Aphasia; treatment; typicality; word production errors; generalization

CONTACT Swathi Kiran kirans@bu.edu Department of Speech, Language and Hearing Sciences, Center for Brain Recovery, Boston University, 111 Cummington Mall, Suite 280, Boston, MA 02215, USA

Background

Word production errors

Aphasia refers to impairments in speaking, listening, reading, and writing after damage to regions of the brain that constitute the language network, typically as a result of left hemisphere stroke, but could also be due to traumatic brain injury or other acquired neurological injury (Gilmore et al., 2022). One hallmark symptom in persons with aphasia (PWA) is naming impairment (Goodglass & Wingfield, 1997), which is characterized by lexical access and retrieval difficulties. Common types of naming or word production errors (Schwartz, 2014) exhibited by PWA include semantic errors (e.g., *cat* for *dog*), phonological errors (e.g., *mat* for *cat*), mixed errors (e.g., *rat* for *cat*), unrelated errors (e.g., *light* for *cup*), and nonwords or neologisms (e.g., *gak*). These word production errors may be attributed to impairment at different stages of the lexical access and retrieval process. Examining how patterns of word production errors may evolve with intervention has the potential to elucidate underlying mechanisms of aphasia recovery.

Detecting and examining error patterns requires an understanding of how word production occurs in the absence of errors. According to the two-step model of lexical access, three general levels (nodes) of linguistic representation are usually involved in word production: semantic, lemma, and phonologic (Dell, 1986; Levelt et al., 1999). Features of the word representing the concept an individual is trying to convey are activated, which leads to selection of, first, the most highly-activated lemma (word node) and, second, the most robustly-activated phonological information (phonemes).

The bidirectional and interactive nature of activation in the two-step model of lexical access described above provides opportunity for errors across all three levels of linguistic representation. Concepts share semantic nodes; and thus, semantic errors may occur due to competing feature activation of semantically-related words (e.g., intending to produce “son” and producing “daughter” as it has similar semantic features to “son”). Phonological errors may emerge due to interference when words that share phonemes are simultaneously activated (e.g., intending to produce “dug” and producing “duck” as the only phonemic difference between the two words is the final consonant). Mixed errors may arise (e.g., intending to produce “dog” and producing “duck” in that they are both animals that begin with the same phoneme /d/; Dell et al., 1997) when there is competing activation of semantically-related words that share phonological features. Unrelated errors, or word substitutions that are not semantically- or phonologically-related to the target (e.g., intending to produce “hairdryer” and producing “owl” as these words are from different semantic categories and do not share phonemes), may manifest due to interference during lemma access from words that are distantly related to the target or inaccurate word selection during the first stage of lexical access (Dell et al., 2004). Finally, disruptions of both lemma and phonological access can lead to the production of nonwords or neologisms – words with replacement of one or more target phonemes (e.g., plied; “I’ve got to plied up. I’ve got to plied again”; Butterworth, 1979). The severity of word production errors is contingent upon the level at which disruptions take place. As such, word production errors resulting from disruptions of lemma and phonological access (i.e., neologisms and unrelated errors) are expected to be more severe than those arising from interference at the semantic level (i.e., semantic errors). In turn, semantic errors are considered more severe than errors stemming from interference at the phonological level (i.e., phonemic errors). Examining the presence of these different types of word production errors in PWA as they progress through aphasia rehabilitation may shed light on the linguistic mechanisms underlying language recovery, and in particular, naming impairment, in PWA.

Semantic-based naming treatment

Aphasia rehabilitation has been shown to be effective for improving language ability, including naming difficulties (Brady et al., 2016; Quique et al., 2019). One of the most widely implemented therapy approaches for lexical impairment is Semantic Feature Analysis (SFA; Boyle & Coelho, 1995; Coelho et al., 2000), which aims to improve access to conceptual information by training semantic features associated with the target word. Targeting the semantic representation level should not only improve access to the target word, but also access to words sharing semantic nodes through spreading activation (Collins & Loftus, 1975). Previous studies have reported robust treatment effects from an SFA approach in that participants demonstrated improved acquisition of the trained stimuli (Boyle & Coelho, 2014; Coelho et al., 2000; Wambaugh et al., 2013). However, generalization effects after SFA treatment have been mixed (Kiran & Bassetto, 2008; Quique et al., 2019) with some studies observing gains in untrained items (Boyle & Coelho, 1995; Coelho et al., 2000) and others not finding evidence of such transfer (Wambaugh et al., 2013).

One framework that has been incorporated into speech and language intervention approaches to promote treatment gains and generalization effects is the Complexity Account for Treatment Efficacy (CATE; Kiran, 2008; Kiran & Thompson, 2003; Thompson et al., 2003) wherein the training of more complex treatment stimuli (e.g., syntactic structures, category exemplars) supports gains in untrained less complex treatment stimuli. Typicality is an aspect of semantic complexity that can be manipulated to drive generalization of untrained within-category exemplars. The concept of typicality posits that, within a category, words may show a graded structure with items that share more features with the category prototype being considered more "typical" and items that share fewer features with the prototype being considered more "atypical" (Rosch & Mervis, 1975). Applying the CATE to semantic complexity and typicality specifically suggests that training semantic features of atypical items should reinforce semantic features of typical items that are more centrally situated within the category (i.e., share more core features with prototype than the atypical item). Based on spreading activation theory as described previously, semantic activation should facilitate phonological access for these untrained items, which would be demonstrated by accurate word production or word production approaching the target (i.e., producing phonological errors rather than semantic, mixed, or unrelated errors).

Previous aphasia treatment studies applying CATE to semantic complexity have shown evidence of successful generalization to untrained typical items in individuals with naming deficits (Gilmore et al., 2018; Kiran & Thompson, 2003). However, they have largely implemented a binary scoring system to measure response accuracy (i.e., 0 = inaccurate, 1 = accurate; Gilmore et al., 2018; Kiran et al., 2011). While this scoring system is objective, it may be insufficiently granular to reveal linguistic mechanisms underlying treatment-induced aphasia recovery (e.g., progression on weekly probes from unrelated errors to semantic paraphasias to phonologic paraphasias to target response). The nature of naming recovery in aphasia may be better understood through a complementary approach that monitors and assesses word production error patterns over the course of treatment. This approach has been applied infrequently and using different methodologies (Drew & Thompson, 1999; Kendall et al., 2015; Kiran, 2008; Kiran & Thompson, 2003; Minkina et al., 2016). Some studies compared the percentage of word production errors from pre-treatment to post-treatment using a paired *t*-test or a chi-square test (Drew & Thompson, 1999; Minkina et al., 2016), whereas other studies investigated the change in the frequency or type of word production errors observed during standardized assessments, considering them an indirect reflection of the treatment effect (Kendall et al., 2015). However, none of these studies examined how word production errors change over the course of treatment at the item level or accounted for additional fluctuations in the naming outcome related to individual- and item-specific factors. This approach is particularly useful in the case of heterogeneous and modestly-sized participant samples (Wiley & Rapp, 2019) – a common feature of aphasia treatment studies, including the present work.

According to the previously mentioned two-step model (Dell, 1986; Levelt et al., 1999), the severity of word production errors depends on the level of lexical access at which disruptions may occur. One study proposed a 20-point error coding scale to examine lexical access in Spanish-English bilinguals with aphasia (Kiran et al., 2014). This scale ranks errors from most to least severe based on the degree of semantic and phonological access. No responses, perseverations, and neologisms are presumed to occur before semantic access of the target lemma; Unrelated word, circumlocutions, and semantic errors occur because of varying degrees of incomplete access at the semantic representation level; Mixed errors occur due to damaged connections between semantic and phonological levels; Phonemic errors and dysarthric responses occur after phonological access has occurred. The current study adopted and modified this previous coding scale to capture the evolution of different types of word production errors in individuals with aphasia.

In sum, naming therapy based on lexical-semantic models aims to improve lexical-retrieval deficits by targeting the semantic representation level. Findings from previous treatment studies that have intentionally manipulated the semantic complexity of treatment stimuli in line with the CATE suggest that explicitly training access to semantic features of atypical items in a category supports naming of untrained typical items in the same semantic category. However, treatment and generalization effects from these studies have predominantly been measured via probe accuracy or standardized language testing, neither of which are directly associated with the targeted linguistic dimension (e.g., semantic features); nor can fully capture the underlying pattern of aphasia recovery (e.g., increased access to semantic representations of target as evidenced by production of more semantically-related errors than unrelated errors).

The current study

The overarching goal of this study, to investigate the evolution of error patterns resulting from typicality-based SFA (Boyle & Coelho, 1995; Gilmore et al., 2018) for individuals with naming deficits, was addressed through two specific research questions:

(1) Do word production error patterns evolve for the trained and untrained items as relative to assessed items (i.e., items that were not in the same semantic categories as the trained and untrained items and exposed only before and after treatment)?

Hypotheses: According to the previous treatment study (Gilmore et al., 2018), in which treatment effects were measured by comparing performance on the trained items with that on the monitored items (i.e., items that were not in the same semantic categories as the trained and untrained items and probed weekly during treatment), significant improvements on the trained items were identified. Hence, it was hypothesized that compared to the assessed items, the typicality-based SFA would result in an evolution of word productions errors for both trained and untrained items, as evidenced by a significant gain on an error coding scale over the course of treatment (i.e., higher scores reflect less errored production).

(2) Do word production error patterns evolve for the typical untrained items as relative to atypical untrained items?

Hypotheses: In the context of semantic complexity, the CATE suggests that training semantic features of atypical items should result in improved lexical access to typical untrained items (Kiran & Thompson, 2003). Findings from previous treatment studies have shown that PWA who received semantic-based naming therapy for atypical items improved significantly on naming the untrained typical items rather than vice versa (Gilmore et al., 2018). Therefore, we hypothesized that word production errors would evolve for the typical untrained stimuli relative to the atypical untrained stimuli, as reflected by a significant increase on an error coding scale over time (Thompson et al., 2003).

Methods

Data were collected from our previous project under the Center for the Neurobiology of Language Recovery (NIH/NIDCD 1P50DC012283; PIL Cynthia Thompson).

Participants

Thirty individuals with chronic aphasia (20 male, M age: 61.5 ± 10.7 years; M months post-onset: 53.0 ± 50.0) due to left-hemisphere stroke were included in the current study and participated in a naming therapy (TX group). Additionally, ten natural history controls (10 male, M age: 56.9 ± 11.9 years; M months post-onset: 85.2 ±

142.0) were tested at baseline and after a three- to six-month period without naming treatment[1] (NH group). Participants were assigned to either TX or NH group in a pseudo-randomized fashion, that is, every fourth patient was asked to consider enrolling in the NH group but was given the option to enroll in the TX group (Johnson et al., 2019). Table 1 presents detailed demographic information for the TX and NH groups. Both groups were matched for age, months post-onset, education, and aphasia severity, two-sample *t*-test *p*-values > .05. Most participants in both groups demonstrated anomic or Broca's aphasia. The TX group additionally included two with conduction aphasia, two with Wernicke's aphasia, one with transcortical motor aphasia, and one with global aphasia. Participants were included if they met the following criteria: 1) had normal or corrected-to-normal hearing and vision, 2) were English-proficient pre-morbidly based on self or family report, 3) presented with stable neurological and medical status, 4) were not receiving concurrent speech and language therapy. None of the participants were diagnosed with other neurodegenerative disease (e.g., Parkinson's Disease, dementia) or active medical conditions. Written consent of participation was obtained in accordance with the Boston University Institutional Review Board (IRB) protocols.

Naming screening

All participants completed a 180-item confrontation naming screener as part of the enrollment criteria, which consisted of items from five semantic categories (i.e., *birds, vegetables, fruit, clothing*, and *furniture*) with 36 exemplars in each category. Exemplars in each category were further divided into half-categories by semantic typicality (i.e., 18 typical; 18 atypical). Pictures were randomly presented and participants were instructed to name each picture during the screener. A binary score of 0 or 1 was assigned to each response. The following responses received full score: intended target, self-corrections, dialectal differences, distortions, or substitutions of one phoneme, and/or correct response with a written self-cue. Participants were included in the study if they demonstrated stable performance (i.e., less than 65% average accuracy) in at least two different half-categories (e.g., typical *birds*, atypical *clothing*) across multiple baselines of the screener. We measured participants' apraxia of speech (AOS) given its common concurrence with aphasia (Gilmore et al., 2018). Individuals with AOS were included if they were stimulable to produce targets with a verbal model provided from a clinician. The AOS rating (Table 1) was based on the Screen for Dysarthria and Apraxia of Speech (S-DAOS) (Dabul, 2000) and clinical judgment from trained speech-language pathologists.

Standardized assessments

Participants in both TX and NH groups were administered a battery of standardized assessments before and after treatment or no-treatment period. Aphasia severity was determined via the Western Aphasia Battery-Revised (WAB-R; Kertesz, 2007). Other assessments included Boston Naming Test (BNT; Kaplan et al., 2001) to test noun naming ability, Pyramids and Palm Trees Test (PAPT; Howard & Patterson, 1992) to assess semantic processing, Cognitive and Linguistic Quick Test (Helm-Estabrooks, 2001) to measure

Table 1. Demographics, baseline performance on the standardized language testing, and classification of treatment responders vs. nonresponders.

ID	Age	Gender	Handedness	MPO	Edu	WAB-AQ (/100)	Aphasia Type	BNT (/60)	PAPT (/52)	PALPA51 (/30)	CLQT-CS (/100)	AOS	Tx Resp.
Treatment participants (TX)													
BU01	55	M	R	12	16	87.2	A	50	50	23	90	AB	R
BU02	50	F	L	29	16	25.2	G	1	49	3	50	UN	NR
BU03	63	F	R	62	16	52.0	C	10	46	21	55	PR	R
BU04	79	M	R	13	16	74.1	C	52	49	18	80	AB	R
BU05	67	M	R	10	18	30.8	W	4	48	9	50	AB	NR
BU06	49	M	R	113	16	66.6	B	44	48	22	80	PR	R
BU07	55	M	R	137	16	48.0	B	6	46	12	60	AB	R
BU08	49	M	R	57	12	82.8	A	51	48	22	100	PR	R
BU09	71	F	R	37	16	95.2	A	45	50	26	95	AB	R
BU10	53	F	R	12	16	80.4	A	37	49	24	80	AB	R
BU11	78	M	R	22	18	92.1	A	41	49	22	75	AB	NR
BU12	68	M	R	104	12	40.0	B	1	46	12	60	PR	NR
BU13	42	M	L	18	14	92.7	A	43	49	21	95	AB	R
BU14	64	F	R	24	13	64.4	B	41	49	16	80	AB	R
BU15	71	F	R	74	12	87.2	A	43	44	16	60	AB	R
BU16	50	M	R	71	12	33.6	B	1	41	3	70	UN	NR
BU17	61	M	R	152	16	74.3	A	54	51	21	90	AB	R
BU18	70	F	R	152	16	78.0	A	24	50	15	85	AB	R
BU19	80	M	R	22	18	28.9	B	1	43	8	70	PR	R
BU20	48	F	R	14	16	13.0	B	0	40	10	45	PR	R
BU21	65	M	R	16	18	11.7	B	0	43	10	45	PR	NR
BU22	62	M	R	12	16	65.4	TCM	1	37	11	45	AB	R
BU23	60	M	R	24	16	45.2	W	6	42	6	45	AB	NR
BU24	69	M	R	169	16	40.4	B	3	49	8	70	PR	NR
BU25	76	F	R	33	18	37.5	B	2	34	19	55	AB	NR
BU26	64	F	R	115	12	58.0	B	15	36	12	30	AB	NR
BU27	65	M	L	17	12	84.3	A	41	50	21	75	AB	R
BU28	62	M	R	15	12	56.0	B	21	51	15	70	n/a	R
BU29	40	M	R	26	16	90.1	A	36	51	18	n/a	n/a	NR
BU30	59	M	R	29	14	60.0	B	16	50	13	75	n/a	NR
M	61.5			53.0	15.2	59.8		23.0	46.3	15.2	68.3		
SD	10.7			50.0	2.1	25.0		20.3	4.7	6.4	18.1		

(*Continued*)

Table 1. (Continued).

ID	Age	Gender	Handedness	MPO	Edu	WAB-AQ (/100)	Aphasia Type	BNT (/60)	PAPT (/52)	PALPA51 (/30)	CLQT-CS (/100)	AOS	Tx Resp.
Control participants (NH)													
BUc01	49	M	R	49	12	85.5	A	53	49	20	85	PR	
BUc02	79	M	R	10	18	26.9	B	3	47	10	n/a	PR	
BUc04	47	M	R	19	16	91.4	A	n/a	47	24	n/a	n/a	
BUc05	49	M	R	67	12	32.3	B	3	44	2	70	UN	
BUc06	69	M	R	165	16	39.3	B	5	48	12	70	PR	
BUc07	39	M	R	18	16	83.6	A	36	52	14	n/a	PR	
BUc08	64	M	L	13	12	79.6	A	n/a	50	17	n/a	AB	
BUc09	62	M	L	21	16	92	A	39	49	21	75	AB	
BUc011	58	M	R	23	14	61.8	B	10	51	18	60	PR	
BUc012	53	M	R	467	17	91.2	A	51	49	22	100	AB	
M	56.9			85.2	14.9	68.4		25.0	48.6	16.0	76.7		
SD	11.9			142.0	2.2	26.2		21.9	2.3	6.6	14.0		

Gender: M (male), F (female); R: right, L: left; MPO: months post-onset; Edu: years of education; WAB-AQ: Western Aphasia Battery – Aphasia Quotient; Aphasia Type: A (Anomic), B (Broca's), W (Wernicke's), C (Conduction), G (Global), TCM (Transcortical Motor); BNT: Boston Naming Test; PAPT: Pyramids and Palm Trees; PALPA51: subtests 51 (Word Semantic Association) of the Psycholinguistic Assessments of Language Processing in Aphasia; CLQT-CS: Cognitive and Linguistic Quick Test Composite Score; Apraxia of speech (AOS), AB: Absent, PR: Present, UN: Undetermined. TX participants were classified into treatment responders (R; effect sizes > 4.0 in at least one trained category) vs. nonresponders (NR; effect sizes < 4.0 in both trained categories) based on Gilmore et al. (2018); n/a: not available.

general cognitive abilities, and subtest 51 (Word Semantic Association) of the Psycholinguistic Assessments of Language Processing in Aphasia (PALPA; Kay et al., 1996) to assess lexical semantic processing. Table 1 shows participants' baseline performance on these tests. The TX and NH groups showed comparable scores on all language assessments, two-sample *t*-test *p*-values > .05, except for the PAPT, $p = .046$, in which the NH group ($M = 48.6$) scored higher than the TX group ($M = 46.3$).

Stimuli

The stimuli included items and their corresponding semantic features, which were selected based on results of separate MTurk pilot tasks (https://www.mturk.com/mturk/). Full details of the study stimuli and how they were derived can be found in Gilmore et al. (2018).

Study design

This study implemented a single-subject design with group-level analyses (Figure 1). All participants completed the 180-item naming screener three times – before treatment for the TX group or before a period of no-treatment for the NH group (Pre 1–3). According to their performance across three baselines, a total of 108 stimuli were selected from the 180-item naming screener for each participant. These items comprised of two trained half-categories, i.e., items that were directly trained during the twice weekly treatment session (e.g., 18 atypical birds, 18 typical furniture); two untrained half-categories, i.e., untrained items that belonged to the same semantic categories as the trained items and exposed weekly during the probe session to assess within-category generalization (e.g., 18 typical birds, 18 atypical furniture); one monitored category, i.e., items that were not in the same categories as the trained and untrained items and exposed weekly during the probe session (e.g., 18 typical and 18 atypical clothing); and one assessed category, i.e., items that were not in the same categories as the trained and untrained items and exposed only before and after treatment during the screener administration (e.g., 18 typical and 18 atypical vegetables). These five categories were assigned in a counterbalanced fashion. To match the procedures applied to the TX group for further group-level comparisons, 108 items were selected from the naming screener for participants in

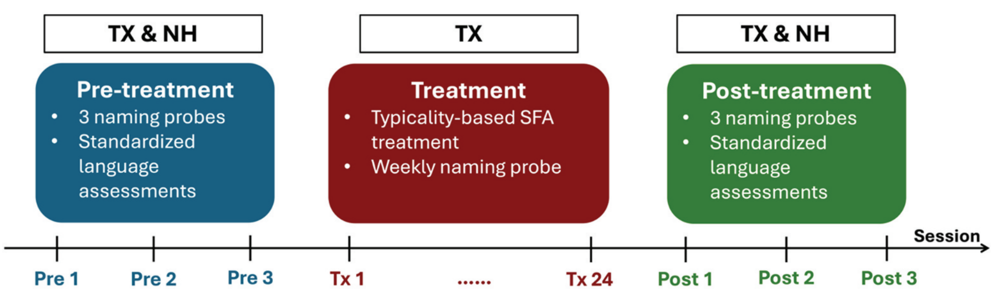

Figure 1. *Study design*. TX: Treatment group, NH: Natural history control group. Session: Pre 1–3 (pre-treatment sessions), T x 1–24 (treatment sessions), Post 1–3 (post-treatment sessions).

the NH group. The categories were labelled as "control-trained" half-categories, "control-untrained" half-categories, "control-monitored" category, and "control-assessed" category. None of these items were used for treatment since the NH group did not undergo therapy as a part of study participation.

Participants in the TX group then participated in 24 two-hour treatment sessions either two or three times per week or until they reached at least 90% accuracy on two consecutive weekly probes for both trained half-categories (Tx 1–24). During the weekly probe, TX participants were assessed on all items in their two trained half-categories, two untrained half-categories, and the monitored category (total $n = 108$). After treatment, participants were readministered the 180-item naming screener three additional times to establish post-treatment accuracy (Post 1–3). After the no-treatment period, participants in the NH group were assessed on their assigned categories.

Treatment

The current study implemented a typicality-based SFA treatment (Boyle & Coelho, 1995, 2014; Gilmore et al., 2018). Participants in the TX group attended individual speech therapy weekly and were trained on the two assigned half-categories (e.g., 18 atypical birds and 18 typical furniture). Treatment was delivered via E-prime tasks (Schneider et al., 2002), and involved the following six steps: 1) category sorting, 2) an initial naming attempt, 3) written feature verification, 4) feature review, 5) auditory feature verification, and 6) a second naming attempt.

Error analysis

Error coding

For each participant, probe responses at both pre- and post-treatment timepoints (Pre 1–3, Post 1–3) were recorded by clinicians. Then the best responses were coded based on a 0–12 scale (see description and examples in Appendix A): 0 (no response); 1 (neologism); 2 (real word perseveration); 3 (unrelated); 4 (inaccurate description); 5 (circumlocution); 6 (superordinate category error); 7 (semantic error); 8 (mixed error); 9 (phonemic error/apraxic); 10 (dysarthric/distortion); 11 (partial target); and 12 (target). This coding scale differed slightly from the one used in Kiran et al. (2014) in that we included three additional error types. First, inaccurate description occurred when utterances provided information that was inaccurate to the target, so that it was scored lower than circumlocution that typically includes accurate description of the target. Second, superordinate category errors occur when the response is a superordinate category of the target. This type of error indicates some semantic access, but to a lesser extent than semantic errors. Third, partial target occurs when part of the response is accurate, indicating more phonological access than phonemic and dysarthric errors, but not as complete as the target response. The coding was completed by trained research assistants who were not blinded to the session or group. The coded results were cross-checked by the first and second authors, and any discrepancies were resolved to 100% agreement. This coding system required a fully accurate response to receive a score of 12 (i.e., did not allow any phonemic errors for the target responses) to capture phonemic errors, a distinction from the binary criterion used to score naming accuracy in Gilmore et al. (2018), in which both fully accurate responses and responses with a phonemic error were given full credit.

To facilitate the accuracy and efficiency of the error coding process, a Python program was created to automate the coding of 0s (no responses), 12s (target), and 6s (superordinate category error). Specifically, "zero = ["not familiar", 'I have no idea'. . .]" was a list of common "no responses". If the response was equal to any term in the "zero" list, then its value was automatically coded as a 0. If the response was the superordinate category of the target (i.e., *bird* for *parakeet*), then the response was coded as a 6. Finally, when the response was the target word (i.e., accurate without any phonemic errors), it was coded as a 12.

Data analysis

All the analyses were performed in R Studio (version 4.2.3) for the TX and NH group separately. The Gilmore et al. study revealed variability in the response to treatment based on the treatment effect sizes. Therefore, we followed the same benchmarks (Beeson & Robey, 2006) to classify participants into treatment responders (i.e., effect sizes > 4.0 in at least one trained category) and nonresponders (i.e., effect sizes < 4.0 in both trained categories; see Table 1 for the individual classification). The same analyses were repeated within each subgroup.

First, to estimate treatment and generalization effects (Research Question 1), we conducted two sets of analysis: 1) logistic mixed-effects model to estimate the probability of a response (versus no response) to examine if participants significantly improved their attempts to name targets agnostic to error type (e.g., "I don't know" ➔ "something to wear"), and 2) linear mixed-effects models to estimate the change of error code (i.e., 1–12). Inherently, nonresponses (e.g., "I don't know") do not convey information about the underlying linguistic impairment that may be driving the access or retrieval issue. Hence, we excluded nonresponses in this analysis because we were interested in the evolution of errors, which required responses to code. In the logistic mixed models, the dependent variable was categorical reflecting response in binary fashion (i.e., "no response" = 0, "response" = 1). Fixed factors included item type (i.e., typical trained, atypical trained, typical untrained, atypical untrained, monitored, and assessed [reference level]), probe session (i.e., Pre 1–3, Post 1–3), and their interaction term. In the linear mixed models, we used probe session (i.e., Pre 1–3, Post 1–3), item type (i.e., trained typical, untrained typical, trained atypical, untrained typical, monitored, and assessed [reference level]), and their interaction to predict error code (i.e., 1–12; dependent variable). Differences between the trained categories and the assessed category would indicate the direct treatment effect, whereas differences between the untrained categories and the assessed category would suggest generalization to untrained related items. If there were no significant difference between the monitored and assessed items, then there would be no generalization for items that were unrelated to the trained or untrained categories. In both sets of models, the WAB-AQ was centered (to $M = 0$) and included as a covariate to account for the contribution of aphasia severity on error evolution. Random intercepts for subject and item were included given the potential for individual variations across subjects and items to influence accuracy of the response. Additionally, a by-subject random slope for the effect of session was added to account for potential inter-subject variability in treatment-induced changes over time.

Then, to estimate the generalization effect from training atypical items to typical untrained items (Research Question 2), a linear mixed-effects model was conducted to predict error code (i.e., 1–12; dependent variable) for the typical untrained and atypical untrained items. Fixed factors included item type (i.e., typical untrained, atypical untrained [reference level]), probe session (i.e., Pre 1–3, Post 1–3), and their interaction. The WAB-AQ was included as a covariate. Random intercepts for subject and item and a by-subject random slope for the effect of session were entered as described in the previous models.

Results

Table 2 illustrates the means, standard deviations, and ranges of error code scores at pre- and post-treatment timepoints for each stimulus condition.

Treatment group (TX)

Treatment and generalization effects

Results from the logistic mixed-effects model (Table 3) revealed that the likelihood of producing a response for the trained stimuli significantly increased over time for both typical trained and atypical trained stimuli relative to the assessed stimuli. However, the likelihood of a response for the typical untrained, atypical untrained, and monitored stimuli decreased over time. These findings suggest that word production for the trained

Table 2. Means (M), standard deviations (SD), minimum (Min.) and maximum (Max.) of error coding scores at pre-treatment (Pre) and post-treatment (Post) for treatment (TX) and control (NH) groups.

	Typical Trained		Atypical Trained		Typical Untrained		Atypical Untrained		Monitored		Assessed	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
TX												
M	6.18	7.80	5.57	7.69	5.57	5.82	5.88	6.22	6.40	6.56	5.89	6.22
SD	4.70	4.83	4.75	5.04	4.47	4.94	5.04	5.07	4.79	5.00	4.87	4.86
Min.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Max.	12.00	12.00	12.00	12.00	12.00	12.00	12.00	12.00	12.00	12.00	12.00	12.00
NH												
M	6.42	7.58	4.20	6.67	4.01	5.10	6.65	7.06	6.54	6.72	6.06	6.90
SD	4.82	4.79	5.14	5.33	4.97	4.97	5.18	5.15	5.14	5.07	5.32	5.06
Min.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Max.	12.00	12.00	12.00	12.00	12.00	12.00	12.00	12.00	12.00	12.00	12.00	12.00

Scores for Pre are averaged across sessions Pre 1–3, scores for Post are averaged across sessions Post 1–3.

Table 3. Logistic mixed-effects model results for the treatment group (TX) and control group (NH).

	TX (all patients)		TX (responders)		TX (nonresponders)		NH	
Predictors	*Coeff. (β)*	*SE*	*Coeff. (β)*	*SE*	*Coeff. (β)*	*SE*	*Coeff. (β)*	*SE*
Session*TyTr	.10**	.03	.17**	.05	.03	.05	.08	.05
Session *AtyTr	.26**	.04	.30**	.05	.22**	.05	−.08	.07
Session *TyUn	−.07*	.03	−.09	.05	−.05	.05	−.17**	.07
Session *AtyUn	−.08*	.03	−.11*	.05	−.08	.05	−.01	.05
Session *Monitored	−.09**	.03	−.11**	.04	−.08*	.04	−.11*	.04

Regression coefficients (β) and standard errors (SE) are reported. TyTr: typical trained, AtyTr: atypical trained, TyUn: typical untrained, AtyUn: atypical untrained; * $p < .05$, ** $p < .01$.

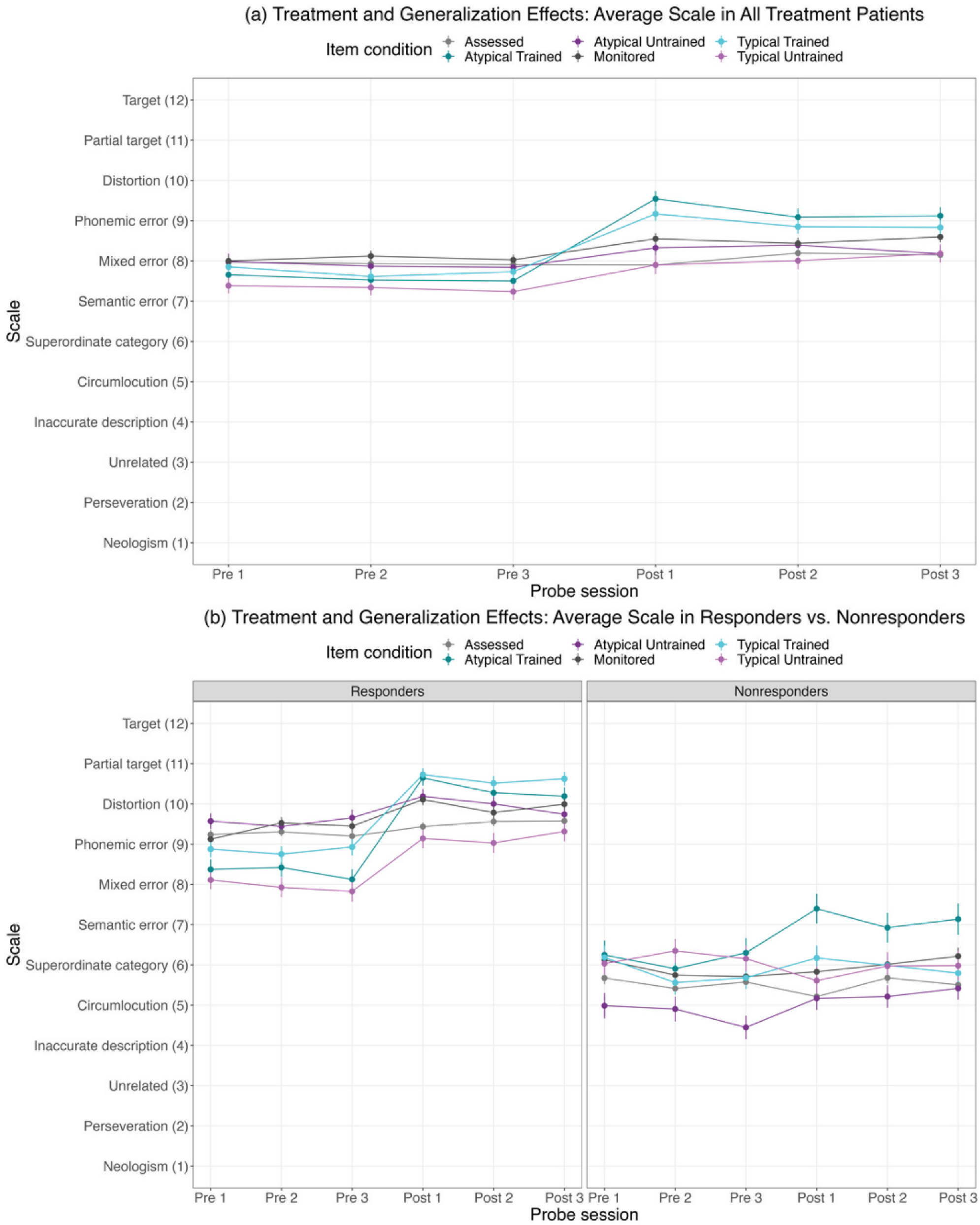

Figure 2. *Evolution of word production errors in the trained and untrained conditions in the treatment group (TX). Note*. Average scale across participants for (a) all treatment patients and (b) treatment responders vs. nonresponders. Y-axis: error coding scale 1 (neologisms) – 12 (target). X-axis: probe sessions (Pre 1–3, Post 1–3).

stimuli was more likely to evolve from no response to response compared to the other stimulus conditions over the course of treatment in the TX group.

Figure 2(a) illustrates the evolution of word production errors. The results showed a significant session*trained interaction for both typical trained, $\beta = .30$, $SE = .03$, $t = 9.40$,

$p < .001$, and atypical trained conditions, β = .38, *SE* = .04, *t* = 10.10, $p < .001$, which suggested that word production of the trained stimuli became closer to the target over time (i.e., from semantic/mixed errors to phonemic errors), supporting a treatment effect. Additionally, a significant session*untrained interaction effect was found for both typical untrained, β = .09, *SE* = .04, *t* = 2.22, *p* = .03, and atypical untrained conditions, β = .09, *SE* = .03, *t* = 2.76, *p* = .01, indicating that word production became closer to the target over time for the typical untrained (i.e., from semantic errors to mixed errors) and atypical untrained (i.e., from mixed errors to phonemic errors) stimuli, emphasizing generalization to the untrained within-category items.

Generalization effect of typicality

Table 2 reports performance on the typical untrained and atypical untrained conditions, which assesses the CATE hypothesis. Figure 3(a) shows the evolution of word production errors in the typical untrained and atypical untrained conditions for all the treatment participants. Results from the linear-mixed model did not reveal a significant session*typical untrained interaction effect, *p* = .42, indicating that training semantic features of atypical items of a semantic category did not generalize to typical untrained items of the same sematic category over time. In other words, when all the treatment participants were grouped together, this finding did not support the CATE for semantic complexity as hypothesized.

Follow-up analysis within the TX group

Figure 2(b) captures the evolution of word production errors in treatment responders and nonresponders groups. When estimating the treatment and generalization effects in treatment responders, the likelihood of producing a response (vs. no response) significantly increased for the typical trained and atypical trained stimuli over time, whereas decreased for the atypical untrained and monitored stimuli (Table 3), suggesting that word production in responders was more likely to evolve from no response to response for the trained stimuli relative to the other conditions.

When estimating the change of error code (i.e., 1–12), results from the linear mixed-model revealed a significant session*trained interaction effect for both typical trained, β = .44, *SE* = .04, *t* = 11.61, $p < .001$, and atypical trained conditions, β = .46, *SE* = .04, *t* = 10.52, $p < .001$, and another significant session*typical untrained interaction, β = .15, *SE* = .04, *t* = 3.33, $p < .001$. These findings suggested that word production in responders became closer to the target for the typical trained (i.e., from phonemic errors to partial target), atypical trained (i.e., from mixed errors to distortion), and typical untrained (i.e., from mixed to phonemic errors) stimuli.

For treatment nonresponders, the likelihood of producing a response significantly increased over time for the atypical trained condition, but decreased for the monitored condition relative to the assessed stimuli (Table 3), indicating that word production was more likely to evolve from no response to response for the trained stimuli than the other conditions even in participants who did not respond favorably to treatment.

The linear mixed-model results showed significant session*atypical trained, β = .24, *SE* = .07, *t* = 3.51, $p < .001$, and session*atypical untrained interactions, β = .20, *SE* = .06, *t* = 3.48, $p < .001$, suggesting that word production became closer to the target over time for both atypical trained (i.e., from superordinate category to semantic errors) and atypical

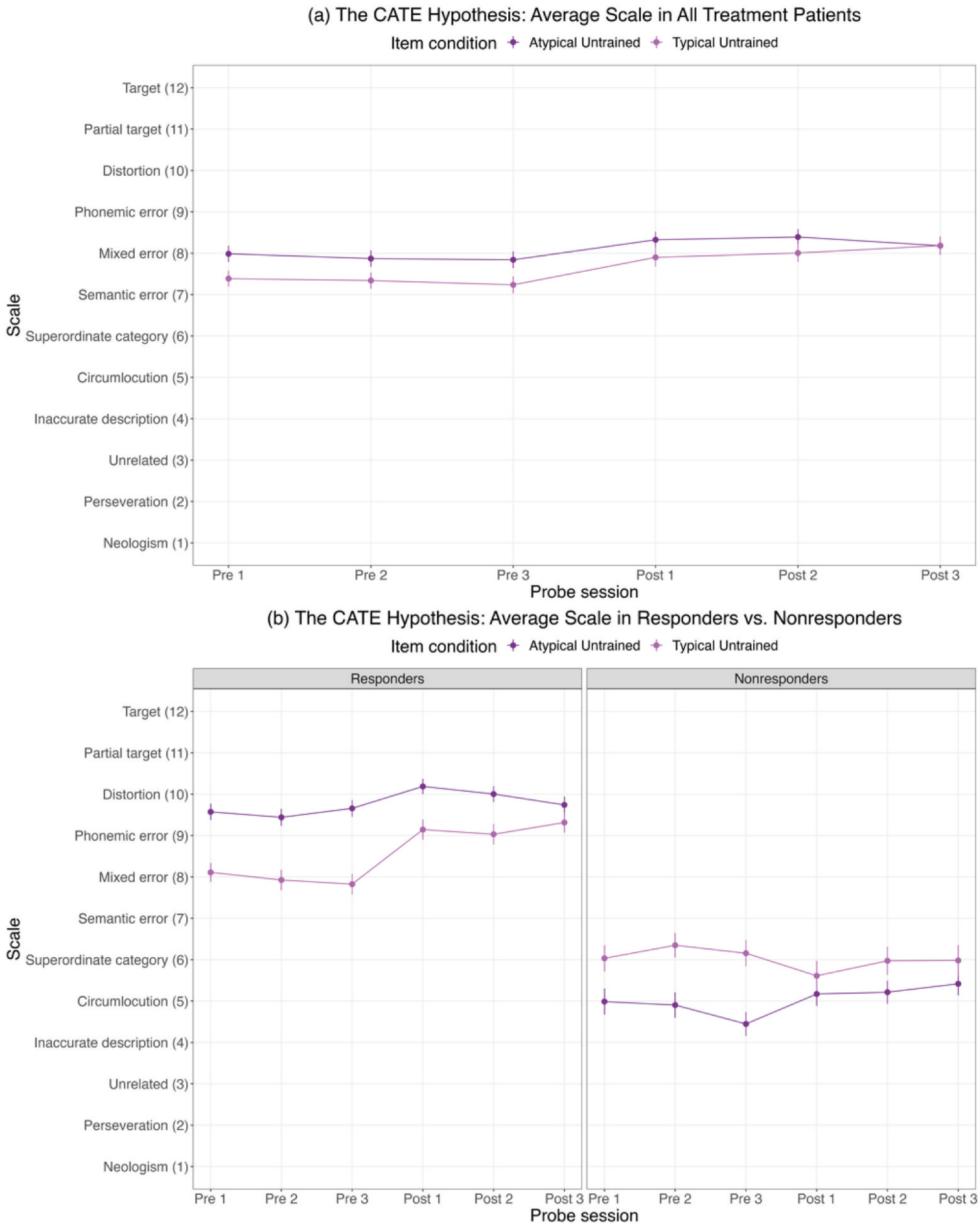

Figure 3. *Evolution of word production errors in atypical untrained and typical untrained conditions in the treatment group (TX). Note.* Average scale across participants for (a) all treatment patients and (b) treatment responders vs. nonresponders. Y-axis: error coding scale 1 (neologisms) – 12 (target). X-axis: probe sessions (Pre 1–3, Post 1–3).

untrained (i.e., from circumlocution/inaccurate description to superordinate category) items. Taken together, the above findings are similar to the whole-group results, supporting a direct treatment effect and generalization to untrained related items in both treatment responders and nonresponders.

When estimating the generalization effect of typicality in treatment responders (Figure 3(b)), there was a significant positive session*typical untrained interaction effect, $\beta = .15$, $SE = .06$, $t = 2.32$, $p = .02$, suggesting that training atypical items promoted generalization to untrained typical items (i.e., from mixed to phonemic errors), which corroborated the CATE hypothesis (Thompson et al., 2003). However, for treatment nonresponders (Figure 3(b)), the results showed a significant negative session*typical untrained interaction effect, $\beta = -.35$, $SE = .12$, $t = -2.81$, $p = .01$, indicating that training typical items facilitated generalization to untrained atypical items (i.e., from circumlocution/inaccurate descriptions to superordinate category errors). Overall, these findings showed that while training semantic features of atypical items appeared to transfer to naming typical untrained items in treatment responders (i.e., supporting the CATE hypothesis), it did not in treatment nonresponders (i.e., did not support CATE hypothesis). The disparate generalization patterns we observed in these two groups will be returned to in the discussion.

Control group (NH)

Treatment and generalization effects

The same analyses above were conducted at the group level for the control participants to examine whether patterns of word production errors for the control-trained and control-untrained items changed during the no-treatment period. Results from the logistic mixed-model showed that the likelihood of producing a response (vs. no response) significantly decreased for typical control-untrained and control-monitored items (Table 3), indicating no evolution from no response to response in any stimulus conditions.

Results from the linear mixed-model (Figure 4(a)) revealed a significant session*trained interaction for both typical control-trained, $\beta = .10$, $SE = .05$, $t = 1.96$, $p = .05$, and atypical control-trained conditions, $\beta = .22$, $SE = .08$, $t = 2.61$, $p = .01$, as compared to the control-assessed items, suggesting an evolution of word errors in the control-trained stimuli (i.e., from mixed errors to phonemic errors) despite the NH group not receiving typicality-based SFA treatment.

Generalization effect of typicality

Figure 4(b) demonstrates the evolution of word production errors for the typical control-untrained and atypical control-untrained conditions in the NH group. Results from the linear mixed-effect model estimating the generalization of typicality did not reveal a significant session*typical untrained interaction, $p = .91$. This finding is consistent with the initial hypothesis and indicates no generalization from atypical control-trained items to typical control-untrained items in the NH group, who did not receive typicality-based SFA treatment.

Follow-up analysis between NH and TX groups

Given that the NH group demonstrated an evolution of word production errors in the control-trained stimuli in the absence of treatment, a follow-up linear mixed-effects model estimating the change of error code (i.e., 1–12) was performed by comparing the NH group with treatment responders and nonresponders (three-way interaction: session*item type*group). Results showed significant session*typical trained*responders,

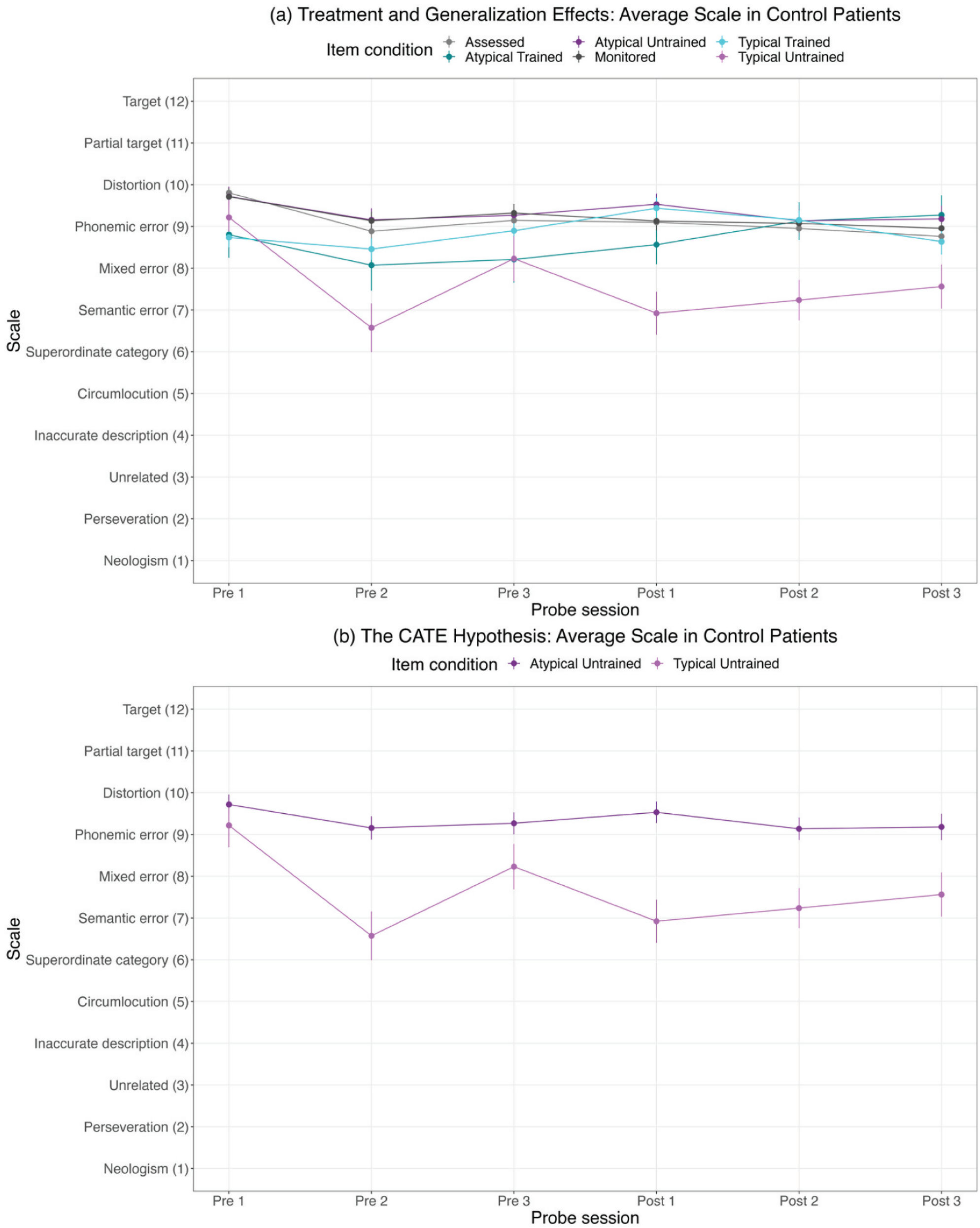

Figure 4. *Evolution of word production errors in the (a) trained and untrained conditions and (b) atypical untrained and typical untrained conditions in the control group (NH). Note.* Average scale across all control participants. Y-axis: error coding scale 1 (neologisms) – 12 (target). X-axis: probe sessions (Pre 1–3, Post 1–3).

$\beta = .33$, $SE = .07$, $t = 4.93$, $p < .001$, and session*atypical trained*responders three-way interactions, $\beta = .22$, $SE = .10$, $t = 2.28$, $p = .02$. These findings indicated that relative to the NH group, the magnitude of change in word production errors for the trained stimuli was significantly greater in treatment responders, in support of a larger treatment effect in this subgroup of participants.

Discussion

The current study investigated patterns of word production errors over the course of a typicality-based SFA treatment in individuals with chronic aphasia. Specifically, we examined whether word production errors evolved 1) for the trained and untrained items vs. assessed items, and 2) for the typical untrained items vs. atypical untrained items. The treatment participants showed significant treatment and generalization effects, as evidenced by significant changes on error coding scores for both trained and untrained items relative to the assessed items. However, the group-level results did not reveal any significant generalization from training atypical items to typical untrained related items consistent with our previous work using a binary scoring system (Gilmore et al., 2018). Follow-up analyses based on classifying treatment participants into responders and nonresponders revealed treatment and generalization effects anticipated from typicality-based SFA treatment. Moreover, treatment responders demonstrated significant gains in naming typical untrained items, whereas treatment nonresponders showed a different generalization pattern – gains in naming atypical untrained items. These findings provide evidence of treatment and generalization effects via the implementation of a fine-grained error coding system, and suggest different linguistic mechanisms underlying aphasia recovery for individuals with varying treatment benefits.

Treatment and generalization effects

For all the treatment participants, the likelihood of a response (vs. no response) significantly increased for the trained items. Additionally, word production errors for the trained items evolved from semantic/mixed errors to phonemic errors as expected given literature suggesting that targeting the semantic system will support improvements in the phonological system. These findings corroborate a significant treatment acquisition effect, as identified in the previous study using a binary scoring system (Gilmore et al., 2018). The decrease of "no responses" (i.e., 0s on the error coding scale) could be due to increased activation in the semantic network as a result of the intervention. The interactive activation model of word production has posited that omissions arise because of insufficient activation of nodes at the semantic level (Dell et al., 2004). Hence, the lemma selection step is not initiated, and an overt response does not occur. The finding that word production errors changed from semantic/mixed errors to phonemic errors further supports this assumption. Semantic errors before treatment were possibly caused by difficulty with activating the target semantic nodes. However, the semantic network was strengthened during intervention to facilitate lexical access of the trained items (Boyle & Coelho, 1995). Therefore, participants might have shifted away from semantic errors towards errors that occur at later stages of the word retrieval process over the course of treatment (i.e., phonemic errors), reflecting a stage closer to the target.

The group-level findings also showed a significant change of word production errors for the untrained stimuli, from semantic to mixed errors for the typical untrained stimuli and from mixed to phonemic errors for the atypical untrained stimuli, suggesting that typicality-based SFA treatment facilitates generalization to semantically related but untrained items (Gilmore et al., 2018). Since words at the semantic representation level share conceptual nodes, training semantic features of an item should not only activate the target word, but also spread its activation to items sharing similar semantic attributes (Collins & Loftus, 1975). As a result of the spreading activation, word production errors evolved from those occurring at earlier stages (i.e., semantic/mixed errors) to those occurring at later stages (i.e., mixed/phonemic errors) of the lexical-retrieval process. These findings indicate that error scoring allows us to tap into the linguistic mechanisms underlying the treatment effects of semantic-based naming intervention (Boyle et al., 2022).

When effect sizes were calculated based on naming accuracy scores derived from a binary scoring system (Gilmore et al., 2018), participants' performance was classified into being representative of a treatment "responder" or "nonresponder". Nevertheless, the logistic regression models demonstrated that both responders and nonresponders made significant improvements in naming trained stimuli (i.e., transitioning from "no response" to "response"), suggesting that targeting the semantic network could enable patients to attempt a response. The separate linear regression models revealed that the evolution patterns of word production errors varied between subgroups. For responders, word production errors evolved from phonemic errors to partial target for the typical trained stimuli, from mixed errors to distortion for the atypical trained stimuli, and from mixed errors to phonemic errors for the typical untrained stimuli. For nonresponders, word production errors changed from superordinate category errors to semantic errors for the atypical trained stimuli, and from circumlocution or inaccurate descriptions to superordinate category errors for the atypical untrained stimuli. Moreover, when both subgroups were compared to the NH group who also demonstrated some improvements of the control-trained items, we found a larger magnitude of gains in the treatment responder group. This suggests that typicality-based SFA treatment has greater potential to improve the overall lexical access in participants who responded favorably to intervention. The unexpected improvements for the control-trained items in the NH group could be attributed to stimuli-related factors, such as word frequency and familiarity. These items may have been encountered more frequently in daily life, which may have resulted in improvements on them during the no-treatment period.

According to the previous study (Gilmore et al., 2018), the differences in naming gains between responders and nonresponders could be associated with pre-treatment cognitive-linguistic abilities, since participants with stronger baseline language and/or cognitive skills responded more favorably to treatment. Our study confirmed this hypothesis, as we found significant differences in the PALPA51 score between responders and nonresponders, Mann-Whitney U test, $p < .01$. This finding suggests less semantic access in nonresponders than responders before treatment, as reflected in the production of more superordinate category errors or circumlocution in nonresponders and more mixed or phonemic errors in responders. This pattern aligns with two-step models of lexical access (Dell, 1986; Levelt et al., 1999), which posit that the severity of word production errors depends on the level of lexical access at which disruptions may occur. For example (Figure 5), to accurately retrieve the word "cat", semantic features of a cat are activated at the semantic level, leading to selection of the most highly-activated lemma and phoneme. Our findings indicate that treatment responders might have disruptions at later stages of lexical access (i.e., phonologic level) given their

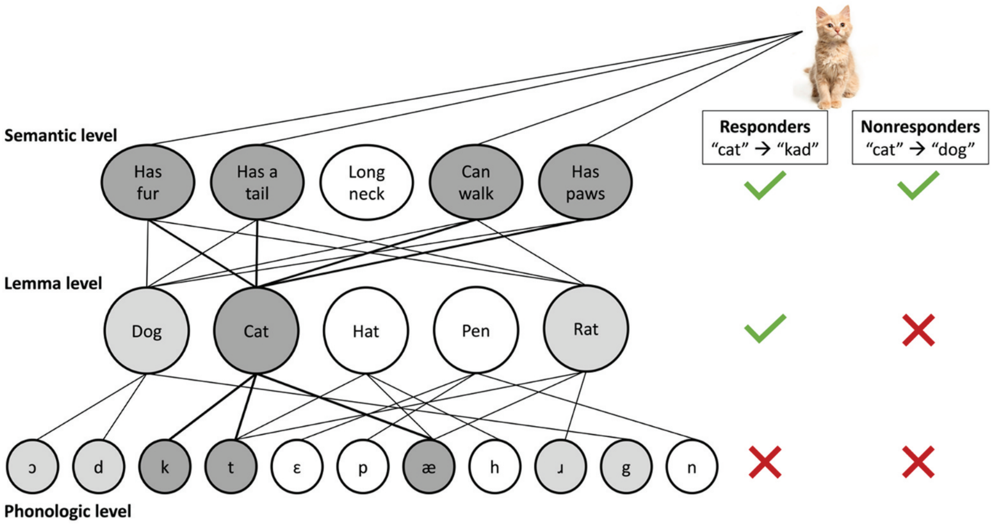

Figure 5. *Possible breakdowns during lexical access in individuals who responded favorably to treatment (responders) and those who did not (nonresponders). Note.* to accurately produce the word "cat", semantic features associated with this object are activated, which then spread activation to the lemma level that includes word forms. Finally, the corresponding phonemes of the word "cat" are activated and produced. Disruptions (red X's) at later stages of lexical access (i.e., phonologic level) may possibly occur in treatment responders due to their better lexical-semantic abilities at baseline, so their word production errors were mostly phonemic errors (e.g., "kad") or mixed errors (e.g., "rat"). In comparison, breakdowns may possibly occur in earlier stages (i.e., semantic or lemma level) in treatment nonresponders due to poorer lexical-semantic abilities at baseline, leading to more severe word production errors, such as superordinate category errors (e.g., "animal") or semantic errors (e.g., "dog").

less severe lexical-semantic deficits at baseline. In comparison, treatment nonresponders might have breakdowns at earlier stages of lexical access (i.e., semantic/conceptual level) due to more severe lexical-semantic deficits. After treatment, we observed a decrease of mixed/phonemic errors in responders and circumlocution/superordinate category errors in nonresponders, suggesting that all participants improved as a function of typicality-based SFA treatment that strengthened both semantic and phonological access.

Generalization effect of typicality

The second research question of the current study tested the CATE applied to semantic complexity (Kiran, 2008; Kiran & Thompson, 2003; Thompson et al., 2003) by investigating if word production error patterns changed for the typical untrained (vs. atypical untrained) stimuli over time. While findings from the whole TX group did not reveal a significant session*typical untrained interaction effect, the follow-up analysis in the responder group showed a significant interaction effect, in support of the application of the CATE to semantic complexity. Since atypical items carry a wider range of semantic features than typical items, training semantic features of atypical items may promote lexical access to untrained typical items through activation of features that are shared, overlapping across items.

Interestingly, we found an opposite pattern in treatment nonresponders (i.e., training the semantic features of typical items promoted generalization to atypical untrained items). Gilmore et al. reported a similar pattern of treatment generalization in the non-responder group when using a binary scoring system. The differences in generalization patterns between responders and nonresponders could be attributed to pre-treatment lexical-semantic processing (i.e., a significant difference in the PALPA51 scores). Treatment responders demonstrated better lexical-semantic processing that may have facilitated their feature analysis during treatment. Thus, reinforcing a variety of semantic features of atypical items (i.e., distinctive, core, and prototypical features) should facilitate the lexical access of the typical untrained stimuli. In contrast, treatment nonresponders tended to exhibit poorer lexical-semantic ability and thus, may have had more difficulties in processing semantic attributes associated with atypical items. In comparison to features of atypical item, semantic features of typical items are central features of a semantic category and more accessible. Therefore, treatment nonresponders appeared to be benefitted from training the features of typical items to promote better overall lexical retrieval (Collins & Loftus, 1975). These findings suggest that individuals with lower cognitive and linguistic skills at baseline may benefit from receiving a generic SFA treatment to improve their overall lexical-semantic abilities before receiving a typicality-based SFA to foster generalization from atypical items to untrained typical items.

Clinical implications

The current study provides evidence of treatment effect following a typicality-based SFA treatment using a fine-grained error coding system. This approach allows us to directly examine the mechanisms underlying semantic-based naming treatment (Boyle et al., 2022). While the commonly used binary accuracy scoring was previously successful in capturing treatment and generalization effects in this sample of PWA (Gilmore et al., 2018), a comprehensive error analysis provided insight into the linguistic mechanisms that may have promoted aphasia recovery. It is important for clinicians to consider individual variations in the response to treatment, as the evolution pattern of word production errors appears to differ between individuals who respond more favorably versus less favorably to typicality-based SFA treatment. Such individual heterogeneity may be attributed to differences in baseline lexical-semantic processing and other general cognitive abilities across participants, highlighting the importance of thoroughly assessing these domains before intervention and considering these abilities when selecting treatment targets for a given individual. Furthermore, clinicians may consider providing a generic SFA treatment to individuals with lower baseline cognitive and linguistic skills before administering a typicality-based SFA to promote the generalization from training atypical items to untrained typical items.

Limitations and future directions

The current study included a diverse participant sample that varied by age, aphasia type, and baseline lexical-semantic abilities and was relatively modest in sample size, both of which made it difficult to further examine which individual factors may be associated with different evolution patterns of word production errors. Due to recruitment challenges, we included eight non-native English speakers who had enough English proficiency to participate in the study. This could lead to potential second-language influence on the treatment outcomes. Hence, future research should include a larger sample size of native English speakers and different subgroups of participants in terms of their aphasia type, lexical-semantic abilities, and general cognitive processing. Additionally, we did not conduct a reliability check for the coding process; however, the coded results were cross-checked by the first and second authors, and any discrepancies were resolved to 100% agreement. The coding was completed without being blinded to session or group, which could increase the potential for bias in the results. Future studies should use blinded coding and conduct reliability check. Moreover, although our error coding scale accounted for a variety of word production errors that are likely to occur in a semantic-based naming therapy, it is unknown whether the same scale would be effective in capturing treatment and generalization effects from other types of intervention (i.e., Phonological Components Analysis; Leonard et al., 2008). Therefore, it is important for future research to replicate our findings in different treatment studies to establish the reliability of this coding scale.

Conclusion

The current study examined the evolution of word production errors following a typicality-based SFA treatment in 30 individuals with chronic aphasia. Treatment participants demonstrated significant improvements in both trained and untrained items, with a higher number of on-target or approaching-the-target responses over time. However, no significant generalization from training atypical items to untrained typical items was observed in the treatment group. Follow-up analyses in the treatment responders and nonresponders revealed different evolution patterns of word production errors for the typical untrained stimuli, as responders demonstrated significant gains in naming untrained typical items from training atypical items, whereas nonresponders exhibited improved naming of untrained atypical items from training typical items. These findings highlight 1) the importance of investigating treatment and generalization effects using a fine-grained error coding system as a complement to a binary scoring system; and 2) the presence of different linguistic mechanisms supporting language recovery for individuals with aphasia.

Note

1. Seven of these natural history control patients served as controls before they were enrolled in the treatment group.

Acknowledgements

We would like to thank the individuals with aphasia who participated in this study. We additionally thank members of the Boston University Aphasia Research Laboratory (now Center for Brain Recovery) for their assistance in data collection and contributions to this project.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by National Institute on Deafness and Other Communication Disorders of the National Institutes of Health (USA) under grant [1P50DC012283].

ORCID

Ran Li http://orcid.org/0000-0002-5034-7294

Data availability statement

The data that support the findings of this study are available on request from the corresponding author, SK. The data are not publicly available due to their containing information that could compromise the privacy of research participants.

References

Beeson, P. M., & Robey, R. R. (2006). Evaluating single-subject treatment research: Lessons learned from the aphasia literature. *Neuropsychology Review*, *16*(4), 161–169. https://doi.org/10.1007/s11065-006-9013-7

Boyle, M., & Coelho, C. A. (1995). Application of semantic feature analysis. *American Journal of Speech-Language Pathology*, *4*(4), 94–98. https://doi.org/10.1044/1058-0360.0404.94

Boyle, M., & Coelho, C. A. (2014). Application of semantic feature analysis as a treatment for aphasic dysnomia. *American Journal of Speech-Language Pathology*, *4*(4), 94–98. https://doi.org/10.1044/1058-0360.0404.94

Boyle, M., Gordon, J. K., Harnish, S. M., Kiran, S., Martin, N., Rose, M. L., & Salis, C. (2022). Evaluating cognitive-linguistic approaches to interventions for aphasia within the rehabilitation treatment specification system. *Archives of Physical Medicine & Rehabilitation*, *103*(3), 590–598. https://doi.org/10.1016/j.apmr.2021.07.816

Brady, M. C., Kelly, H., Godwin, J., Enderby, P., & Campbell, P. (2016). Speech and language therapy for aphasia following stroke. *Cochrane Database of Systematic Reviews*, *2016*(6). https://doi.org/10.1002/14651858.CD000425.pub4

Butterworth, B. (1979). Hesitation and the production of verbal paraphasias and neologisms in jargon aphasia. *Brain and Language*, *8*(2), 133–161. https://doi.org/10.1016/0093-934X(79)90046-4

Coelho, C. A., Mchugh, R. E., & Boyle, M. (2000). Semantic feature analysis as a treatment for aphasic dysnomia: A replication. *Aphasiology*, *14*(2), 133–142. https://doi.org/10.1080/026870300401513

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407

Dabul, B. (2000). *Apraxia battery for adults* (2nd ed.). Pro-Ed.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283–321. https://doi.org/10.1037/0033-295X.93.3.283

Dell, G. S., Lawler, E. N., Harris, H. D., & Gordon, J. K. (2004). Models of errors of omission in aphasic naming. *Cognitive Neuropsychology*, *21*(2–4), 125–145. https://doi.org/10.1080/02643290342000320

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, *104*(4), 801–838. https://doi.org/10.1037/0033-295X.104.4.801

Drew, R. L., & Thompson, C. K. (1999). Model-based semantic treatment for naming deficits in aphasia. *Journal of Speech, Language, & Hearing Research*, *42*(4), 972–989. https://doi.org/10.1044/jslhr.4204.972

Gilmore, N., Braun, E. J., & Kiran, S. (2022). Neurology, connectivity, and the nature of aphasia. In I. Papathanasiou & P. Coppens (Eds.), *Aphasia and related neurogenic communication disorders* (pp. 38–55). Jones & Bartlett Publishers.

Gilmore, N., Meier, E. L., Johnson, J. P., & Kiran, S. (2018). Typicality-based semantic treatment for anomia results in multiple levels of generalization. *Neuropsychological Rehabilitation*, *30*(5), 802–828. https://doi.org/10.1080/09602011.2018.1499533

Goodglass, H., & Wingfield, A. (1997). Word-finding deficits in aphasia: Brain-behavior relations and clinical symptomatology. In *Anomia* (pp. 3–27). Academic Press. https://doi.org/10.1016/B978-012289685-9/50002-8

Helm-Estabrooks, N. (2001). *Cognitive linguistic quick test: CLQT*. Psychological Corporation.

Howard, D., & Patterson, K. E. (1992). *The pyramids and palm trees test*.

Johnson, J. P., Meier, E. L., Pan, Y., & Kiran, S. (2019). Treatment-related changes in neural activation vary according to treatment response and extent of spared tissue in patients with chronic aphasia. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *121*, 147–168. https://doi.org/10.1016/j.cortex.2019.08.016

Kaplan, E., Goodglass, H., & Weintraub, S. (2001). *Boston naming test*.

Kay, J., Lesser, R., & Coltheart, M. (1996). Psycholinguistic assessments of language processing in aphasia (PALPA): An introduction. *Aphasiology*, *10*(2), 159–180. https://doi.org/10.1080/02687039608248403

Kendall, D., Edmonds, L., Van Zyl, A., Odendaal, I., Stein, M., & Van der Merwe, A. (2015). What can speech production errors tell us about cross-linguistic processing in bilingual aphasia? Evidence from four English/Afrikaans bilingual individuals with aphasia. *The South African Journal of Communication Disorders*, *62*(1), 1–10. https://doi.org/10.4102/sajcd.v62i1.111

Kertesz, A. (2007). *Western aphasia battery: Revised*. Pearson.

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

Kiran, S., Balachandran, I., & Lucas, J. (2014). The nature of lexical-semantic access in bilingual aphasia. *Behavioural Neurology*, *2014*, 1–18. https://doi.org/10.1155/2014/389565

Kiran, S., & Bassetto, G. (2008, February). Evaluating the effectiveness of semantic-based treatment for naming deficits in aphasia: What works? *Seminars in Speech and Language*, *29*(1), 071–082. https://doi.org/10.1055/s-2008-1061626

Kiran, S., Sandberg, C., & Sebastian, R. (2011). Treatment of category generation and retrieval in aphasia: Effect of typicality of category items. *Journal of Speech, Language, & Hearing Research*, *54*(4), 1101–1117. https://doi.org/10.1044/1092-4388(2010/10-0117)

Kiran, S., & Thompson, C. K. (2003). The role of semantic complexity in treatment of naming deficits: Training semantic categories in fluent aphasia by controlling exemplar typicality. *Journal of Speech, Language, & Hearing Research*, *46*(4), 773–787. https://doi.org/10.1044/1092-4388(2003/061)

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

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(1), 1–38. https://doi.org/10.1017/S0140525X99001776

Minkina, I., Oelke, M., Bislick, L. P., Brookshire, C. E., Hunting Pompon, R., Silkes, J. P., & Kendall, D. L. (2016). An investigation of aphasic naming error evolution following phonomotor treatment. *Aphasiology*, *30*(8), 962–980. https://doi.org/10.1080/02687038.2015.1081139

Quique, Y. M., Evans, W. S., & Dickey, M. W. (2019). Acquisition and generalization responses in aphasia naming treatment: A meta-analysis of semantic feature analysis outcomes. *American Journal of Speech-Language Pathology*, *28*(1S), 230–246. https://doi.org/10.1044/2018_AJSLP-17-0155

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605. https://doi.org/10.1016/0010-0285(75)90024-9

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime reference guide*. Psychology Software Tools, Incorporated.
Schwartz, M. F. (2014). Theoretical analysis of word production deficits in adult aphasia. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634), 1–10. https://doi.org/10.1098/rstb.2012.0390
Thompson, C. K., Shapiro, L. P., Kiran, S., & Sobecks, J. (2003). The role of syntactic complexity in treatment of sentence deficits in agrammatic aphasia. *Journal of Speech, Language, & Hearing Research*, *46*(3), 591–607. https://doi.org/10.1044/1092-4388(2003/047)
Wambaugh, J. L., Mauszycki, S., Cameron, R., Wright, S., & Nesslera, C. (2013). Semantic feature analysis: Incorporating typicality treatment and mediating strategy training to promote generalization. *American Journal of Speech-Language Pathology*, *22*(2), S334–S369. https://doi.org/10.1044/1058-0360(2013/12-0070)
Wiley, R. W., & Rapp, B. (2019). Statistical analysis in small-N designs: Using linear mixed-effects modeling for evaluating intervention effectiveness. *Aphasiology*, *33*(1), 1–30. https://doi.org/10.1080/02687038.2018.1454884

Appendices

Appendix A

Table A1. Description of the error coding scale and examples of errors.

Scale	Response	Description	Examples
0	No response	No response, "I don't know", head shake, unintelligible speech, automatic utterance	Target: *hat* Response: *No, not sure*
1	Neologism	Less than 50% of the word resembles the target	Target: *dresser* Response: *boosleeve*
2	Real word perseveration	Needs to be repeated three times or more.	Target: *shoe* Response: *pop, pop, pop*
3	Unrelated	Word semantically and phonologically unrelated to the target	Target: *couch* Response: *hook*
4	Inaccurate description	Description providing information inaccurate about the target	Target: *helmet* Response: *for a house*
5	Circumlocution	Description providing information accurate about the target	Target: *underpants*, Response: *bathroom, bedroom*
6	Superordinate category error	Response belongs to the superordinate category of the target	Target: *robin* Response: *bird*
7	Semantic error	Semantic substitution or paraphasia	Target: *chair* Response: *sofa*
8	Mixed error	Real word response that qualifies as a semantic error and that meets the criterion for phonological similarity	Target: *skirt* Response: *shirt*
9	Phonemic error/ apraxic	More than one phonemic substitution or omission	Target: *dog* Response: *dock*
10	Dysarthric/ distortion	Dysarthric: slurred/imprecise production; Distortion: when the sound is not left out or substituted but does not sound right. There is an attempt to make the sound but is misarticulated.	
11	Partial target	Part of the response is accurate	Target: *pepper* Response: *green pepper*
12	Target	Accurate response without any phonemic errors	Target: robin Response: robin

Note. Adopted and modified from Kiran et al. (2014). Each increasing number indicates lexical access closer to the target response.