Determinants of Multilevel Discourse Outcomes in Anomia Treatment for Aphasia

by

Robert Benjamin Cavanaugh

B.A. Communication Sciences and Disorders, University of Pittsburgh, 2013

M.S. Speech and Hearing Sciences, University of North Carolina at Chapel Hill, 2015

Submitted to the Graduate Faculty of the

School of Health and Rehabilitation Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2023

UNIVERSITY OF PITTSBURGH

SCHOOL OF HEALTH AND REHABILITATION SCIENCES

This dissertation was presented

by

Robert Benjamin Cavanaugh

It was defended on

April 17, 2023

and approved by

Michael Walsh Dickey, Professor, Communication Sciences and Disorders, School of Health and Rehabilitation Sciences, University of Pittsburgh

William D. Hula, Adjunct Assistant Professor, Communication Sciences and Disorders, School of Health and Rehabilitation Sciences, University of Pittsburgh

Davida Fromm, Special Faculty Researcher, Department of Psychology, Dietrich College of Humanities and Social Sciences, Carnegie Mellon University

Dissertation Director: William S. Evans, Assistant Professor, Communication Sciences and Disorders, School of Health and Rehabilitation Sciences, University of Pittsburgh

Determinants of Multilevel Discourse Outcomes in Anomia Treatment for Aphasia

Robert Benjamin Cavanaugh, M.S. CCC-SLP

University of Pittsburgh, 2023

Communication is fundamental to the human condition but is impaired in life-altering ways for more than 2.4 million individuals with aphasia in the United States. Individuals with aphasia identify discourse-level communication (i.e., language in use) as a high priority for treatment. The central premise of most aphasia treatments is that restoring language at the phoneme, word, and/or sentence level will generalize to discourse. However, treatment-related changes in discourse-level communication are modest, poorly understood, and vary greatly between individuals with aphasia. In response, this study consisted of a multilevel discourse analysis of archival, monologic discourse outcomes across two high-intensity Semantic Feature Analysis clinical trials (combined n = 60). Aim 1 evaluated changes in theoretically motivated discourse outcomes representing lexical-semantic processing, lexical diversity, grammatical complexity, and discourse informativeness across study enrollment, entry, exit, and 1-month follow-up. Aim 2 explored the potential moderating role of non-language cognitive factors (semantic memory, divided attention, and executive function) on discourse outcomes in a subsample of participants (n = 44). The present study found no evidence for meaningful or statistically reliable improvements in monologue discourse performance after Semantic Feature Analysis. There was weak and inconsistent evidence that non-language cognitive factors may play a role in moderating treatment response. These findings underscore the need to intentionally link treatment mechanisms, discourse elicitation tasks, and outcome measures. Furthermore, there is a clear need to examine how established treatments, restorative or compensatory, can

better facilitate generalization to discourse-level communication. These priorities are critical for meaningfully improving everyday communication and reducing the profound communication and psychosocial consequences of aphasia.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CAT	Comprehensive Aphasia Test
CI	Credible Interval
CIU	Correct Information Unit
COR	Correlation
LKJ	Lewandowski, Kurowicka, and Joe (Prior)
LUNA	Linguistic Underpinnings of Narrative in Aphasia
MATTR	Moving Average Type-Token Ratio
MLU	Mean Length of Utterance
NARNIA	Novel Approach to Real-life communication: Narrative Intervention in Aphasia
PD	Probability of Direction
PPT	Pyramids and Palm Trees
QPA	Quantitative Analysis of Agrammatic Production
SD	Standard Deviation
SFA	Semantic Feature Analysis
SFA-1	Semantic Feature Analysis clinical trial #1 (n=44; 5I01RX000832-05)
SFA-2	Semantic Feature Analysis clinical trial #2 (n =16; 1R01DC017475-01)
TEA	Test of Everyday Attention
WCST	Wisconson Card Sorting Test

Preface

Acknowledgements

I fear that if I tried make an exhaustive list of everyone who has supported me thus far, I'd still be writing. So, here's the abridged version:

I'm infinitely grateful that the past five years (and those preceding) have been filled with friendship, laughs, shared adversity, and collective accomplishment more than anything else. To Amanda, for being my emotional and intellectual partner, for saying yes to this adventure and refusing to let me give up, and for encouraging whatever my most recent obsession is. I can't imagine a world where I've made it this far without you. To Will, for striving to have your priorities in the right order and setting the expectation from day one that I do the same. Also, for teaching me the rules of the game, how to play it with high standards, and for the right reasons. To my committee and many mentors, for selflessly giving your time to help me navigate research, career, and life challenges. To the aphasia research community and my clinical colleagues, for being welcoming and inclusive, for your insatiable aspiration to peel back the next layer of understanding, and your passion for trying to make a difference. To the Pitt PhD CSD cohort, for being so genuine and supportive, without hesitation, in the midst of a global pandemic. To Mom and Dad and Katie, my family, and friends, for being the foundation on which everything else is built. And to Murphy and Willa, for the unlimited fluff therapy.

Funding Sources

This research was supported by the following awards, for which I am immensely grateful: VA Rehabilitation Research and Development Award I01RX000832 awarded to Michael Walsh Dickey and Patrick J. Doyle. National Institute on Deafness and Other Communication Disorders 5R01DC017475-04 awarded to William D. Hula and Michael Walsh Dickey. National Center for Advancing Translational Sciences NRSA TL1TR001858 awarded to Robert Cavanaugh. National Institute on Deafness and Other Communication Disorders NRSA F31DC019853-01 awarded to Robert Cavanaugh. The Audrey Holland Endowed Student Research Award awarded to Robert Cavanaugh. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

1.0 Introduction

Communication is fundamental to the human condition but is impaired in life-altering ways for more than 2.4 million people with aphasia in the US (Simmons-Mackie, 2018). Aphasia affects one-third of stroke survivors, with more than 180,000 new cases annually (Pedersen et al., 2004). Aphasia has a profound impact on health-related quality of life, including greater rates of social isolation and depression compared to stroke survivors without aphasia (Hilari, 2011).

Impairment-focused clinical interventions for aphasia aim to ameliorate underlying linguistic deficits, with the intention that improvements will generalize (Thompson, 2006) to discourse, or language use above the sentence-level. However, many evidence-based aphasia interventions lack a clear theoretical basis to describe how treatment effects are expected to generalize to discourse (Dipper et al., 2020; Webster et al., 2015). The absence of clearly specified treatment theory may contribute to the modest, poorly understood, and highly variable discourse outcomes in aphasia interventions to date (Dipper et al., 2020; Webster et al., 2015). Improving our understanding of discourse generalization in aphasia is a critical step in ongoing efforts to improve everyday communication outcomes (Stark et al., 2020), which may help to reduce the profound psychosocial consequences of aphasia. The aims of this study are to (1) evaluate the contributions of key mechanisms affecting discourse outcomes in Semantic Feature Analysis (SFA) treatment for aphasia through multilevel discourse analysis and (2) explore potential non-language cognitive processes which support discourse outcomes in aphasia. The broader purpose of the study is to support the larger community-wide effort seeking to improve meaningful communication outcomes in aphasia by developing and refining theoretically based interventions.

The following introductory sections will introduce the theoretical basis and underlying evidence for SFA as a behavioral treatment for aphasia. I will review the evidence for the effects of SFA on discourse production and describe preliminary data from a completed SFA clinical trial which was the impetus for the present study. Finally, I will introduce the potential benefits of viewing SFA's impact on discourse from the lens of multilevel discourse theory which form the basis for the aims and hypotheses of the present study.

1.1 Semantic Feature Analysis Treatment for Aphasia

Many treatments for aphasia target anomia (word-finding deficits) under the premise that improving anomia will improve communication function at the discourse level. Anomia is the cardinal deficit in aphasia (Goodglass, 1980) and a common target of intervention studies (Brady et al., 2016) and the clinical management of aphasia (Brogan et al., 2020). Semantic Feature Analysis (SFA) is one of the most common anomia treatments in clinical research (e.g., Boyle, 2010; Efstratiadou et al., 2018; Gravier et al., 2018; D. L. Kendall et al., 2019; Quique et al., 2019). SFA is comprised of interleaved, effortful retrieval of target words and semantic features (Boyle & Coelho, 1995) in several semantic categories, typically superordinate, use, action, properties, location, and personal association.

The foundational mechanism of action in SFA is based on a spreading activation hypothesis of semantic processing (Collins & Loftus, 1975). In this view, elicitation and production of semantic features spreads activation of the features within the semantic network to their associated concepts and ultimately to associated lexical items. The extent of activation for a given lexical item stems from the activation of many semantic features or a few distinguishing features (Boyle, 2010). The lexical item with the highest activation is selected for production.

A more recent interpretation of SFA's restorative mechanism of action is that repeated production of target words and semantically related features strengthens the connections between conceptual and lexical representations, consistent with the two-step interactive activation model of lexical access (Dell, 1986; Foygel & Dell, 2000). Reinforcement of these connections is thought to spread activation between semantically related items in the lexical-semantic network, via feedback from activated (target) lexical representations to associated conceptual representations, which in turn engenders feedforward activation from those conceptual representations to other semantically related lexical items. Similarly, it may also be that repeated feature generation and naming of target items improves the resting activation for both the target item and other items within a semantic category. In either case, the core prediction of these restorative hypotheses is that SFA will improve lexical access to both trained and semantically related, untrained words via repeated activation. If this hypothesized mechanism of action is indeed behind SFA's benefits, SFA need only focus on a limited subset of words to improve lexical access to a broader range of semantically related words, resulting in widespread improvements in word finding.

Several reviews and meta-analyses have evaluated the effects of SFA on naming of both treated and semantically related, untreated words. While these reviews have employed different methods, many have evaluated overlapping datasets from the same treatment studies and have generally arrived at the same conclusion. Boyle (2010) found that 16 out of 17 individuals with aphasia improved on naming of treated words after SFA, and 13 of the 17 demonstrated some evidence of generalization to semantically related words. Reviews by Oh et al. (2016) and

Efstratiadou et al. (2018) reported similar conclusions. In a meta-analysis of single-subject design studies which comprised the majority of the SFA literature at the time, Quique et al. (2019) found that SFA improved performance on both treated and semantically related, untreated words across studies, with effect sizes greater for treated words. Quique et al. (2019) also found that treatment effects were moderated by treatment dosage and pre-treatment language impairment indices. Larger group-level studies have largely corroborated these findings (Evans, Cavanaugh, Gravier, et al., 2021; D. L. Kendall et al., 2019), though the question of whether SFA has clear and robust effects on semantically related but untrained words is still outstanding (see Nickels, 2002).

Recent work on learning and retrieval practice in aphasia, in both effortful and errorless contexts, provides a potential complementary restorative mechanism of action in SFA. Middleton and colleagues (2016) have demonstrated that repeated retrieval practice, whether massed or distributed, effortful or errorless, can improve naming performance on trained words. Retrieval practice is thought to strengthen both lexical-semantic and semantic-phonological connections, improving word retrieval specifically for trained words. Thus, improvements in naming treated words in SFA may result from successful effortful practice during the target-naming component. This mechanism would also suggest that repeated feature generation should improve an individual's ability to retrieve features in forms similar to how they are generated during treatment (e.g., retrieval of the phrase "used for cutting" repeatedly generated as a function/use semantic feature for the target item "knife."). Under this premise, strengthened lexical-semantic and semantic-phonological connections may improve access not only to the limited treated word set, but also the vast number and diversity of features generated throughout treatment, which often include short phrases within verb-object or prepositional structures.

Improvement to semantically related words may occur under this mechanism, if they are incidentally included in feature generation (SFA does not constrain the features that participants are allowed to produce, though studies ideally take care not to train semantically related words being probed as features).

Moreover, there is evidence that SFA may also operate via a compensatory mechanism, teaching self-cueing and strategic, compensatory responses to instances of anomia. Antonucci (2009) argued that SFA "promotes habituation of semantic self-cueing and semantically appropriate circumlocution, strategies that facilitate communication even if specific lexical retrieval fails" (p. 855). This compensatory mechanism suggests that, even when lexical retrieval fails, SFA may help individuals with aphasia convey their intended message via retrieval and production of relevant semantically related content (Falconer & Antonucci, 2012). Retrieval of semantically related content may also help individuals with aphasia navigate to their intended lexical item.

Two studies have incorporated explicit strategy training to the traditional SFA treatment, hoping to take advantage of this potential compensatory mechanism. Wambaugh et al. (2013) demonstrated improvements to trained words following the implementation of an explicit self-cueing mediating strategy training phase in SFA, but no gains were seen on generalization items. In this case, strategy training might either be limited to treated items or more relevant to less structured language tasks such as discourse production (though notably, 6/8 participants receiving the strategy training had co-morbid apraxia of speech which may have contributed to limited generalization findings). Tilton-Bolowsky et al., (2022) added a metacognitive component focused on the use of self-cueing and circumlocution in SFA and found that all four participants increased their awareness and use of strategic communication after SFA. However, the impact of this additional component on word retrieval, particularly for untreated words, was not clear. In summary, there is evidence of a complementary, compensatory mechanism supporting SFA that may improve language use instead of, or in addition to, restoration of the underlying word-finding deficits.

1.2 Effects of Semantic Feature Analysis on Discourse Production

The candidate restorative and compensatory mechanisms have clear implications for improvements to word finding deficits in the trained context (i.e., confrontation picture naming). But how might these mechanisms engender improvements in discourse production? Assuming a mechanism framed under the two-step interactive activation model, restoration of access to targets and a range of semantically related words within the semantic network might be sufficient to produce semantically diffuse gains relevant to general-topic discourse stimuli. In other words, SFA might improve word finding for many trained and related but untrained words across multiple semantic categories, and these gains could be relevant to discourse production even when the topic of discourse is not explicitly trained during SFA. Similarly, a retrieval-practice restorative mechanism in SFA might improve discourse production via improving access to a wide array of targets and semantic features regardless of their semantic relationships or the semantic similarities between the trained categories and discourse topic. Contrary to either restorative mechanism, generalization from a compensatory mechanism is not theoretically limited to trained or semantically related words. Instead, a compensatory account of SFA improves individuals' potential to strategically adapt when they experience an instance of anomia irrespective of any underlying improvements to the language system or the topic of discourse at hand.

There are two primary measurement paradigms for examining discourse outcomes in SFA. The *status quo* of discourse outcome measurement paradgims in aphasia, and by extension SFA, uses “general topic” picture stimuli (most often, Nicholas & Brookshire, 1993) to elicit discourse monologue (Bryant et al., 2016). In the context of SFA and other anomia studies, this paradigm is a test of “far generalization” of treatment effects to connected speech contexts which are not explicitly related to treatment targets. Thus improvements must be general enough to be elicited by general-topic stimuli. A less common alternative is to use discourse stimuli to develop treatment targets, which would be more consistent with a near-transfer paradigm. The evidence for SFA on discourse production in the context of both approaches are discussed below.

In general, the evidence for generalization of SFA’s treatment effects to discourse production is inconsistent at best. In one of the largest published SFA studies to date, Silkes and colleagues (2021) found that SFA improved discourse informativeness by a non-significant 3 percentage points at treatment exit on a general-topic story retell task. Across 30 individuals who received SFA in the study, there was substantial variability in SFA outcomes, with a non-trivial subset (roughly one-third) of participants demonstrating positive change after treatment. However, Silkes et al., (2021) did not evaluate the statistical significance of individual-level change.

The finding that SFA had weak group-level effects on discourse production and may have only produced robust change for a subset of participants is consistent with single subject and small-N SFA studies. In the three early single-subject design studies on SFA, only one out of four participants demonstrated meaningful improvements in discourse informativeness (Boyle, 2004; Boyle & Coelho, 1995; Coelho et al., 2000) on the Nicholas and Brookshire protocol. To directly measure generalization of trained words after SFA, Rider et al. (2008) trained lists of words drawn from 3 participants' summaries of short TV sit-com clips and procedural discourse samples ("How to make a Peanut Butter and Jelly Sandwich"). Rider et al. (2008) found that participants marginally increased the number of trained words produced during discourse samples (4-6 additional words). However, no changes were seen on a general measure of word finding in discourse (lexical diversity). Additionally, multiple target words produced at baseline were not produced after treatment – demonstrating the challenges imposed by the non-obligatory nature of discourse elicitation tasks. Peach & Reuter (2010) employed a similar approach, training nouns and verbs that 2 participants were unable to retrieve during picture description or procedural discourse tasks at the start of each treatment session. Like Rider et al. (2008), discourse outcomes were highly varied and demonstrated only weakly positive trends. Only Antonucci (2009) and Falconer & Antonucci (2012) have reported consistent improvements at the discourse level using a version of the standard Nicholas and Brookshire (1993) protocol. Both studies implemented SFA in group settings and focused on circumlocution and self-cueing. Six out of seven participants demonstrated improvements in discourse informativeness, discourse efficiency, or both.

1.3 Preliminary Data

In a preliminary analysis of a recently completed SFA clinical trial (n = 44; Evans, Cavanaugh, Gravier, et al., 2021; Gravier et al., 2018), individuals with aphasia demonstrated improvements in discourse informativeness, the percentage of accurate, relevant words to total words produced on general-topic, monologue-based discourse tasks (Nicholas & Brookshire, 1993) after SFA (Figure 1: posterior probability entry to exit: >99%; entry to follow-up: >99%). This change corresponded to a model-estimated average gain of 5 percentage points at exit (42.5% to 47.5% CIUs) and 7 percentage points at follow-up (42.5% to 49.5% CIUs). The change was not moderated by aphasia severity (CAT mean T-score; posterior probability <60%) but was highly variable. Model estimates of adjusted individual effects suggested that 19/44 participants exhibited improvements (posterior probability of positive change >90%) at exit and/or follow-up.

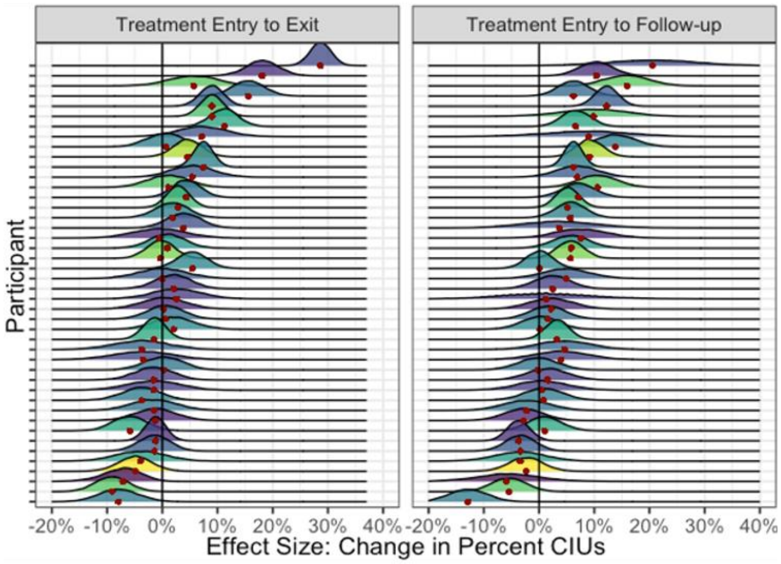


Figure 1. Preliminary effects of SFA on discourse informativeness in 44 individuals with aphasia Vertical line indicates no change. Aphasia severity is represented by color from severe (dark/purple) to mild (light/yellow)

There is reason to believe that this SFA clinical trial may be more likely to engender improvements in discourse production compared to aforementioned SFA studies. This study included a sentence generation component, which required participants to generate a sentence using the target word and two other elements with semantically rich content (verb; subject or object), typically involving semantic features generated during the trial, with cues as needed from the clinician. This component was included to increase the likelihood of generalization outside of picture naming by providing an opportunity to practice at the sentence level. Other aphasia interventions have used similar “loose” and generative training approaches, expanding on word- and phrase-level production to foster generalization to discourse production, with some success (e.g., Wambaugh & Martinez, 2000). Eliciting language that is more closely aligned with general communicative contexts, with greater variability in patient-generated responses and appropriate modeling and shaping by a clinician may foster generalization to more general contexts. It is plausible that the addition of this sentence-level component, combined with the substantial dosage of both SFA trials, could increase the likelihood of generalization in this specific instance of SFA.

However, informativeness can be influenced by restoring word-retrieval ability and/or improving compensatory communication strategy use. A restorative mechanism that improves underlying lexical retrieval ability might enable participants to retrieve and produce more words that are both correct and more relevant to the discourse monologue context. Yet improvements in compensatory communication strategy use might achieve a similar result through overt or covert self-cueing and circumlocution. As a result, examining changes only in discourse informativeness in these data does little to clarify the degree to which either mechanism might be responsible for discourse gains after SFA.

1.4 Multilevel Discourse Theory and Treatment for Aphasia

Multilevel discourse theory provides a much-needed theoretical framework for describing how aphasia treatments might generalize to discourse production. Multilevel discourse analysis is the integrated analysis of discourse microstructure (within-sentence lexical and syntactic features that support connected speech) and macrostructure (between-sentence features that promote overall effectiveness) (Harris Wright & Capilouto, 2012; Marini et al., 2011; Sherratt, 2007). Marini et al. (2011) demonstrated a method by which multilevel discourse analysis can inform treatment mechanisms using two contrasting case studies. In the first, they demonstrated that a reduction in paraphasias was associated with resolution of grammatical deficits and improvements in discourse informativeness, suggesting that intervention improved underlying linguistic structures which propagated to higher-order discourse levels. This claim underlies the motivation for restorative anomia treatments such as SFA – improving underlying word retrieval ability will translate to broader improvements in communication across discourse levels. In the second case, an individual with aphasia improved in informativeness and coherence through a compensatory-focused treatment, which did not improve word-level errors or syntactic structures, consistent with a compensatory mechanism.

These claims are based on a multilevel processing model (Frederiksen et al., 1990), which conceives discourse production as a bidirectional, interactive process, from activating conceptual representations to formulating language, supported by necessary cognitive functions (Sherratt, 2007). This model is supported by established relationships between discourse micro- and macrostructure. Sherratt (2007) provided initial support for this model by demonstrating the interrelatedness of different levels of discourse processing, from discourse measures tied to frame/schema generation to linguistic formulation (Figure 2). Marini et al. (2011) and Andreetta et al. (2012) both interpreted correlations between measures of lexical impairment and higher-order discourse measures (e.g. in terms of utterance-completeness, cohesion, and coherence) as evidence that lexical impairment may be responsible for impairments in these higher-order measures. Similarly, Harris Wright & Capilouto (2012) found strong correlations between lexical diversity, grammatical complexity, informativeness, and global coherence. Recent factor analytic work by Gordon (2020) reported that 17 micro-linguistic discourse measures could be condensed into 6 latent factors comprising phrase building, narrative productivity, semantic anomaly, grammatical error, grammatical complexity, and repair. These factors were often intercorrelated, sharing as high as 34% of variance. Together, these findings provide strong, converging evidence for the interrelatedness of different levels of discourse production and the benefits of examining discourse across micro- and macro-structural levels.

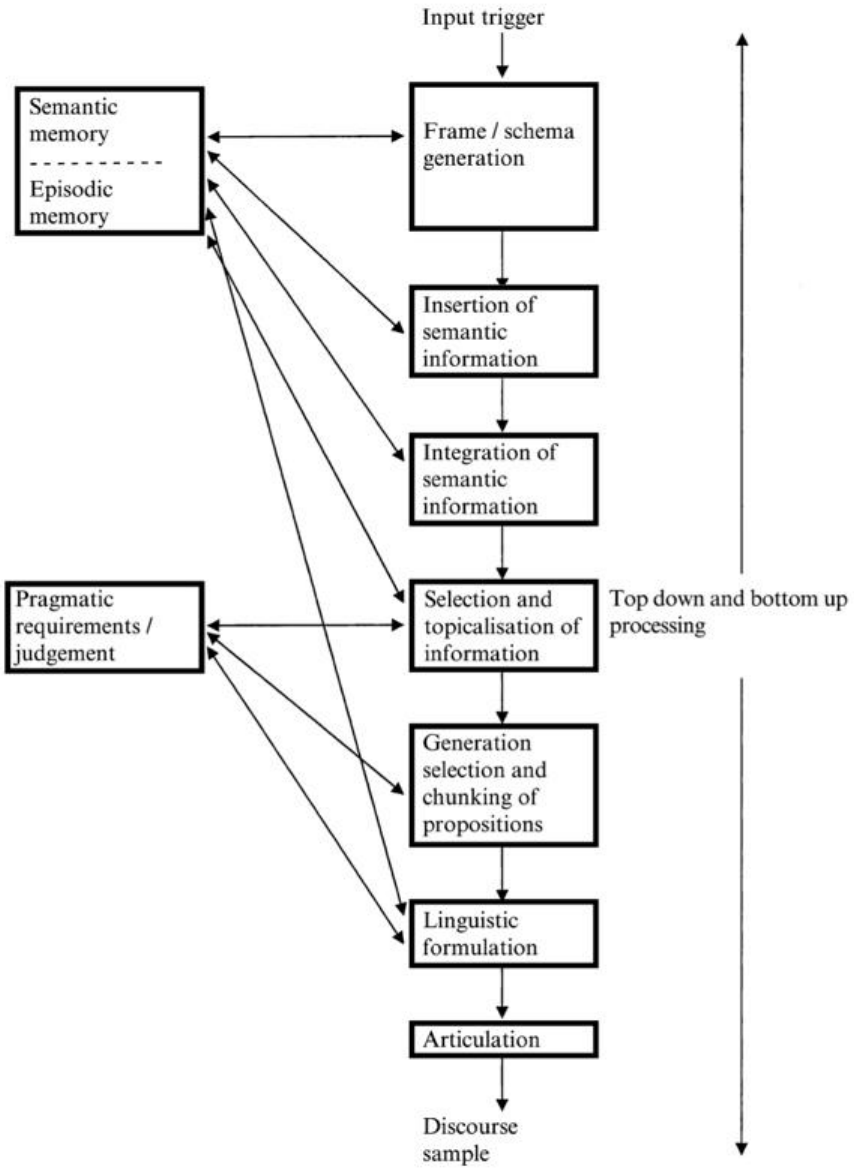

Input trigger
Semantic memory
Episodic memory
Frame / schema generation
Insertion of semantic information
Integration of semantic information
Pragmatic requirements / judgement
Selection and topicalisation of information
Top down and bottom up processing
Generation selection and chunking of propositions
Linguistic formulation
Articulation
Discourse sample

Figure 2. Schema Multilevel Discourse Framework

Sheratt, 2007; Reprinted with permission

There is emerging evidence for the utility of multilevel discourse analysis in characterizing discourse outcomes in aphasia treatment studies. Whitworth et al. (2015) reported the results of a pilot treatment NARNIA (a Novel Approach to Real-life communication: Narrative Intervention in Aphasia). NARNIA combined a verb mapping-style treatment focused on discourse microstructure (e.g., lexical retrieval for agents, verbs, and patients and sentence building with specific argument structure) and narrative instruction focused on macrostructure (e.g., practice using cohesive ties). Improvements after NARNIA were seen across verb usage, argument structure, and orientation elements within discourse, suggesting that the treatment may have generalized to broader narrative discourse production. However, the relatively small sample size (n = 8 participants received NARNIA) limits strong conclusions about treatment effectiveness. Furthermore, Whitworth et al. (2015) did not report whether the same participants were improving across measures, thus it is not clear whether improvements at the linguistic level led to improvements in propositional or macro-structural discourse, as predicted by the multilevel theory laid out by Sherratt (2007).

A more recent approach, Linguistic Underpinnings of Narrative in Aphasia (LUNA), has been piloted by Cruice, Dipper, and colleagues (2020). LUNA views the relationship between linguistic components using a "Russian doll" analogy where each lower-level process is nested within the adjacent higher level. Each level has the potential to influence others via "upwards" or "downwards" feedback. This framework predicts that failures in linguistic processing (e.g., lexical retrieval) may derail the macro-structure or pragmatic component of discourse production – proving further rationale for how treatments such as SFA might improve discourse production. Predictions from the LUNA framework are similarly consistent with that of Sherratt (2007), Marini et al. (2011), and Whitworth (2015): resolution of underlying linguistic deficits (in the linguistic component) may produce "knock-on" improvements at higher levels of language. An unpublished pilot treatment approach based on the LUNA framework has found emerging evidence for the utility of applying this multilevel framework to treatments for aphasia, but also advised cautious interpretation due to lack of statistical power (Dipper et al., 2022).

Despite preliminary evidence for the utility of multilevel frameworks in aphasia, the evidence for a strong interpretation of multilevel discourse outcomes in aphasia is limited. Marini et al. (2011) evaluated 14 total discourse measures, with a substantial risk of overinterpreting false positives in noisy data. Though Whitworth et al. (2015) included a total of 14 individuals with aphasia (8 received NARNIA), the sample size is unlikely to be powered for either between or within group changes across multiple discourse outcome measures. The present claim is that multilevel analysis of discourse outcomes in a large SFA trial provides a unique opportunity and tractable basis for 1) characterizing the linguistic context surrounding gains in discourse informativeness in preliminary data and more broadly potential mechanisms supporting discourse generalization in SFA and 2) validating the claims laid out by Marini et al. (2011) in a large sample of individuals with aphasia.

1.5 Evaluating Monologue Discourse Outcomes in Semantic Feature Analysis Treatment for Aphasia (Aim 1)

Examining patterns of correspondence and non-correspondence across theoretically motivated outcomes has the potential to characterize mechanisms supporting discourse outcomes in SFA (Aim 1). This study sought to evaluate four components of discourse production which are interrelated, can be conceptualized within multilevel discourse theory (Sherratt, 2007), and tied to the hypothesized mechanisms in SFA. These components (measures) are lexical-semantic processing (% semantic paraphasias), lexical diversity or word retrieval at the discourse-level (moving average type-token ratio), grammatical complexity (mean utterance length), and discourse informativeness (%CIUs). While gains in informativeness and lexical diversity are plausible with either the restorative or compensatory mechanism, gains in lexical-semantic processing, and grammatical complexity are aligned primarily with a restorative mechanism. This multilevel approach will better characterize preliminary findings which show modest improvements in discourse informativeness in a large sample of individuals with aphasia. Under this premise, this proposal laid out the following hypotheses:

(1) If the preliminary findings in discourse informativeness are driven by the restorative mechanism, SFA should engender correlated improvements across lexical-semantic processing, lexical diversity, grammatical complexity, and informativeness. Multilevel discourse theory suggests that the resolution of lexical-semantic processing deficits should reduce the frequency of semantic errors and increase the diversity of words produced, in turn allowing participants to increase utterance length and complexity and produce overall more informative responses to discourse stimuli. This proposition is based on (a) the known effects of SFA; (b) the extensive number and diversity of semantic features (composed of verbs, adjectives, and nouns in single words or short phrases) produced during 60 hours of treatment over 4 weeks, which may vastly extend the semantic space for spreading activation and consequently the number of words and phrases accessible to individuals with aphasia during discourse tasks; and (c) the presence of an additional sentence-level component designed to engender generalization beyond the word level.

(2) Alternatively, discourse improvements limited only to informativeness and lexical diversity, but not lexical-semantic processing, and/or grammatical complexity would be more consistent with a compensatory account of discourse generalization in SFA (Antonucci, 2009; Falconer & Antonucci, 2012). In other words, participants are increasing the rate of producing correct and relevant words in their monologues, but this increase is not tied to a reduction in semantic errors or changes in utterance length or complexity. As Marini (2011) noted, this

pattern is better aligned with the deployment of compensatory strategies, but minimal resolution of the underlying impairment. Such a finding might encourage treatment modifications to facilitate strategic communication use in SFA but precludes a strong account of the restorative mechanism as responsible for improvements in general monologue discourse production.

(3) Alternatively, uncorrelated improvements across measures would not support the standardly assumed restorative motivation for anomia treatments: that resolution of word-finding deficits will engender more effective discourse-level communication. This finding would primarily suggest that the mechanism supporting discourse generalization in SFA largely depends on individual characteristics, which is the focus of Aim 2.

1.6 Examining Cognitive-linguistic Moderators of Monologue Discourse Outcomes in Semantic Feature Analysis Treatment for Aphasia (Aim 2)

Key cognitive processes may be necessary for generalization to discourse-level communication in aphasia. The multilevel processing model described by Sherratt (2007) suggests that (1) semantic memory supports the process of discourse production from frame/schema generation to linguistic formulation, whereas judgment processes (2) attention and (3) executive function support macrostructural-type processes such as selection of information, chunking of propositions, and linguistic formulation (Pritchard et al., 2018; Sherratt, 2007). Other discourse production frameworks have hypothesized similar relationships between non-linguistic cognitive abilities and discourse production. For example, the LUNA model (Dipper et al., 2021) posits that cognitive skills are relevant to all aspects of discourse production, emphasizing the role of working memory, episodic memory, and executive function in macrostructure planning and organization, sequencing, and attention in the propositional aspects of spoken discourse.

Accordingly, deficits in semantic memory, attention, and executive function may moderate discourse outcomes in language-based interventions like SFA. Although theoretical discourse frameworks have long hypothesized strong relationships between cognitive abilities and discourse production (Dipper et al., 2021; Frederiksen et al., 1990), there is limited empirical evidence for how these factors influence discourse production in aphasia. Current evidence suggests that abilities such as working memory construct an important foundation for discourse production (Cahana-Amitay & Jenkins, 2018).

Provided that cognitive-linguistic skills are critical for discourse production cross-sectionally, it follows that they may also influence response to treatment as well. For example, these cognitive skills are associated with both direct training and stimulus generalization effects in anomia treatments, (Dickey et al., 2016; Lambon Ralph et al., 2010; Simic et al., 2020) which has been attributed to their importance in human learning processes (Meyer et al., 2007; Villard & Kiran, 2017). If these cognitive deficits constrain improvement on word-level tasks, there may be fewer improvements available to generalize to the discourse level. Furthermore, executive function abilities, which reflect a wide range of skills including problem solving, cognitive flexibility, and self-monitoring, are associated with functional communication (Olsson et al., 2019; Ramsberger, 2005) and specifically successful compensatory strategy use (Dean et al., 2017; Purdy & Koch, 2006), which are central elements of the compensatory mechanism in SFA.

Of the studies which have demonstrated links between cognitive factors and treatment outcomes in aphasia, none have evaluated associations with discourse outcomes at the group level. Examining the relationships between each cognitive factor and discourse outcomes will establish how each cognitive process might differentially support restorative and/or compensatory mechanisms and change at multiple levels of discourse. This finding will enable future latent-factor predictive models of discourse-oriented treatment response, thereby improving our knowledge of treatment candidacy and the cognitive processes which support discourse-level communication outcomes.

Because no studies have evaluated cognitive predictors of discourse outcomes, aim 2 explored whether such relationships might exist in the context of SFA and multilevel discourse analysis. The following broad patterns of results are plausible, given multilevel models of discourse described above and the cognitive capacities assumed to support them: (1) Associations between a given cognitive predictor and restorative-focused discourse outcomes (lexical-semantic processing, grammatical complexity) would provide evidence for that factor's role in the restorative treatment mechanism, consistent with established associations between cognitive factors and naming outcomes in aphasia (Gilmore et al., 2019; Lambon Ralph et al., 2010; Simic et al., 2020) (2) Associations between cognitive predictors and lexical diversity and informativeness would reflect that factor's role in restorative and/or compensatory mechanisms; identifying relationships between cognitive processes and restorative vs. compensatory treatment mechanisms requires experimental manipulation absent in this study, but these findings would support future work in this area. (3) Examining whether cognitive factors are predictive of some, but not other discourse levels and examining the relative predictive strength between cognitive factors for a given discourse level would establish preliminary findings necessary to support future latent factor analyses that could more robustly characterize the cognitive processes underlying discourse outcomes in aphasia.

2.0 Methods

2.1 Overview

Despite an extensive evidence-base supporting SFA's efficacy for picture naming over the past 25 years, and its widespread clinical use, it remains unclear how and to what extent SFA improves discourse-level communication (Boyle, 2011). The purpose of this retrospective study is to evaluate the contributions of potential treatment mechanisms affecting general-topic monologue discourse outcomes in SFA through multilevel discourse analysis and (2) explore potential non-language cognitive processes which support discourse outcomes in aphasia.

2.2 Description of Study Data Sources

The proposed project leveraged de-identified archival behavioral testing data from the aphasia research program at VA Pittsburgh. It included data from 60 individuals with chronic aphasia due to unilateral left-hemisphere stroke who completed an intensive SFA treatment across two sequential clinical trials. Participant data was included from 44 individuals with aphasia from *Dosage and predictors of naming treatment response in aphasia* (SFA-1; VA RR&D 5I01RX000832-05, IRB: STUDY1617169) and the first 16 individuals with aphasia to complete an ongoing trial, *Optimizing and understanding semantic feature analysis treatment for aphasia: A randomized controlled comparative effectiveness trial* (SFA-2; NIDCD 1R01DC017475-01A1, IRB: STUDY1617342). Both trials employ an SFA protocol with differences that are orthogonal to discourse outcomes but will be addressed by the statistical design. Unless otherwise noted, the following description of the SFA study protocol applies to both studies, and is summarized from existing publications of SFA-1 (Evans, Cavanaugh, Gravier, et al., 2021; Gravier et al., 2018). Methods undertaken for the present dissertation study are described in sections 2.3 (Discourse Analysis) and 2.4 (Statistical Analysis).

2.2.1 Eligibility Criteria

For both SFA-1 and SFA-2, participants 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). To qualify, participants were required to be least 6 months post-onset of left-hemisphere stroke and without progressive neurological disorder or severe motor speech disorder (Duffy, 2013). Participants were required to score below 70 on the modality mean T-score (overall score) and above 40 on the auditory comprehension and naming subtests on the Comprehensive Aphasia Test (CAT; Swinburn et al., 2004). A T-score of 50 indicates average performance on the CAT normative sample of individuals with aphasia; the standard deviation of the normative sample was 10 T-score points. Participants were also not permitted to receive any concurrent speech-language treatment outside of study activities for the duration of the study.

2.2.2 Characteristics of the Participant Sample

Demographic characteristics of participants from SFA-1, SFA-2, and both studies overall are reported in Table 1. Participants averaged 62 years old and were majority white and male, consistent with primary recruitment population (Veterans of the United States armed forces). Nearly all participants (58) reported having at least 12 years of education and nearly half (25) reported they had at least 16 years of education. There was a wide range of time post stroke onset, ranging from the minimum 6 months to over 20 years. Participant's aphasia severity was approximately average (52.6) relative to the expected distribution of the CAT, with all mean T-scores falling above 40 and below 70 (2 standard deviations above the mean).

Table 1 Demographic Summary of Participants from SFA-1 and SFA-2

	SFA-1 (N=44)	SFA-2 (N=16)	Overall (N=60)
Age			
Mean (SD)	61.6 (12.2)	62.3 (14.8)	61.8 (12.8)
Median [Min, Max]	65.5 [24.0, 78.0]	66.5 [26.0, 79.0]	65.5 [24.0, 79.0]
Sex			
Female	5 (11.4%)	4 (25.0%)	9 (15.0%)
Male	39 (88.6%)	12 (75.0%)	51 (85.0%)
Race/Ethnicity			
White	36 (81.8%)	14 (87.5%)	50 (83.3%)
Black	5 (11.4%)	1 (6.3%)	6 (10.0%)
Alaska Native	0 (0%)	1 (6.3%)	1 (1.7%)
Black / Native American	2 (4.5%)	0 (0%)	2 (3.3%)
Hispanic	1 (2.3%)	0 (0%)	1 (1.7%)
Years of Education			
Mean (SD)	14.9 (3.19)	15.3 (2.21)	15.0 (2.95)
Median [Min, Max]	14.0 [10.0, 25.0]	16.0 [12.0, 18.0]	14.0 [10.0, 25.0]
Months Post-onset			
Mean (SD)	62.5 (57.2)	53.9 (37.1)	60.2 (52.5)
Median [Min, Max]	47.5 [6.00, 245]	36.5 [17.0, 125]	38.5 [6.00, 245]
Cat Mean T-Score			
Mean (SD)	52.3 (4.72)	53.3 (5.47)	52.6 (4.90)
Median [Min, Max]	51.8 [44.3, 64.2]	53.1 [44.8, 65.2]	52.0 [44.3, 65.2]

2.2.3 Assessment

A speech-language pathologist administered and scored an extensive assessment battery as part of both SFA-1 and SFA-2 at study enrollment, treatment entry, exit, and 1-month follow-up. As noted above, to assess language ability, the CAT was administered to all participants upon study enrollment. Assessments from the battery relevant to this dissertation work are described below.

Monologue-based discourse production was elicited using the Nicholas and Brookshire protocol for all participants at all time points. This monologue-based discourse elicitation protocol (Brookshire & Nicholas, 1994; Bryant et al., 2016; Nicholas & Brookshire, 1993) is commonly used in aphasia and especially in studies examining the effects of Semantic Feature Analysis (Boyle, 2011). It is generally thought to elicit language samples that reflect potential “far generalization” of treatment effects to a connected speech context which is not explicitly related to treatment targets (though this is not tightly controlled). This protocol is composed of two matched sets (Brookshire & Nicholas, 1994) of 5 structured discourse stimuli across multiple genres: 2 picture descriptions, 1 narrative, 1 personal story, and 1 procedural discourse sample. Sets alternated between time points to minimize test-retest effects and order was counter-balanced to minimize ordering effects.

Only individuals with aphasia in SFA-1 (n = 44) completed the cognitive testing battery at treatment entry. This battery examined pre-treatment non-linguistic cognitive skills, including semantic memory, attention, and executive function for SFA-1 (n=44). The following assessments were administered: 1. semantic memory: Pyramids and Palm Trees (PPT; Howard & Patterson, 1992). 2. attention: Test of Everyday Attention, elevator counting with and without distraction (TEA; Robertson et al., 1994). 3. executive function (cognitive flexibility and problem solving): Wisconsin Card Sorting Task (WCST; Grant & Berg, 1948). These assessments were chosen for SFA-1 because they have been previously shown to be associated with aphasia treatment outcomes (Lambon Ralph et al., 2010). They are further pertinent to the present proposal given the hypothesized relationships between semantic memory, attention, and executive function on spoken discourse production.

(1) Semantic Memory: The Pyramids and Palm Trees (Howard & Patterson, 1992) is a commonly used assessment of semantic memory where participants are asked to match a stimulus to one of two target images. The Pyramids and Palm Trees test is considered a "relatively pure measure of conceptual semantic processing" (Martin et al., 2006 p. 158), which has been used successfully in aphasia and related disorders.

(2) Divided Attention: The Test of Everyday Attention (Robertson et al., 1994) is established as an aphasia-friendly, ecologically valid assessment of multiple forms of attention that has been standardized on a group of stroke patients. The Test of Everyday Attention subtests, elevator counting with and without distraction, are thought to index sustained and divided attention, respectively.

(3) Executive Function: The Wisconsin Card Sorting Task (Grant & Berg, 1948) is an established assessment of complex executive function which has been used previously in both aphasia and other neurological cognitive-communication disorders. While considered an assessment of executive function, it involves a variety of executive processes (e.g., attention switching, updating working memory) and non-executive processes (e.g., phonological retention and rehearsal; Allen et al., 2012).

2.2.4 Semantic Feature Analysis Treatment

2.2.4.1 Stimuli

Treatment lists were generated for all participants based on performance on a pre-treatment naming battery. For SFA-1, this battery included 194 items across eight semantic categories and was administered three times. Items that were named incorrectly at least twice were eligible for treatment. For SFA-2, naming ability was estimated from performance on the full or adaptive version of the Philadelphia Naming Test, and a published algorithm was used to generate eligible items of appropriate difficulty (Fergadiotis et al., 2015). Items were then selected to fill semantic categories with sufficient numbers of qualifying items. Four lists of ten items were generated for participants 1-4 in SFA-1, three lists of five items were generated for participants 5-44 in SFA-1, and three lists of ten items were generated for all 16 participants in SFA-2. The number and size of treatment lists were reduced during SFA-1 in order to minimize the burden of administering daily naming probes.

2.2.4.2 Experimental Design

For both studies, treatment was administered to each list sequentially. For SFA-1, treatment for a given list was discontinued when participants named 90% (Participants 1-4 in SFA-1) or 80% (participants 5-44 in SFA-1. and all participants in SFA-2) of treated items accurately on three of four consecutive probes, or if the list was trained for a maximum of eight days. A minimum of four treatment days was set for each list regardless of treatment probe accuracy to ensure sufficient exposure across lists. For SFA-2, participants received an equal number of sessions of treatment on each list and there were no probes during treatment.

Additionally for SFA-2, the last 3 days of treatment were sampled randomly from all 3 lists until all items were treated again.

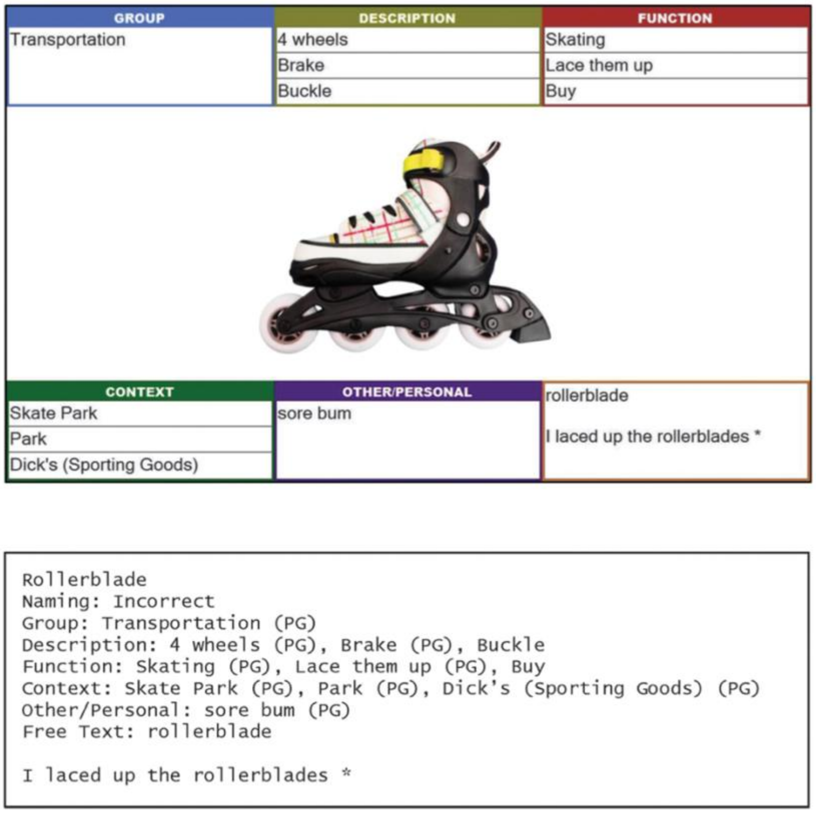
2.2.4.3 Procedures

All individuals with aphasia received SFA intervention four to five days per week over four weeks for a total of ~ 60 hours. Trained words were selected based on a picture-naming task. For the first four individuals with aphasia in SFA-1, treatment was provided for four lists of 10 words, but was decreased to three lists of five words for the remaining 40 individuals with aphasia to reduce the time burden for probes. In SFA-2, treatment was provided for three lists of 10 words. The SFA paradigm included the following components: 1 – naming: participants were presented with a picture and asked to name the item. 2 – feature generation: individuals with aphasia were asked to generate semantic features within five semantic categories (context, use, description, superordinate category, and personal association). A cueing hierarchy was used to facilitate successful feature production. If participants repeatedly generated the same features, clinicians encouraged the individual with aphasia to generate new features. 3 – naming: participants attempted to name the item again, with feedback. 4 – sentence generation: participants were asked to generate a sentence using the target word and two other elements with rich semantic content (verb; subject or object), typically involving semantic features generated during the trial. The clinician provided up to two cues to assist with sentence generation or modification; the sentence was repeated by the participant if necessary. This component is not standard in SFA but was included to increase the likelihood of generalization beyond naming tasks by providing a practice opportunity at the sentence level.

In SFA-1, treatment was delivered sequentially to each list in a multiple-baseline design. Treatment was provided for each list for a minimum of four days. Treatment progressed to the next list when participants named 90% (S1-S4) or 80% (S5-44) of words correctly on three of four consecutive daily naming probes. For SFA-2, each list was treated in succession for four consecutive days, with the final three days randomly sampled as indicated above.

2.2.4.4 Treatment Fidelity

Treatment fidelity was monitored in both SFA-1 and SFA-2. The procedures have minor differences between studies, but generally were as follows. At least 30 minutes of daily treatment sessions were video or audio recorded. A member of the study staff not providing treatment randomly selected 15 minutes of the recording for review of adherence to the protocol using an established fidelity checklist. Any checklist elements not checked were considered to be deviations. Deviations were reviewed and clarified with the treating clinicians. Treatment fidelity was monitored over the course of intervention for each participant.



GROUP	DESCRIPTION	FUNCTION
Transportation	4 wheels	Skating
	Brake	Lace them up
	Buckle	Buy

CONTEXT	OTHER/PERSONAL	
Skate Park	sore bum	rollerblade I laced up the rollerblades *
Park		
Dick's (Sporting Goods)		

```
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 *
```

Figure 3. Semantic Feature Analysis Treatment Worksheet

Gravier et al., 2018, reprinted with permission

2.3 Discourse Analysis

2.3.1 Discourse Transcription.

Discourse samples at each time point (treatment enrollment, entry, exit, and 1-month follow-up) were orthographically transcribed by a research speech-pathologist blinded to time point and manually scored for correct information units and words, which formed the basis for our preliminary analysis. 10% of samples were examined by a separate research speech-language pathologist and examined for transcription reliability using percent agreement on a word-by-word basis, which was found to exceed 90% and deemed acceptable. When frank transcription errors were found during the coding process (described below) and a correction was agreed upon by both independent raters, transcripts were subsequently amended. These amendments were extremely rare, and their prevalence was not recorded.

2.3.2 Discourse Outcome Measures

The discourse measures applied to archival language samples elicited from the Nicholas and Brookshire (1993) protocol were selected based on mentorship team expertise, a review of psychometric data (Boyle, 2020; Marini et al., 2011; Pritchard et al., 2018) demonstrating correlations between the selected measures (Andreetta et al., 2012; Fergadiotis et al., 2013; Harris Wright & Capilouto, 2012; Larfeuil & Dorze, 1997; Marini et al., 2011), and theoretical ties to SFA components and hypothesized mechanisms of action.[1] While there are ongoing efforts to expand test-retest stability data for discourse measures, it is not widely available (Pritchard et al., 2018). There are also known differences between discourse stimulus and genre. These concerns are explicitly addressed within the statistical approach.

[1] In the initial dissertation prospectus, outcome measures included predicate argument structure in lieu of mean utterance length as well as global coherence. These measures were modified mid study due to concerns about their psychometric properties and to facilitate study feasibility.

2.3.2.1 Lexical Semantic Processing

Lexical-semantic processing, or the ability to select semantically-appropriate words, was measured by the percentage of semantic paraphasias to the total number of phonologically well-formed content words (Marini et al., 2011). A restorative SFA mechanism should reduce the frequency of semantic and verbal errors while compensatory SFA mechanisms predict no changes in lexical semantic processing or the frequency of such errors.. Semantic errors were operationalized as real word errors with a plausible semantic relationship to the target word when the target word was clearly apparent. Errors with both a semantic and phonological component (i.e., mixed errors) were excluded from this measure. Real-word errors (i.e., real-word productions that were clearly not consistent with the context or content of the utterance) without a clear target were also excluded. Semantic relatedness was identified based on the raters' experience with scoring the Philadelphia Naming test, and potential semantic relationships with the target word based on the relationships identified in the Semantic Feature Analysis protocol (i.e., a contextual, functional, description, or categorical relationship to the target). Personal association relatedness was not considered for the purpose of identifying semantic errors given the lack of familiarity between participants and raters.

Content words were conceptually defined as words with an opportunity to make a semantic error and operationalized based on prior definitions (Fergadiotis et al., 2013). Content words included any noun, verb, adjective, adverb, and pronoun. Pronouns were included as content words because errors in pronoun selection (e.g., "*he* was washing dishes" in the cookie theft picture) were considered to be semantic errors. This definition of content words aligned best with the intended construct: the incidence of semantic errors in a sample, given the opportunities to make semantic errors.

2.3.2.2 Lexical Diversity

Lexical Diversity is "the range of vocabulary deployed in a text" reflecting the speaker's "capacity to access and retrieve target words from the lexicon" (Fergadiotis & Wright, 2011, p.145). Lexical diversity was measured using the moving average type-token ratio (MATTR; Covington, 2007), a length-invariant measure of lexical diversity which calculates a type-token ratio (the ratio of unique lexical items divided by the total number of words) for sequential but not overlapping segments of a sample. MATTR has been shown to be a unbiased estimate of lexical diversity in individuals with aphasia (Fergadiotis et al., 2013) used previously to measure discourse outcomes (Bunker et al., 2018). In this study, MATTR was calculated using a 10-word window (hereafter, MATTR-10) regardless of sample length to increase the variability in sample-level MATTR scores (see statistical analysis for additional details). A restorative mechanism should increase lexical diversity as measured by the MATTR-10 by improving the ability to retrieve a wider range of words while a compensatory mechanism should increase lexical diversity by giving individuals with aphasia strategies which facilitate more successful access to words already available in the lexicon.

2.3.2.3 Grammatical Complexity

Grammatical complexity was estimated using mean utterance length in words (MLU). MLU is an established proxy for grammatical complexity given known correlations between MLU and other indices of grammatical complexity (Gordon, 2020; Stark, 2019). The key hypothesis is that MLU would increase as participants were engaged in producing grammatically and semantically rich utterances on each trial of the SFA treatments, and improvements in lexical-semantic processing and lexical diversity would be reflected in longer and more complex utterances. MLU was calculated by the number of words (after removing repetitions and revisions, and leading carrier words) divided by the number of utterances in each sample. Utterance boundaries were determined following guidelines established in the Quantitative Analysis of Agrammatic Production Appendix A section IV (Saffran et al., 1989).

2.3.2.4 Discourse Informativeness

Informativeness was measured by calculating the percentage of correct information units (Brookshire & Nicholas, 1994) as utilized in the preliminary analysis and following the protocols for SFA-1 and SFA-2. Informativeness is well established in the aphasia treatment literature (Bryant et al., 2016) and incorporates micro- and macro-structural elements, incorporating both the accuracy of each word in a sample and the word's relevance to the topic. As demonstrated in Marini et al. (2011), it may be sensitive to changes in underlying linguistic structures or solely improvements in strategic communication use.

2.3.3 Reliability

Rather than calculating reliability against a second rater with a subsample of all transcripts, all samples were independently coded by two certified speech-language pathologists who were blinded to time point. A discourse scoring codebook was established prior to coding and iteratively revised throughout discourse coding to account for edge-cases and grey areas not addressed by initial coding rules. The codebook, including coding rules, rationales, and examples is reported in Appendix D. Guides from the Quantitative Analysis of Agrammatic Production (QPA; Saffran et al., 1989) were used to identify narrative and non-narrative words, repetitions, revisions, and utterance boundaries in each sample for the calculation of lexical-semantic processing, MLU, and lexical diversity. After coding, both speech-language pathologists met to resolve disagreements. A third expert rater also blinded to time point (committee member Dr. Davida Fromm) was engaged to resolve disagreements or ambiguous instances when the first two raters could not come to an agreement. The exception to the dual-rater approach is discourse informativeness. For this measure, transcripts originally scored as part of the SFA-1 and SFA-2 protocols were used. These transcripts were scored by research speech-language pathologists trained in scoring CIUs and blinded to time point. Reliability for scoring CIUs has been established as part of the SFA-1 and SFA-2 protocol and is incorporated into training.

2.4 Statistical Analysis

2.4.1 Overview

Changes in discourse outcome measures and moderators of such change were analyzed using Bayesian generalized mixed-effects models (Bürkner, 2017; Carpenter et al., 2017; McElreath, 2020). The Bayesian hierarchical model approach is well-suited to this study because (1) it can appropriately characterize the crossed and nested data structure (i.e., measurements within participants and across stimuli) and (2) leverages partial pooling (i.e., shrinkage) to produce more generalizable estimates of group-level adjusted effects (McElreath, 2020). Bayesian frameworks permit a more intuitive interpretation of results (for example, a 95% credible interval is interpreted as a 95% probability that the parameter of interest falls within the interval, given the data and prior assumptions). Bayesian modeling also often allows for greater model complexity (i.e. group-level effect structures) than frequentist models.

All analyses were executed in R (R Core Team, 2020) version 4.2.2. via Stan (Carpenter et al., 2017) accessed through the R package brms (Bürkner, 2017) and cmdstanr (Gabry & Češnovar, 2020). Although data were not permitted to be shared, R code for statistical models is reported in Appendix F. While the models did not exactly follow the anticipated analysis articulated prior to the start of this study for practical reasons, the initial Bayesian sample size estimation and reasons for modification are reported in Appendix C.

2.4.2 Aim 1 Analysis

For question 1, separate item/stimuli-level models were implemented for each discourse outcome measure with performance on the outcome measure as the dependent variable. A population-level effect (i.e., fixed effect) of time point was included as a four-level categorical variable (study enrollment, treatment entry, treatment exit, and 1-month follow-up). Treatment entry was used as the reference category such that each time point parameter describes change relative to treatment entry. Models used theoretically-informed (Matuschek et al., 2017) group-level effect (i.e., random effect) structures established a-priori. Group-level effects included stimuli as a random intercept, to account for the pseudorandomized order of stimuli presentation across participants. Group-level effects also included a group-level intercept for participant and a by-participant slope for time point. Taken together, this group-level effect structure appropriately characterizes the crossed and nested nature of the data (stimuli crossed with participants but delivered in different order across subjects) and permits participants not only to begin at their own level (intercept) but also change at different rates.

Each model was run with weakly informative, regularizing priors to improve sampling efficiency. To ensure priors reflected a reasonable sampling space, prior predictive checks were conducted by running each model while sampling only the prior distributions and visualizing the resulting parameter estimations for each class of parameters. Because priors are defined on the scale of the outcome measure, given the link function, specific priors for each model are described within the context of each model below. For each model, four independent Hamiltonian Markov Chain Monte Carlo (MCMC) chains were run with 3000 iterations. The initial 1000 iterations were discarded as a warmup and not included in parameter estimation.

2.4.2.1 Lexical-Semantic Processing (% of semantic errors to content words)

A beta-binomial mixture probability distribution with a logit link function was used to model the proportion of content words that were considered semantic errors. The beta-binomial mixture model has an advantage over beta regression in that it accounts for sample length and an advantage over aggregated binomial regression in that it has an additional model parameter to account for overdispersion. The prior distribution for the population-level effects of time point consisted of a normal distribution with mean of zero (no effect) and standard deviation of 1 logit. This distribution is informed by prior knowledge that semantic errors are relatively uncommon in general and in discourse samples. A one-logit decrease from 10% errors is roughly equivalent to a 6-percentage point change. Therefore, the *Normal (0, 1)* prior distribution reasonably constrains the sampling space while permitting a wide range of possible values. Default wide but regularizing priors on the group-level standard deviation were used, consistent of a student-t distribution with three degrees of freedom, a location value of 0, and a scale of 2.5. The default LKJ (Lewandowski, Kurowicka, and Joe) prior with lkj = 1 was utilized for the correlation parameter. With a value of 1, extreme correlation values become less likely.

2.4.2.2 Lexical Diversity (Moving Average Type-Token Ratio)

A zero-inflated beta regression model probability distribution with a logit link function was used to model lexical diversity as measured by the MATTR-10. This two-step model is well suited to the MATTR because a non-trivial number of samples did not exceed the requisite 10 tokens to be calculated. The zero-inflated beta regression approach models whether the sufficient tokens were produced to calculate the MATTR-10 (*zi* parameter). Provided that at least 10 tokens were produced, it then estimates lexical diversity as measured by the MATTR-10. Both the *zi* parameter (odds of producing at least 10 tokens) and the MATTR-10 measure were included as multivariate outcome measures, allowing the model to estimate changes both in producing sufficient tokens and increasing the lexical diversity of the tokens produced. The marginal effects of time point can then be estimated across both parts of the model. This approach permits modeling lexical diversity with a larger window (10 tokens) to produce more variability in the metric, while avoiding the omission of samples that were not of sufficient length. The prior distribution for the population-level effects of time point consisted of a normal distribution with mean of zero (no effect) and standard deviation of 1 logit. This prior is informed by expectations that the MATTR-10 will be constrained to the upper third of possible values (e.g., scores generally greater than .67) based on previous data (Cunningham & Haley, 2020; Fergadiotis & Wright, 2011) and the expectation that average change would be modest in size. The prior distribution for the *zi* parameter was a normal distribution with a mean of zero and standard deviation of three, the wider prior reflecting greater uncertainty about the potential effects of treatment on this parameter without existing data or studies. Default wide but regularizing priors on the group-level standard deviation were used, consistent of a student-t distribution with three degrees of freedom, a location value of 0, and a scale of 2.5. The default

LKJ prior with lkj = 1 was utilized for the correlation parameter, such that extreme correlation values are thought to be less likely.

2.4.2.3 Grammatical Complexity (Mean Length of Utterance)

A shifted lognormal probability distribution with an identity link function was used to model the mean length of utterance. This model was chosen because it is appropriate for non-negative continuous data, and demonstrated best fit compared with similar probability distributions (lognormal, gamma) as determined by posterior predictive checks and leave-one-out cross-validation. The prior distribution for the population-level effects of time point consisted of a normal distribution with mean of zero (no effect) and standard deviation of 1, which reasonably constrains the sampling space while permitting a wide range of possible values. Default wide but regularizing priors on the group-level standard deviation were used, consistent of a student-t distribution with three degrees of freedom, a location value of 0, and a scale of 2.5. The default LKJ prior with lkj = 1 was utilized for the correlation parameter. With a value of 1, extreme correlation values become less likely.

2.4.2.4 Informativeness (% CIUs)

A beta-binomial mixture probability distribution with a logit link function was used to model the proportion of words that were CIUs. Based on prior modeling of informativeness using similar models (Boyle et al., 2022), the prior distribution for the population-level effects of time point consistent of a normal distribution with mean of zero (no effect) and standard deviation of 1 logit. This prior is informed by existing data on treatment effects on informativeness (Boyle et al., 2022), that the mean change is unlikely to exceed 10-15 percentage points from the initial scores. A 1 logit change from 50% informativeness (50% CIUs to words) is equivalent to a 23.1 percentage point change. Therefore, the *Normal (0, 1)* prior distribution reasonably constrains the sampling space while permitting a wide range of possible values. Default wide but regularizing priors on the group-level standard deviation were used, consistent of a student-t distribution with three degrees of freedom, a location value of 0, and a scale of 2.5. The default LKJ prior with lkj = 1 was utilized for the correlation parameter.

2.4.2.5 Correlations Between Individual Change Estimates

The initial analysis plan included multivariate models permitting estimation of correlations of treatment effects across outcome measures, which would describe the extent to which discourse outcome measures changed together. However, this approach was not possible due to the need to use different probability distributions for each outcome measure to best capture the data generation process. As a compromise, individual change scores were calculated for each measure between treatment enrollment and treatment entry, entry and treatment exit, and entry and 1-month follow-up. Bayesian linear correlations with a wide, non-informative prior were estimated between each outcome measure for all individual change scores (Morey & Rouder, 2022). This method is limited by the fact that there are four sources of measurement error for each correlation (measurement error at each time point), and interpretation of these correlations should be considered exploratory and treated with caution.

2.4.3 Aim 2 Analysis

For question 2, models from Aim 1 were extended to include three moderators of interest, semantic memory, attention, and executive function. These moderators were added to each model by including population-level effects for each variable, as well as interactions between the moderating variable and time point. All three variables were added to the same model for each outcome measure, such that the effects of each moderator would reflect any effects above and beyond the moderating effects of the others. Aphasia severity (CAT mean modality T-score, z-scored and centered) was included as a covariate in all four models.

Semantic memory was defined as a two-level categorical variable derived from the Pyramids and Palm Trees overall score, defined as "impaired" if the PPT score was less than 0.9 and "normal" if greater than or equal to 0.9. The PPT score was binned into these two groups given concerns about the psychometric properties of the PPT overall score (e.g., Hula et al., 2012) and evidence supporting an established cut-off of 0.9 for impaired semantic memory performance (PPT; Howard & Patterson, 1992). The PPT indicator variable was sum-coded such that the effect of time point continued to reflect the overall effect of time point. For attention, the score from the distraction subtest of the Test of Everyday Attention (TEA) was z-scored by the standard deviation of the study participants and centered such that the effect of attention would reflect the change in the discourse outcome measure for a 1-standard deviation change in the distraction subtest of the TEA. Finally, for executive function, the highest category achieved on the WCST was used as a monotonic, ordinal predictor (Bürkner & Vuorre, 2019), with levels including 0, 1, 2, 3, and greater than or equal to 4. There were only five participants who achieved category 4 (8.3%) and five participants who achieved category 5 on the WCST.

Therefore, these categories were collapsed into a single category to maintain more balanced cells. The monotonic, ordinal predictor approach has the advantage of returning a single parameter estimate (whereas a categorical predictor would return multiple estimates), while avoiding any assumptions about the interval status of the data. Prior distributions on all moderating population-level effects were equivalent to priors on the main effects, resulting in wide, uninformative priors.

2.4.4 Model Convergence and Fit

Models were assessed for convergence by ensuring the split-half potential scale reduction factor ($\hat{R}$), was less than 1.01 and the number of effective samples were greater than 400 for each parameter estimate (Gelman et al., 2013). The split-half potential scale reduction factor can be interpreted as the ratio of the variance within each chain to the variance pooled across chains. Values that exceed 1.01 indicate a lack of convergence and inadequate mixing of the Markov Chain Monte Carlo chains. The number of effective samples represents the number of independent draws that were not autocorrelated with previous draws. A greater number of effective samples indicates more stability in the distributions from which the parameter estimates are derived. After establishing that the estimated $\hat{R}$ values were less than 1.01, and the number of effective sample sizes exceeded 400 for all parameters, models were judged to have adequate convergence. Diagnostic and convergence plots for all models are reported in Appendix B.

Finally, posterior predictive checks were conducted to confirm that the models adequately fit the data. Posterior predictive checks represent the process of estimating many simulations of the data and comparing these simulations to the observed data. 500 predictive distributions were estimated for each dependent variable. All estimated distributions closely followed the density distribution pattern for the observed data, indicating that the estimated models adequately fit the data.

3.0 Results

3.1 Evaluating Monologue Discourse Outcomes in Semantic Feature Analysis Treatment for Aphasia (Aim 1)

3.1.1 Lexical Semantic Processing

Results for changes in lexical-semantic processing are reported in Table 2 and visualized in Figure 4. The intercept estimate was -3.92 (95% CI: [-4.50, -3.34]), indicating that semantic errors, were extremely uncommon upon treatment entry; roughly 2% of total content words. No changes between treatment time points were reliably different from zero. The estimated change from enrollment to treatment entry was -0.09 (95% CI: [-0.36, 0.14]). The estimated change from entry to treatment exit was -0.06 (95% CI: [-0.30, 0.17]) and from entry to 1-month follow-up was, -0.08 (95% CI: [-0.31, 0.15]). These results indicate that semantic errors were relatively rare across all treatment time points, with no evidence for meaningful reduction as a result of treatment on average.

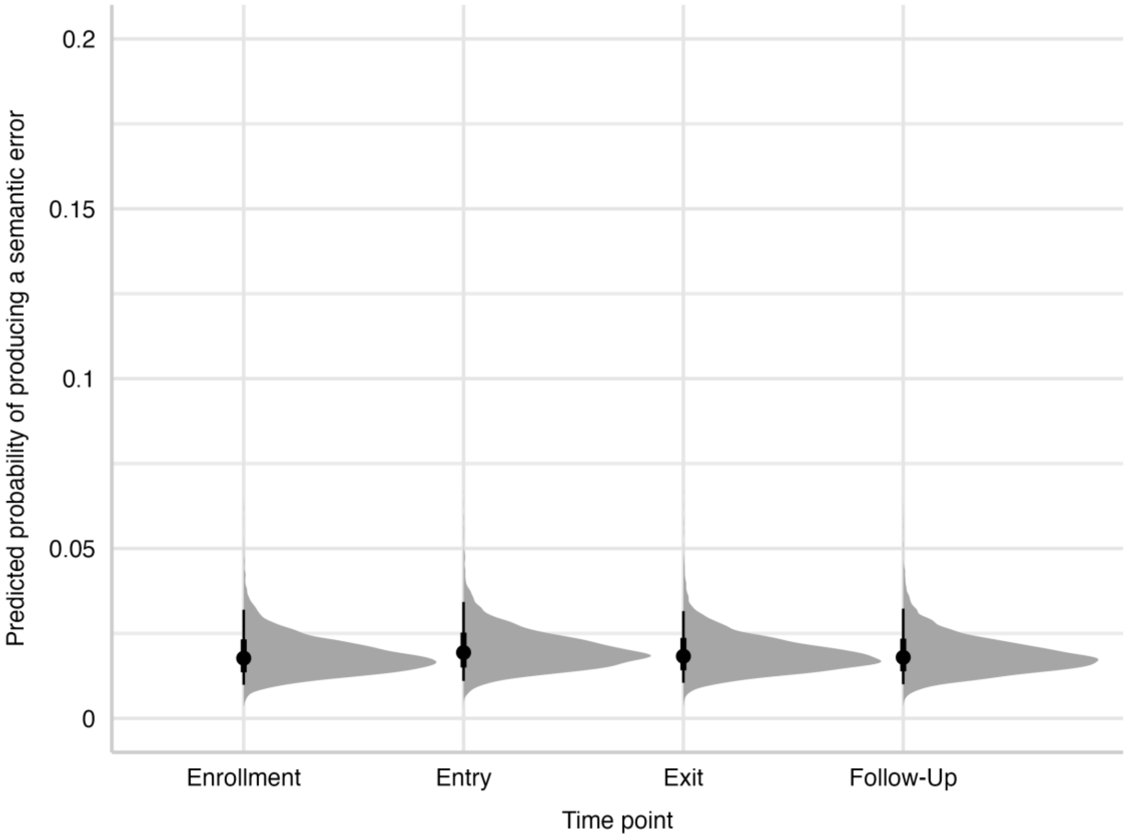


Figure 4. Conditional effects of time point on lexical-semantic processing

Point estimates show the median estimate. Error bars represent 68% and 95% credible intervals. Density plots represent the entire posteror distribution at each time point.

Table 2. Model results for lexical-semantic processing

Effects	Parameter	Median	95% CI	pd
Population	Intercept	-3.92	[-4.50, -3.34]	100%
	Time point [Enrollment - Entry]	-0.09	[-0.36, 0.14]	76.53%
	Time point [Exit - Entry]	-0.06	[-0.30, 0.17]	70.21%
	Time point [Follow-up - Entry]	-0.08	[-0.31, 0.15]	74.58%
Group	SD (Intercept: participant)	0.55	[0.38, 0.76]	100%
	SD (Time point [Enrollment - Entry]: participant)	0.31	[0.02, 0.70]	100%
	SD (Time point [Exit - Entry]: participant)	0.34	[0.04, 0.63]	100%
	SD (Time point [Follow-up - Entry]: participant)	0.20	[0.01, 0.55]	100%
	cor (Intercept ~ Time point [Enrollment - Entry]: participant)	0.13	[-0.57, 0.78]	63.62%
	cor (Intercept ~ Time point [Exit - Entry]: participant)	-0.70	[-0.95, 0.09]	96.67%
	cor (Time point [Enrollment - Entry] ~ Time point [Exit - Entry]: participant)	0.01	[-0.78, 0.73]	51.16%
	cor (Intercept ~ Time point [Follow-up - Entry]: participant)	0.15	[-0.68, 0.84]	63.15%
	cor (Time point [Enrollment - Entry] ~ Time point [Follow-up - Entry]: participant)	0.04	[-0.79, 0.82]	53.05%
	cor (Time point [Exit - Entry] ~ Time point [Follow-up - Entry]: participant)	1.87e-03	[-0.79, 0.75]	50.12%
	SD (Intercept: stimuli)	0.78	[0.49, 1.42]	100%
Population	phi	73.86	[54.55, 104.75]	100%

Note: CI: Credible Interval, pd: probability of direction, SD: standard deviation, cor: correlation

3.1.2 Lexical Diversity

Results for changes in lexical diversity are reported in Table 3 and visualized in Figure 5. The intercept estimate was 1.75 (95% CI: [1.62, 1.88]) indicating that the average lexical diversity score (as estimated by the MATTR-10) was about 0.85 at treatment entry. No changes between treatment time points were reliably different from zero. The estimated change from enrollment to treatment entry was -0.05 (95% CI: [-0.15, 0.04]). The estimated change from entry to treatment exit was -0.06 (95% CI: [-0.13, 0.01]) and from entry to 1-month follow-up was, -0.02 (95% CI: [-0.11, 0.06]). These results indicate that lexical diversity was relatively stable across all treatment time points with no evidence for meaningful change as a result of treatment. Notably, the zero inflated parameters were large and negative with a substantial amount of posterior distributions less than zero, suggesting that there was weak evidence for an effect between treatment entry and exit (-1.08, 95% CI: [-3.92, 0.55]) and treatment entry and follow-up (-0.64, 95% CI: [-2.50, 0.78]), such that the probability of not producing a sample long enough to calculate the MATTR-10 decreased over time.

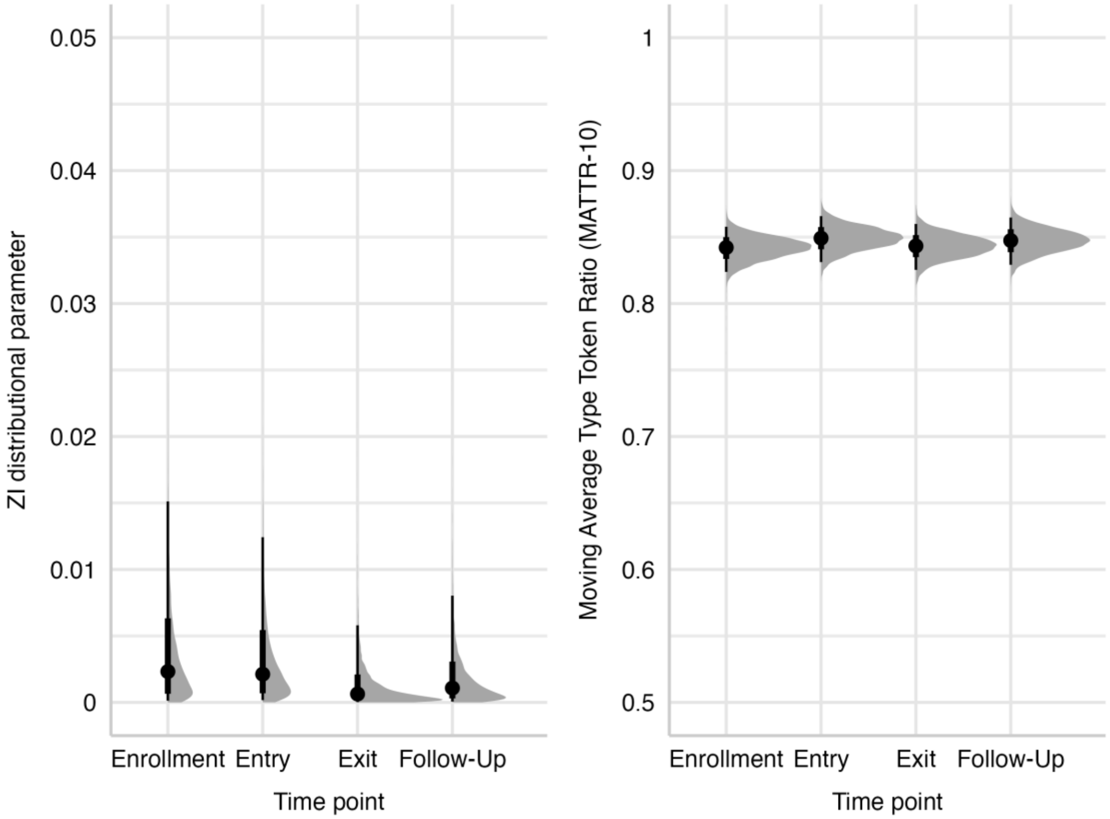


Figure 5. Conditional effects of time point on lexical diversity

Point estimates show the median estimate. Error bars represent 68% and 95% credible intervals. Density plots represent the entire posteror distribution at each time point.

Table 3. Model results for lexical diversity

Effects	Parameter	Median	95% CI	pd
Population	Intercept	1.75	[1.62, 1.88]	100%
	Time point [Enrollment - Entry]	-0.05	[-0.15, 0.04]	87.64%
	Time point [Exit - Entry]	-0.06	[-0.13, 0.01]	94.09%
	Time point [Follow-up - Entry]	-0.02	[-0.11, 0.06]	71.59%
Group	SD (Intercept: participant)	0.41	[0.33, 0.51]	100%
	SD (Time point [Enrollment - Entry]: participant)	0.22	[0.09, 0.33]	100%
	SD (Time point [Exit - Entry]: participant)	0.06	[0.00, 0.17]	100%
	SD (Time point [Follow-up - Entry]: participant)	0.12	[0.01, 0.25]	100%
	cor (Intercept ~ Time point [Enrollment - Entry]: participant)	-0.47	[-0.77, -0.09]	98.81%
	cor (Intercept ~ Time point [Exit - Entry]: participant)	-0.24	[-0.82, 0.66]	71.35%
	cor (Time point [Enrollment - Entry] ~ Time point [Exit - Entry]: participant)	0.27	[-0.67, 0.86]	70.21%
	cor (Intercept ~ Time point [Follow-up - Entry]: participant)	-0.02	[-0.60, 0.65]	53.10%
	cor (Time point [Enrollment - Entry] ~ Time point [Follow-up - Entry]: participant)	0.02	[-0.75, 0.71]	51.80%
	cor (Time point [Exit - Entry] ~ Time point [Follow-up - Entry]: participant)	0.03	[-0.79, 0.81]	52.05%
	SD (Intercept: stimuli)	0.08	[0.04, 0.15]	100%
Population	phi	40.31	[36.76, 44.27]	100%
Population: ZI	Intercept	-6.15	[-8.66, -4.37]	100%
	Time point [Enrollment - Entry]	0.08	[-2.06, 1.93]	54.24%
	Time point [Exit - Entry]	-1.08	[-3.92, 0.55]	90.18%
	Time point [Follow-up - Entry]	-0.64	[-2.50, 0.78]	82.24%
Group: ZI	SD (Intercept: participant)	3.53	[2.26, 5.48]	100%
	SD (zi_Time point [Enrollment - Entry]: participant)	1.38	[0.11, 3.20]	100%
	SD (zi_Time point [Exit - Entry]: participant)	1.25	[0.07, 3.32]	100%
	SD (zi_Time point [Follow-up - Entry]: participant)	0.55	[0.03, 2.05]	100%
	cor (zi_Intercept ~ zi_Time point [Enrollment - Entry]: participant)	-0.19	[-0.83, 0.65]	66.66%
	cor (zi_Intercept ~ zi_Time point [Exit - Entry]: participant)	0.10	[-0.71, 0.80]	58.31%

Effects	Parameter	Median	95% CI	pd
	cor (zi_Time point [Enrollment - Entry] ~ zi_Time point [Exit - Entry]: participant)	0.03	[-0.78, 0.79]	52.38%
	cor (zi_Intercept ~ zi_Time point [Follow-up - Entry]: participant)	-0.01	[-0.84, 0.82]	51.00%
	cor (zi_Time point [Enrollment - Entry] ~ zi_Time point [Follow-up - Entry]: participant)	6.76e-03	[-0.81, 0.81]	50.69%
	cor (zi_Time point [Exit - Entry] ~ zi_Time point [Follow-up - Entry]: participant)	0.03	[-0.80, 0.82]	52.15%
	SD (Intercept: stimuli)	0.92	[0.31, 2.01]	100%

Note: CI: Credible Interval, pd: probability of direction, SD: standard deviation, cor: correlation

3.1.3 Mean Length of Utterance

Results for changes in mean length of utterance in words are reported in Table 4 and visualized in Figure 6. The intercept estimate was 1.70 (95% CI: [1.56, 1.83]), indicating that the median mean length of utterance was estimated to be about 5.47 words at treatment entry. No changes between treatment time points were reliably different from zero. The estimated change from enrollment to treatment entry was -0.02 (95% CI: [-0.06, 0.03]). The estimated change from entry to treatment exit was 0.00 (95% CI: [-0.03, 0.04]) and from entry to 1-month follow-up was 0.00 (95% CI: [-0.03, 0.04]). These results indicate that mean length of utterance in words was relatively stable across all treatment time points with no evidence for meaningful change as a result of treatment.

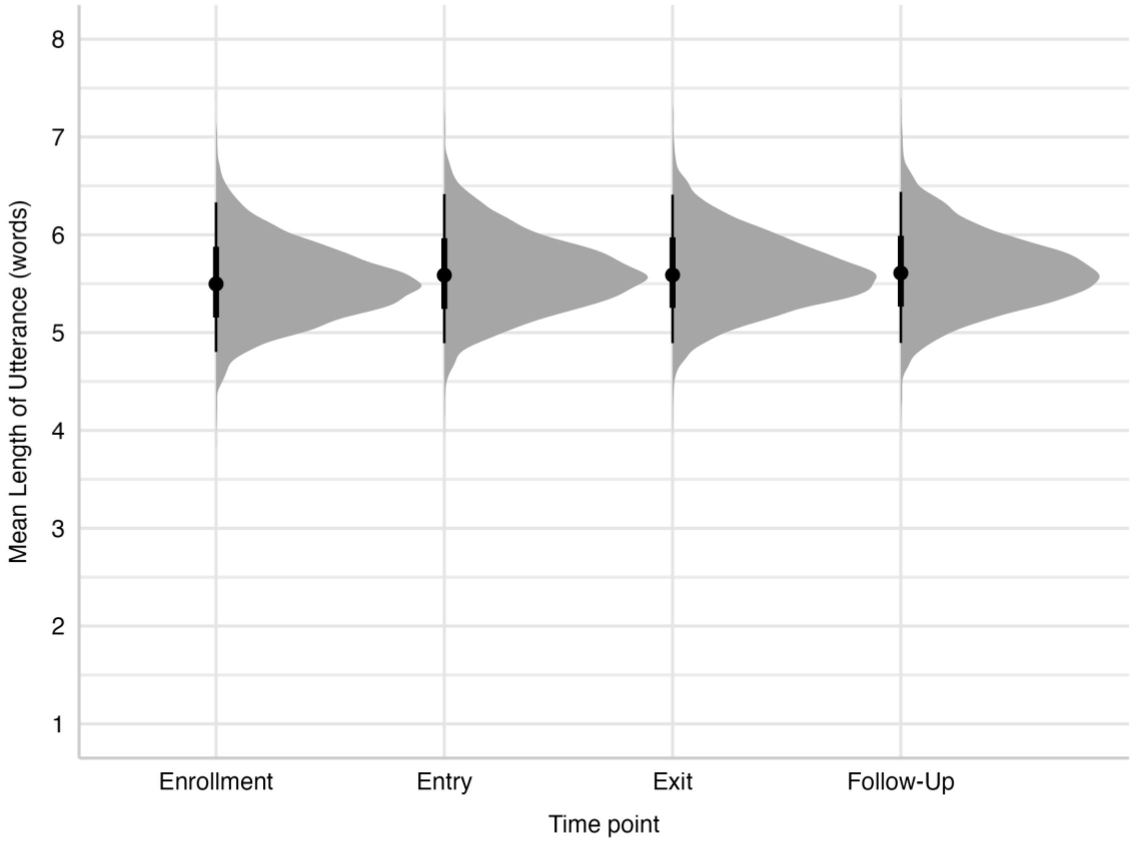


Figure 6. Conditional effects of time point on mean length of utterance

Point estimates show the median estimate. Error bars represent 68% and 95% credible intervals. Density plots represent the entire posteror distribution at each time point.

Table 4. Model results for mean length of utterance

Effects	Parameter	Median	95% CI	pd
Population	Intercept	1.70	[1.56, 1.83]	100%
	Time point [Enrollment - Entry]	-0.02	[-0.06, 0.03]	76.80%
	Time point [Exit - Entry]	1.12e-03	[-0.04, 0.04]	52.30%
	Time point [Follow-up - Entry]	4.51e-03	[-0.03, 0.04]	58.81%
Group	SD (Intercept: participant)	0.48	[0.40, 0.59]	100%
	SD (Time point [Enrollment - Entry]: participant)	0.09	[0.03, 0.14]	100%
	SD (Time point [Exit - Entry]: participant)	0.08	[0.01, 0.13]	100%
	SD (Time point [Follow-up - Entry]: participant)	0.03	[0.00, 0.09]	100%
	cor (Intercept ~ Time point [Enrollment - Entry]: participant)	-0.01	[-0.42, 0.45]	51.79%
	cor (Intercept ~ Time point [Exit - Entry]: participant)	-0.20	[-0.64, 0.32]	78.74%
	cor (Time point [Enrollment - Entry] ~ Time point [Exit - Entry]: participant)	-0.44	[-0.90, 0.27]	88.54%
	cor (Intercept ~ Time point [Follow-up - Entry]: participant)	-0.11	[-0.79, 0.71]	60.11%
	cor (Time point [Enrollment - Entry] ~ Time point [Follow-up - Entry]: participant)	0.04	[-0.78, 0.80]	52.90%
	cor (Time point [Exit - Entry] ~ Time point [Follow-up - Entry]: participant)	0.31	[-0.70, 0.90]	71.17%
	SD (Intercept: stimuli)	0.08	[0.05, 0.14]	100%
Population	sigma	0.23	[0.22, 0.24]	100%

Note: CI: Credible Interval, pd: probability of direction, SD: standard deviation, cor: correlation

3.1.4 Discourse Informativeness

Results for changes in discourse informativeness are reported in Table 5 and visualized in Figure 7. The intercept estimate was -0.21 (95% CI: [-0.45, 0.04]) indicating that percentage of words that were correct information units was approximately 0.45 at treatment entry. The estimated change from enrollment to treatment entry was 0.00 (95% CI: [-0.10, 0.11]) indicating stability between treatment enrollment and entry. The estimated change from entry to treatment exit was 0.04 (95% CI: [-0.07, 0.14]), providing no reliable evidence of improvement. However, estimated informativeness did improve from entry to 1-month follow-up (b = 0.10, 95% CI: [-0.01, 0.21]) indicating that informativeness did improve to a small degree by treatment follow-up. The posterior probability that this effect is greater than zero was 96.49%. These results provide minimal evidence for improvements in informativeness at treatment exit and modest evidence for improvements at 1-month follow-up.

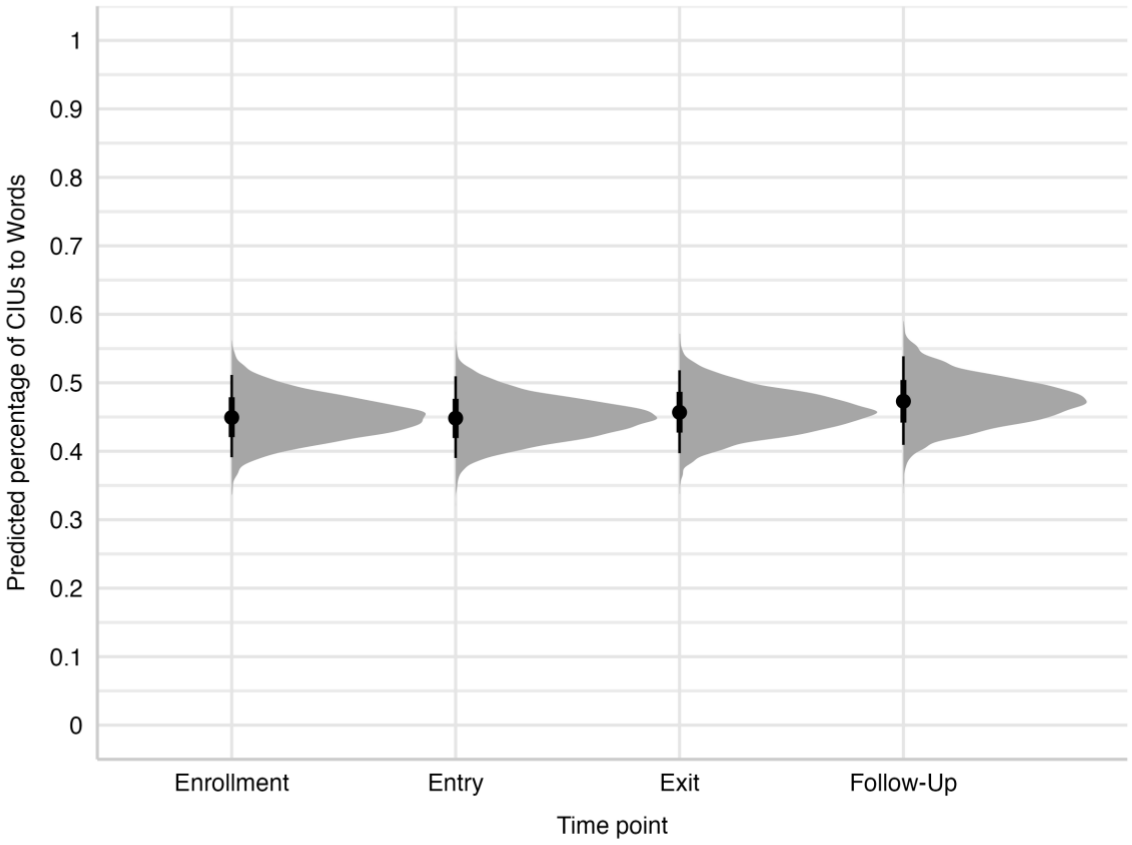


Figure 7. Conditional effects of time point on discourse informativeness

Point estimates show the median estimate. Error bars represent 68% and 95% credible intervals. Density plots represent the entire posteror distribution at each time point.

Table 5. Model results for discoures informativeness

Effects	Parameter	Median	95% CI	pd
Population	Intercept	-0.21	[-0.45, 0.04]	95.14%
	Time point [Enrollment - Entry]	5.53e-03	[-0.10, 0.11]	54.36%
	Time point [Exit - Entry]	0.04	[-0.07, 0.14]	75.15%
	Time point [Follow-up - Entry]	0.10	[-0.01, 0.21]	96.49%
Group	SD (Intercept: participant)	0.76	[0.63, 0.95]	100%
	SD (Time point [Enrollment - Entry]: participant)	0.19	[0.02, 0.33]	100%
	SD (Time point [Exit - Entry]: participant)	0.20	[0.02, 0.34]	100%
	SD (Time point [Follow-up - Entry]: participant)	0.21	[0.04, 0.37]	100%
	cor (Intercept ~ Time point [Enrollment - Entry]: participant)	-0.13	[-0.63, 0.52]	67.39%
	cor (Intercept ~ Time point [Exit - Entry]: participant)	0.02	[-0.50, 0.60]	52.90%
	cor (Time point [Enrollment - Entry] ~ Time point [Exit - Entry]: participant)	0.21	[-0.67, 0.78]	69.00%
	cor (Intercept ~ Time point [Follow-up - Entry]: participant)	0.22	[-0.32, 0.71]	78.97%
	cor (Time point [Enrollment - Entry] ~ Time point [Follow-up - Entry]: participant)	-0.28	[-0.86, 0.54]	74.84%
	cor (Time point [Exit - Entry] ~ Time point [Follow-up - Entry]: participant)	0.51	[-0.48, 0.92]	87.28%
	SD (Intercept: stimuli)	0.20	[0.12, 0.36]	100%
Population	phi	18.90	[16.73, 21.30]	100%

Note: CI: Credible Interval, pd: probability of direction, SD: standard deviation, cor: correlation

3.1.5 Correlations between Outcome measures

Correlations between individual change scores are reported in Figure 8 (treatment enrollment to treatment entry), Figure 9 (treatment entry to treatment exit), and Figure 10 (treatment entry to 1-month follow-up). None of the correlations were credibly different from zero, and these findings should be interpreted with considerable caution given that these correlation estimates are influenced by multiple sources of measurement error omitted from the model. Furthermore, there were no cases of individuals who clearly demonstrated improvements across three or four treatment outcome measures.

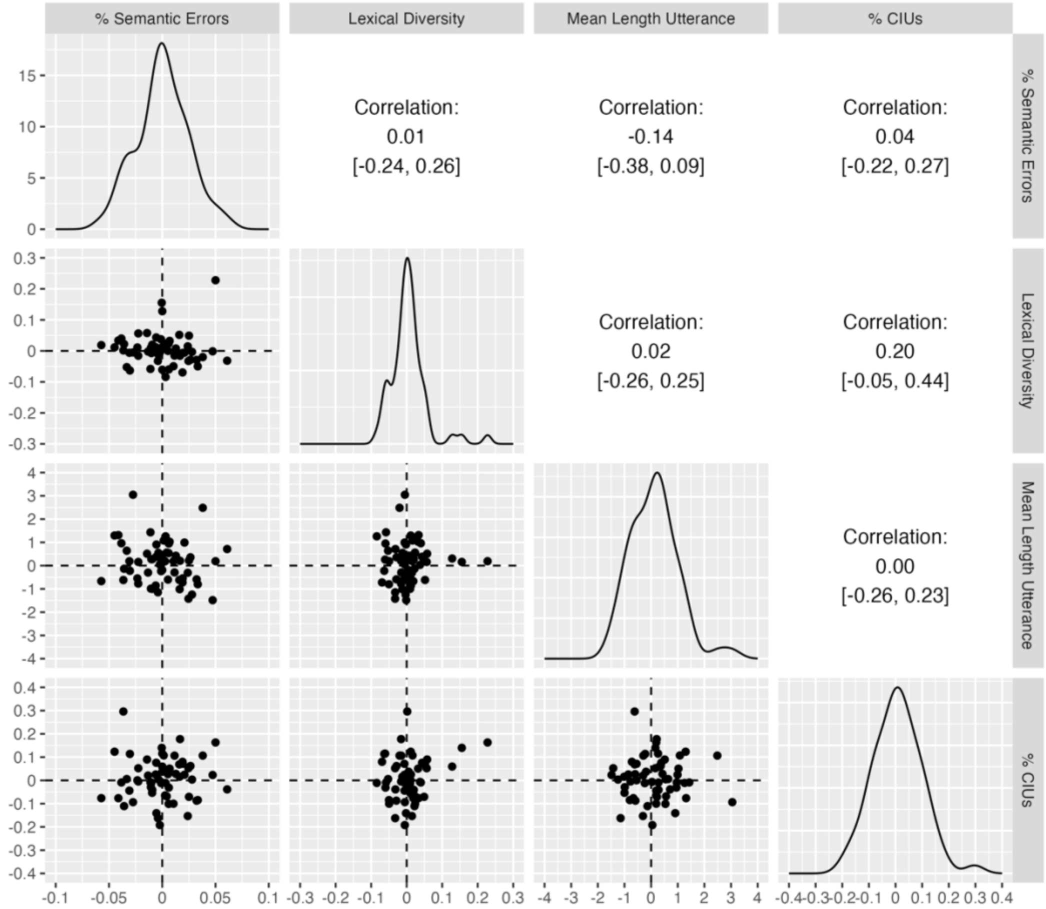


Figure 8. Correlations between change scores from treatment enrollment to treatment entry

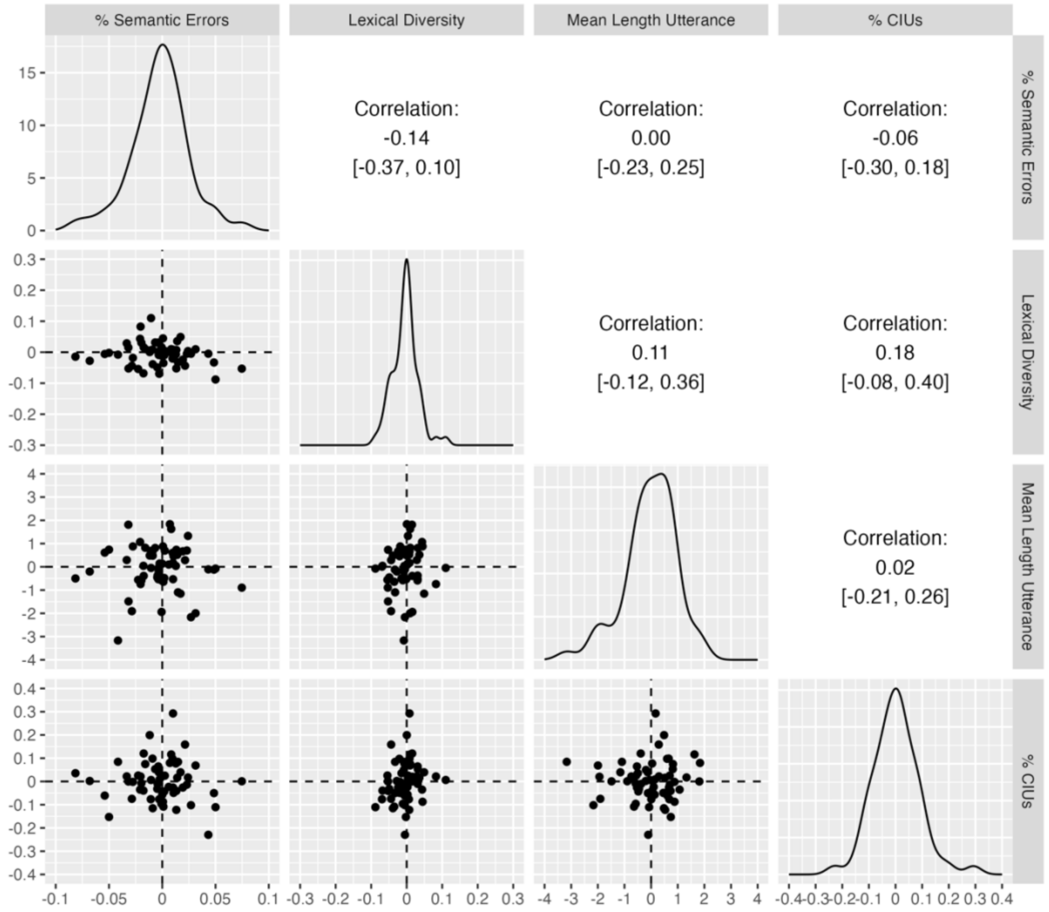


Figure 9. Correlations between change scores from treatment entry to treatment exit

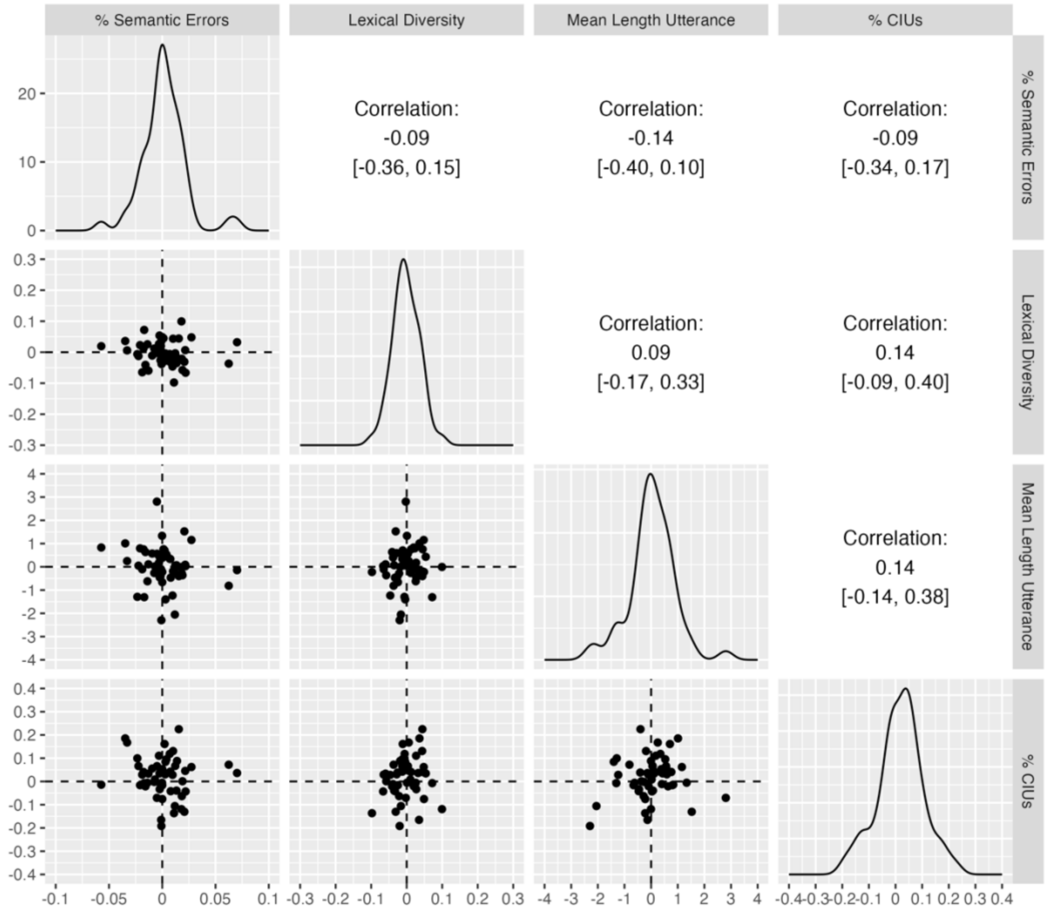


Figure 10. Correlations between change scores from treatment entry to 1-month Follow-up

3.2 Examining Cognitive-linguistic Moderators of Monologue Discourse Outcomes in Semantic Feature Analysis treatment for Aphasia (Aim 2)

3.2.1 Lexical Semantic Processing

Results are reported in Table 6 and visualized in Figure 11. For lexical-semantic processing, the intercept estimate was -3.95 (95% CI: [-4.66, -3.23]), indicating that semantic errors, were similarly uncommon upon treatment entry for the SFA-1 participants only (about 2% of total content words. For a participant with average PPT, WCST, and TEA scores, changes between treatment time points were not reliably different from zero. For this average participant, the estimated change from enrollment to treatment entry was -0.10 (95% CI: [-0.62, 0.43]). The estimated change from entry to treatment exit was -2.44e-03 (95% CI: [-0.45, 0.42]) and from entry to 1-month follow-up was -0.40 (95% CI: [-1.00, 0.09]). However, the 94.11% of the posterior for change from treatment entry to 1-month follow-up was less than zero, providing weak evidence that a participant with average non-language cognitive test scores may have demonstrated a small reduction in the proportion of semantic errors produced. Given the contrast coding of the cognitive-linguistic predictors, these effects of time point should be interpreted as the effect for the average score on the PPT, WCST, and TEA, accounting for variance explained by aphasia severity. Differences between the effects of time point for these models and those reported under aim 1 may result from the presence of moderating variables in the model, the impact of the interaction terms, the different group of participants (n = 44), or a combination of these.

Results for moderating non-language cognitive variables are as follows (Figure 11). For the TEA, the interaction between TEA performance and change between enrollment and entry was positive, but not reliably different from zero (0.16, 95% CI: [-0.14, 0.47]). The interaction between TEA performance and change between entry and exit was also positive (0.25 95% CI: [0.00, 0.50]), with 97.70% of the posterior greater than zero, suggesting that change from treatment entry to treatment exit might be related to TEA scores. The interaction between TEA performance and entry and 1 month follow-up was negative but not reliably different from zero

(-0.15 95% CI: [-0.43, 0.12]), indicating that the TEA score was unlikely to be a clear moderator of any potential treatment effect on the production of semantic errors.

For the Pyramids and Palm Trees, the interaction between PPT performance and change between enrollment and entry was negative, but not reliably different from zero (-0.24, 95% CI: [-0.97, 0.44]). The interaction between PPT performance and change from entry and exit was negative (-0.49 95% CI: [-1.08, 0.10]) with 94.80% of the posterior less than zero and the interaction between performance and change from entry to follow-up was negative (-0.85 95% CI: [-1.56, -0.14]) with 99.00% of the posterior less than zero providing potentially weak evidence that participants with higher PPT scores were more likely to produce fewer semantic errors after treatment, with the caveat that this change is unlikely to be meaningfully different from the change between treatment enrollment and treatment entry.

For the Wisconsin Card Sorting Test, the interaction between WCST performance and change between enrollment and entry was negative, but not reliably different from zero (-0.08, 95% CI: [-0.35, 0.17]). The interaction between PPT performance and change from entry to exit was negligible (0.03 95% CI: [-0.18, 0.25]). The interaction between performance and change from entry to follow-up was robustly positive (0.28 95% CI: [0.05, 0.55]), with 98.95% of the posterior greater than zero providing potentially moderate evidence that participants who achieved a higher category on the WCST increased the proportion of semantic errors produced at 1-month follow-up relative to participants who scored poorly on the WCST.

Table 6. Model results for moderation lexical-semantic processing changes (population-level effects)

Parameter	Median	95% CI	pd
Intercept	-3.95	[-4.66, -3.23]	100%
Time point [Enrollment - Entry]	-0.10	[-0.62, 0.43]	65.36%
Time point [Exit - Entry]	2.44e-03	[-0.45, 0.42]	50.51%
Time point [Follow-up - Entry]	-0.40	[-1.00, 0.09]	94.11%
TEA (z)	-0.23	[-0.45, -0.01]	98.00%
PPT	0.27	[-0.28, 0.78]	83.99%
WCST	2.33e-03	[-0.18, 0.19]	50.96%
CAT T-Score (z)	-0.13	[-0.33, 0.07]	90.39%
Time point [Enrollment - Entry] x TEA (z)	0.16	[-0.14, 0.47]	85.56%
Time point [Exit - Entry] x TEA (z)	0.25	[0.00, 0.50]	97.70%
Time point [Follow-up - Entry] x TEA (z)	-0.15	[-0.43, 0.12]	86.98%
Time point [Enrollment - Entry] x PPT	-0.24	[-0.97, 0.44]	75.31%
Time point [Exit - Entry] x PPT	-0.49	[-1.08, 0.10]	94.80%
Time point [Follow-up - Entry] x PPT	-0.85	[-1.56, -0.14]	99.00%
Time point [Enrollment - Entry] x WCST	-0.08	[-0.35, 0.17]	74.45%
Time point [Exit - Entry] x WCST	0.03	[-0.18, 0.25]	61.48%
Time point [Follow-up - Entry] x WCST	0.28	[0.05, 0.55]	98.95%

Note: CI: Credible Interval, pd: probability of direction, SD: standard deviation, cor: correlation. Full model results are reported in Appendix A

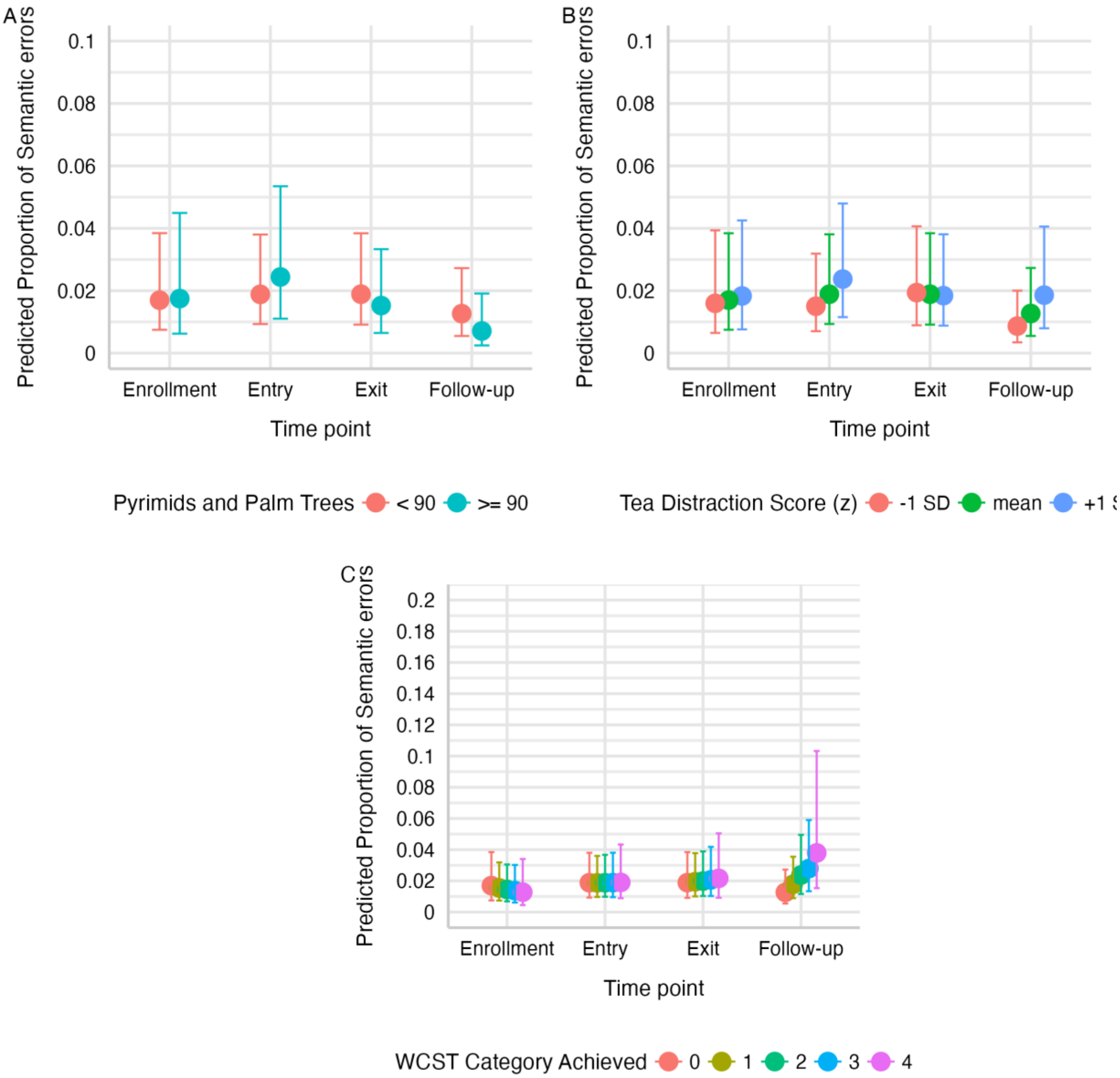


Figure 11. Relationships between lexical-semantic processing outcomes and non-langage cognitive scores.

Panel A: Semantic Memory; Panel B: Divided Attention; Panel C: Executive Function

3.2.2 Lexical Diversity

For lexical diversity, results are reported in Table 7 and visualized in Figure 11. The intercept estimate was 1.76 (95% CI: [1.51, 2.01]), indicating that the average MATTR-10 score at treatment entry was about 0.85. For a participant with average PPT, WCST, and TEA scores, changes between treatment time points were not reliably different from zero. For this average participant, the estimated change from enrollment to treatment entry was -0.12 (95% CI: [-0.32, 0.07]). The estimated change from entry to treatment exit was 0.05 (95% CI: [-0.12, 0.25]) and from entry to 1-month follow-up was, -0.12 (95% CI: [-0.28, 0.04]).

None of the three non-language cognitive variables appeared to have a relationship between change across time points (Figure 12). For the TEA, the interaction between TEA performance and change was negligible from enrollment to treatment entry (-0.03, 95% CI: [-0.14, 0.09]), treatment entry and exit (0.03 95% CI: [-0.06, 0.12]), and from treatment entry to follow-up (0.04 95% CI: [-0.06, 0.13]), indicating that the TEA score was unlikely to be a clear moderator of any potential treatment effect for lexical diversity. Similar findings were present for the Pyramids and Palm Trees from enrollment to entry (-0.08, 95% CI: [-0.34, 0.18]), treatment entry to exit (0.01 95% CI: [-0.20, 0.22]), and treatment entry to follow-up (0.08 95% CI: [-0.16, 0.32]). For the Wisconsin Card Sorting Test, the interaction between WCST performance and change between enrollment and entry (0.03, 95% CI: [-0.07, 0.13]), treatment entry to exit (-0.06 95% CI: [-0.16, 0.02]), and treatment entry to follow-up (0.06 95% CI: [-0.04, 0.20]) were not reliably different from zero.

Table 7. Model results for lexical diversity moderation (population-level effects)

Parameter	Median	95% CI	pd
Intercept	1.76	[1.51, 2.01]	100%
Time point [Enrollment - Entry]	-0.12	[-0.32, 0.07]	89.96%
Time point [Exit - Entry]	0.05	[-0.12, 0.25]	70.23%
Time point [Follow-up - Entry]	-0.12	[-0.28, 0.04]	93.89%
TEA (z)	0.04	[-0.10, 0.18]	72.91%
PPT	-0.03	[-0.38, 0.32]	57.04%
WCST	-9.80e-03	[-0.14, 0.10]	56.71%
CAT T-Score (z)	0.14	[0.00, 0.27]	97.81%
Time point [Enrollment - Entry] x TEA (z)	-0.03	[-0.14, 0.09]	68.23%
Time point [Exit - Entry] x TEA (z)	0.03	[-0.06, 0.12]	72.39%
Time point [Follow-up - Entry] x TEA (z)	0.04	[-0.06, 0.13]	76.39%
Time point [Enrollment - Entry] x PPT	-0.08	[-0.34, 0.18]	73.00%
Time point [Exit - Entry] x PPT	0.01	[-0.20, 0.22]	53.94%
Time point [Follow-up - Entry] x PPT	0.08	[-0.16, 0.32]	74.34%
Time point [Enrollment - Entry] x WCST	0.03	[-0.07, 0.13]	75.30%
Time point [Exit - Entry] x WCST	-0.06	[-0.16, 0.02]	92.36%
Time point [Follow-up - Entry] x WCST	0.06	[-0.04, 0.20]	88.61%

Note: CI: Credible Interval, pd: probability of direction, SD: standard deviation, cor: correlation. Full model results are reported in Appendix A

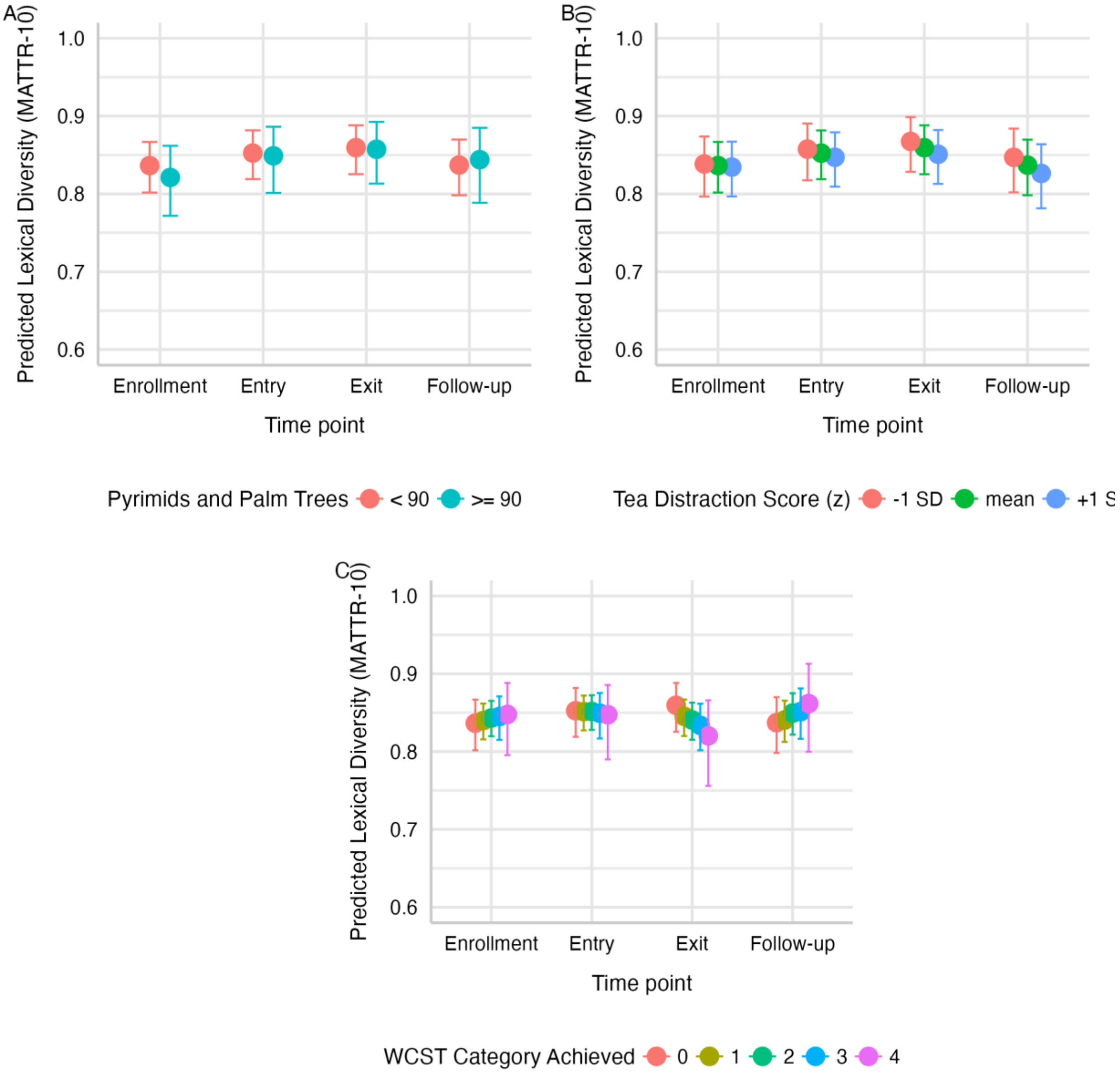


Figure 12. Relationships between lexical diversity outcomes and non-langage cognitive scores

Panel A: Semantic Memory; Panel B: Divided Attention; Panel C: Executive Function

3.2.3 Mean Length of Utterance

For mean length of utterance in words, results are reported in Table 8 and visualized in Figure 13. The intercept estimate was 1.76 (95% CI: [1.53, 2.00]), indicating that the average

MLU at treatment entry was about 5.81. For a participant with average PPT, WCST, and TEA scores, changes between treatment time points were not reliably different from zero. For this average participant, the estimated change from enrollment to treatment entry was -0.01 (95% CI: [-0.12, 0.09]). The estimated change from entry to treatment exit was 0.00 (95% CI: [-0.12, 0.07]) and from entry to 1-month follow-up was, 0.02 (95% CI: [-0.06, 0.10]).

For the TEA, the interaction between TEA performance and change was negligible from enrollment to treatment entry (5.97e-03, 95% CI: [-0.05, 0.06]) and small and unreliable from treatment entry and exit (0.02 95% CI: [-0.03, 0.07]). There was weak evidence for a moderating relationship between TEA performance and change from entry to 1-month follow-up (0.04 95% CI: [-0.01, 0.08]), providing weak evidence (94.55% posterior probability) that attention ability may relate to treatment changes at follow-up.

No reliable findings were present for the Pyramids and Palm Trees from enrollment to entry (0.02, 95% CI: [-0.11, 0.16]), treatment entry to exit (-0.03 95% CI: [-0.15, 0.09]), and treatment entry to follow-up (-0.03 95% CI: [-0.14, 0.09]). Similar null results were evident for the Wisconsin Card Sorting Test, the interaction between WCST performance and change between enrollment and entry (-0.03, 95% CI: [-0.07, 0.03]), treatment entry to exit (-0.02 95% CI: [-0.11, 0.04]), and treatment entry to follow-up (-0.02 95% CI: [-0.08, 0.02]) were not reliability different from zero.

Table 8. Model results for mean length of utterance moderation (population-level effects)

Parameter	Median	95% CI	pd
Intercept	1.76	[1.53, 2.00]	100%
Time point [Enrollment - Entry]	-0.01	[-0.12, 0.09]	58.14%
Time point [Exit - Entry]	-5.80e-03	[-0.12, 0.07]	55.95%
Time point [Follow-up - Entry]	0.02	[-0.06, 0.10]	70.89%
TEA (z)	-2.29e-03	[-0.14, 0.14]	51.28%
PPT	0.10	[-0.27, 0.45]	70.60%
WCST	-0.05	[-0.18, 0.06]	83.30%
CAT T-Score (z)	0.26	[0.11, 0.40]	99.89%
Time point [Enrollment - Entry] x TEA (z)	5.97e-03	[-0.05, 0.06]	58.80%
Time point [Exit - Entry] x TEA (z)	0.02	[-0.03, 0.07]	82.11%
Time point [Follow-up - Entry] x TEA (z)	0.04	[-0.01, 0.08]	94.55%
Time point [Enrollment - Entry] x PPT	0.02	[-0.11, 0.16]	62.10%
Time point [Exit - Entry] x PPT	-0.03	[-0.15, 0.09]	69.55%
Time point [Follow-up - Entry] x PPT	-0.03	[-0.14, 0.09]	67.06%
Time point [Enrollment - Entry] x WCST	-0.03	[-0.07, 0.03]	85.96%
Time point [Exit - Entry] x WCST	-0.02	[-0.11, 0.04]	71.81%
Time point [Follow-up - Entry] x WCST	-0.02	[-0.08, 0.02]	85.89%

Note: CI: Credible Interval, pd: probability of direction, SD: standard deviation, cor: correlation. Full model results are reported in Appendix A

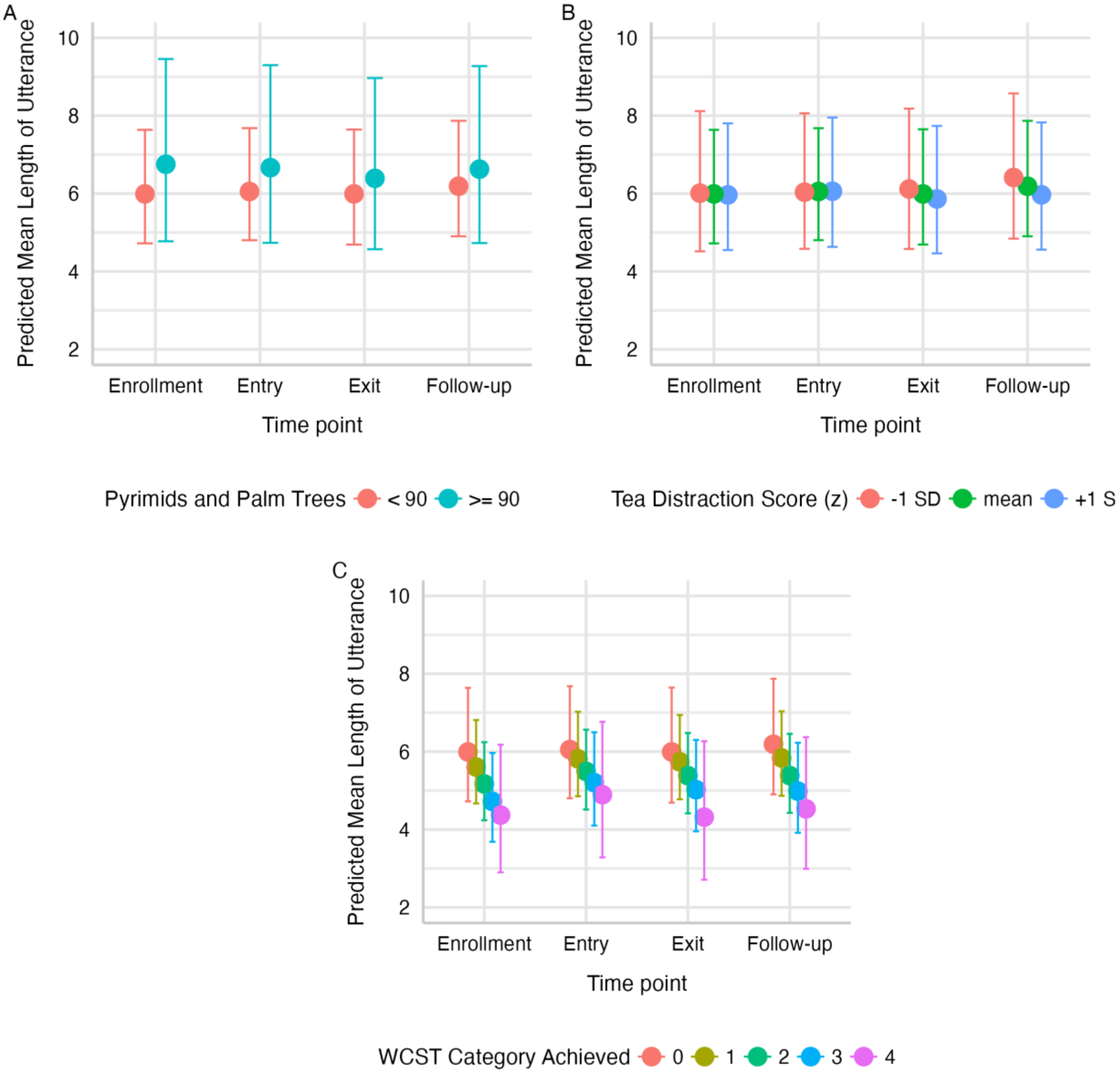


Figure 13. Relationships between mean length of utterance outcomes and non-langage cognitive scores

Panel A: Semantic Memory; Panel B: Divided Attention; Panel C: Executive Function

3.2.4 Discourse Informativeness

For discourse informativeness, results are reported in Table 9 and visualized in Figure 14. The intercept estimate was -0.23 (95% CI: [-0.59, 0.13]), indicating that about 44% of words were scored as CIUs at treatment entry. For a participant with average PPT, WCST, and TEA scores, the estimated change from enrollment to treatment entry was 0.04 (95% CI: [-0.14, 0.22]). The estimated change from treatment entry to treatment exit was 0.02 (95% CI: [-0.15, 0.18]) and from entry to 1-month follow-up was 0.05 (95% CI: [-0.14, 0.23]), with a posterior probability of direction of 70.54%. These results indicate that a participant with average scores on the PPT, TEA, and WCST did not demonstrate a reliable difference in discourse informativeness between time points.

Interactions between TEA performance and change was small and unreliable from enrollment to treatment entry (-0.08, 95% CI: [-0.18, 0.03]), treatment entry and exit (-0.05 95% CI: [-0.14, 0.04]), and from treatment entry to 1-month follow-up (-2.20e-03 95% CI: [-0.10, 0.10]). No reliable findings were present for the Pyramids and Palm Trees from enrollment to entry (-2.91e-03, 95% CI: [-0.29, 0.28]), treatment entry to exit (-0.02 95% CI: [-0.27, 0.23]), and treatment entry to follow-up (-0.14 95% CI: [-0.41, 0.14]).

For the Wisconsin Card Sorting Test, the interaction between WCST performance and change between enrollment and entry (-9.10e-04, 95% CI: [-0.12, 0.09]) was negligible. However, there was a weak interaction between WCST category and change from treatment entry to exit (0.04 95% CI: [-0.04, 0.13]) and a large and more reliable interaction from treatment entry to follow-up (0.14 95% CI: [0.01, 0.37]) with a posterior probability of direction of 98.20%. These findings suggest that the category obtained on the WCST may be related to improvements in discourse informativeness over time, though the effects are modest in size and only apparent at 1-month follow-up.

Table 9. Model results for discourse informativeness moderation (population-level effects)

Parameter	Median	95% CI	pd
Intercept	-0.23	[-0.59, 0.13]	90.56%
phi_Intercept	3.22	[2.66, 3.79]	100%
Time point [Enrollment - Entry]	0.04	[-0.14, 0.22]	64.88%
Time point [Exit - Entry]	0.02	[-0.15, 0.18]	60.85%
Time point [Follow-up - Entry]	0.05	[-0.14, 0.23]	70.54%
TEA (z)	0.09	[-0.09, 0.28]	82.84%
PPT	0.02	[-0.48, 0.48]	52.81%
WCST	-0.26	[-0.67, 0.14]	89.76%
CAT T-Score (z)	0.49	[0.30, 0.68]	100%
Time point [Enrollment - Entry] x TEA (z)	-0.08	[-0.18, 0.03]	91.49%
Time point [Exit - Entry] x TEA (z)	-0.05	[-0.14, 0.04]	87.49%
Time point [Follow-up - Entry] x TEA (z)	-2.20e-03	[-0.10, 0.10]	51.95%
Time point [Enrollment - Entry] x PPT	-2.91e-03	[-0.29, 0.28]	50.98%
Time point [Exit - Entry] x PPT	-0.02	[-0.27, 0.23]	55.95%
Time point [Follow-up - Entry] x PPT	-0.14	[-0.41, 0.14]	83.75%
Time point [Enrollment - Entry] x WCST	0.10	[-0.33, 0.53]	68.42%
Time point [Exit - Entry] x WCST	0.14	[-0.33, 0.59]	71.74%
Time point [Follow-up - Entry] x WCST	1.02e-03	[-0.17, 0.17]	50.62%

Note: CI: Credible Interval, pd: probability of direction, SD: standard deviation, cor: correlation. Full model results are reported in Appendix A

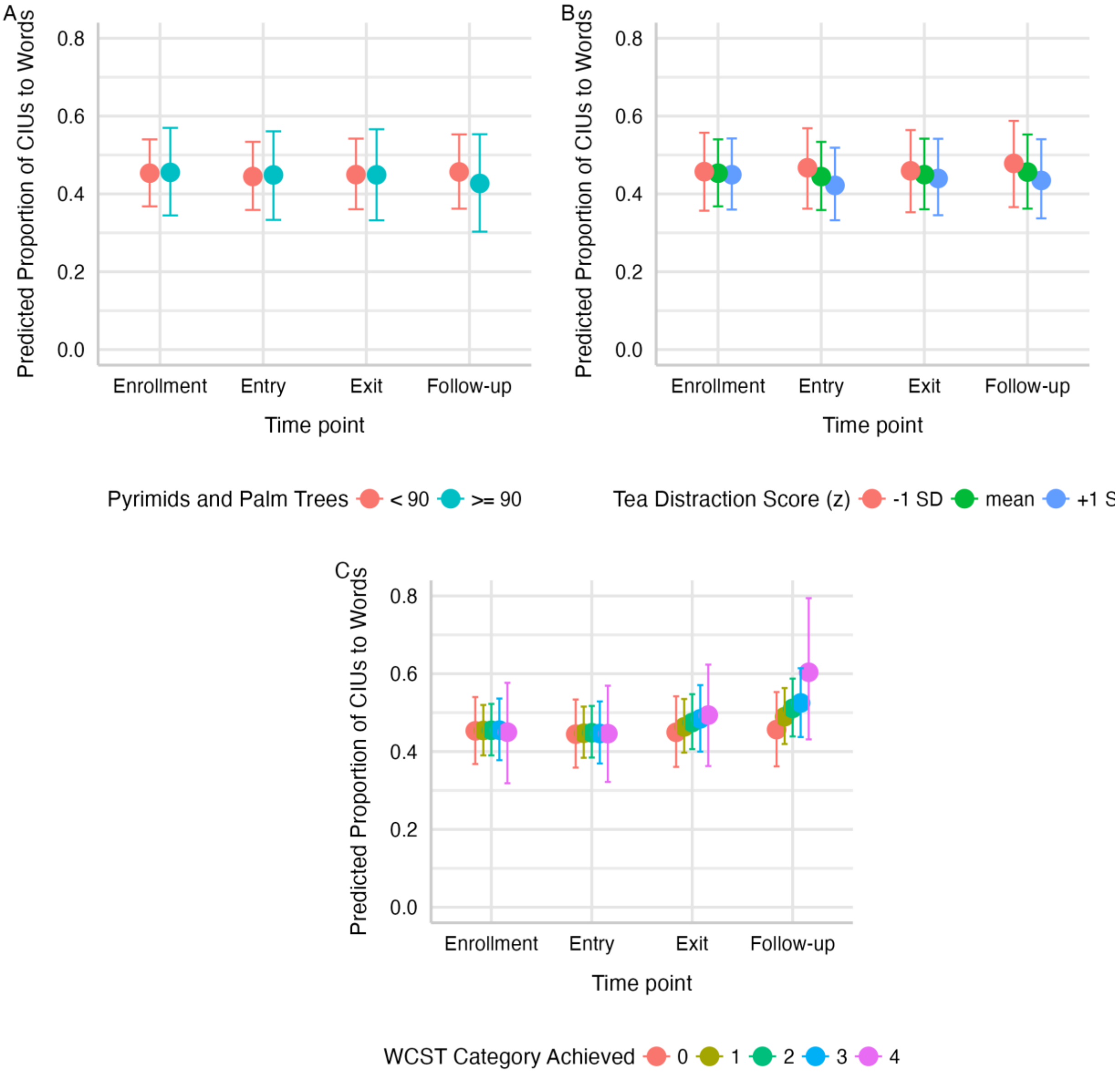


Figure 14. Relationships between discourse informativeness outcomes and non-langage cognitive scores

Panel A: Semantic Memory; Panel B: Divided Attention; Panel C: Executive Function

4.0 Discussion

A central premise of most aphasia treatments is that restoring access to language at the phoneme, word, or sentence level will generalize to more meaningful contexts such as connected speech. However, treatment-related changes in discourse-level communication are generally modest, poorly understood, and vary significantly between individuals with aphasia. This study analyzed archival, monologic discourse outcomes across two simislar clinical trials of intensive SFA (n = 60). A preliminary analysis revealed small but statistically reliable improvements in discourse informativeness in a subset of this sample (n = 44), raising theoretical and clinical questions about potential mechanisms underlying change in monologue discourse performance after SFA.

Aim 1 evaluated the potential contributions of key mechanisms toward discourse outcomes in Semantic Feature Analysis treatment for aphasia through multilevel discourse analysis. Concurrent changes in lexical-semantic processing, lexical diversity, grammatical complexity, and discourse informativeness were evaluated at study enrollment, entry, exit, and follow-up. The working hypothesis was that patterns of improvement across measures would inform the role of restorative vs. compensatory mechanisms on discourse generalization in SFA. Instead, there was no evidence for meaningful or reliable change on any discourse outcome measure from treatment entry to treatment exit or 1-month follow-up, and no evidence for correlated changes across outcome measures. Aim 1 provides little evidence for either mechanism, but this null finding warrants thoughtful introspection of discourse generalization expectations in word-level aphasia treatments.

Aim 2 explored potential non-language cognitive processes which support discourse outcomes in aphasia by examining the potential moderating role of non-language cognitive factors (divided attention, semantic memory, and executive function) on discourse outcomes in a subsample of participants (n = 44). These factors are often associated with successful discourse production and response to behavioral interventions. Null findings in aim 1 do not preclude the possibility of moderating relationships in aim 2; main effects may be null in the presence of strong crossover effects. However, results revealed only weak evidence for moderation effects, which should be interpreted with considerable caution. The following sections discuss these findings in detail.

4.1 No Evidence for Changes in Monologue Discourse Outcomes

Multiple candidate mechanisms of action may underlie SFA's treatment effects. To summarize, the "spreading activation" account of SFA suggests that repeated feature generation has an underlying restorative effect on the semantic system by spreading activation of the features within the semantic network to their associated concepts and, ultimately, to associated lexical items. A modern interpretation under the two-step interactive activation model hypothesizes that the production of target words and semantically related features strengthen the connections between conceptual and lexical representations, ultimately increasing activation from those conceptual representations to other semantically related lexical items (Boyle, 2010). Additionally, evidence from the retrieval practice literature suggests that repeated target and feature generation could result in improvements both to lexical-semantic and semantic-phonological connections (Middleton et al., 2016). Finally, a compensatory account of SFA suggests that SFA improves language production by helping participants habituate self-cueing strategies to improve word finding in instances of anomia (Falconer & Antonucci, 2012).

This study sought to examine the role these mechanisms might play in discourse-level improvements after SFA. The working hypothesis was that improvements across lexical-semantic processing (the rate of semantic errors produced in connected speech), lexical diversity, grammatical complexity, and informativeness would support the role of the primarily restorative mechanisms. Multilevel-discourse theory, and the implicit assumptions of many aphasia treatments, hypothesize that the restorative improvements reflected in lexical-semantic processing and lexical diversity should lead to knock-on effects in grammatical complexity and discourse informativeness, especially given the sentence-level generalization task included in this variant of SFA. This study also proposed a complementary alternative: improvements limited to lexical diversity and informativeness would be more consistent with a compensatory account of SFA. Ultimately, however, the lack of strong evidence for improvements in *any* of the outcome measures for the full sample of 60 participants suggests that the traditional form of SFA has a minimal impact on general-topic discourse-level communication, using the *status quo* measurement paradigm. This appears to be the case despite a high treatment intensity and additional sentence-level treatment component. In the following, these null effects for each outcome measure are discussed in turn.

A lack of evidence for improvements in lexical-semantic processing at the discourse level could be explained by several factors. First, the rate of semantic errors was low to begin with, which makes detecting improvements challenging, even with non-linear models which emphasize changes at the tails of the distribution (i.e., small changes in the percentage of semantic errors to content words would still reflect large changes on the log-odds scale). It is still plausible that SFA could reduce the rate of semantic error production for participants who are more prone to producing them. However, there is no clear improvement in lexical-semantic processing for this heterogenous participant population *on average*.

The challenge of detecting improvements in lexical-semantic processing at the discourse level is exacerbated in participants with more severe aphasia and sparse linguistic output. The probability of producing a single semantic error in a language sample with fewer than ten content words was low across all participants. Moreover, participants with severe aphasia often have comorbid phonological impairments impacting language production, which could obscure potential semantic errors by also impairing phonological selection for incorrectly retrieved lexical items. Stimuli and tasks that elicit longer discourse samples or scaffold spoken language production (e.g., story retell) help to better understand the impact of semantic treatments on semantic error production in connected speech by increasing the number of opportunities for error production.

Similarly, there was no evidence for an increase in lexical diversity, which suggests that participants did not meaningfully increase the range or variety of content words used in response to the monologue discourse elicitation task. As noted earlier, the Nicholas and Brookshire protocol includes stimuli not intentionally related to targets trained in SFA. Thus, one explanation for this finding is that the discourse stimuli used in this study may not sufficiently elicit the use of target words or practiced semantic features to detect improvements in word finding. If the effects of SFA are largely item-specific, it is reasonable to hypothesize that the improvements in lexical diversity might require discourse stimuli that are likely to elicit the same words and semantic features trained during treatment. Given the generally modest and inconsistent response generalization in SFA studies in the confrontation naming context

(Nickels, 2002; Webster et al., 2015), perhaps it is not reasonable to presume that anomia treatments will simultaneously generalize to discourse-level communication (stimulus generalization) and related, untrained words (response generalization; Thompson, 1989). Additionally, because SFA is not expected to produce cross-category generalization to unrelated, untrained words, any stimulus generalization of practiced features would require evidence that discourse elicitation tasks would, in fact, elicit practiced targets and features.

The null finding for lexical diversity comes in the context of more complex analytical techniques for examining changes in the MATTR. This study used a MATTR with a 10-token window and a statistical parameter to adjust for samples with less than ten tokens. This zero-inflated beta modeling approach helps to generate a greater spread in lexical diversity estimates more closely resembling established work (Fergadiotis & Wright, 2011) rather than the minimum window (MATTR-5), which is subject to substantial ceiling effects (Cunningham & Haley, 2020). While there were no changes in lexical diversity estimates, results indicated that participants were somewhat more likely to generate samples long enough to calculate the MATTR-10 immediately following treatment. This may suggest that the treatment protocol did have a small effect on the total amount of content produced by participants with more severe aphasia.

Given the lack of changes in lexical-semantic processing and lexical diversity, robust improvements in grammatical complexity would not be expected under the present hypothesis because such changes were hypothesized to be a downstream consequence of improved word retrieval. Consistent with this claim, there was no evidence to suggest any meaningful changes in MLU, and it remains an open question whether improvements to word retrieval in discourse-level communication might have positive effects on grammatical complexity.

Preliminary data provided evidence for small but statistically reliable improvements to discourse informativeness. However, the present analysis with all 60 participants suggests a smaller and less reliable effect, with neither the credible interval for change at treatment exit nor 1-month follow-up excluding zero. There was weak evidence for improvement at follow-up, with a 96% posterior probability of an effect, though the effect size remained small (0.10 logits). The most likely explanation for this change from preliminary data is modest regression to the mean (i.e., no effect) with a larger sample. It is also possible that the more robust modeling strategy (i.e., using a beta-binomial model to account for any possible overdispersion in the data) resulted in more conservative estimates of precision, which limit strong conclusions for or against any effects.

It is also plausible that minor changes in study design between SFA-1 and SFA-2 could have had an unanticipated effect on discourse informativeness. Originally, this study suggested that the primary difference between studies (that about half of the participants would produce fewer features *per trial*) was orthogonal to the outcome measures and research questions at hand. In retrospect, this may not have been the case. Individuals in the "few features" group might be producing less variety in their features, spending more time on rote retrieval practice of the target and the same features relative to the "many features" group. Even though there is no current evidence for an effect of feature diversity on naming treatment outcomes in SFA (Evans, Cavanaugh, Gravier, et al., 2021), it remains a possibility that this change to the SFA protocol may have had some impact on the discourse outcome measures.

Taken together, the lack of changes across all four outcome measures does not support the role of either candidate mechanism (restorative or compensatory) on general monologue discourse outcomes in SFA. Instead, these results replicate recently published discourse outcomes in SFA, where Silkes and colleagues (2021) found that SFA improved discourse informativeness (using a story retell procedure) by a non-significant three percentage points at treatment exit and that any gains were not maintained at follow-up. The findings are also consistent with Boyle (2011), who reported a positive improvement in discourse outcomes for only 1/5 of participants after SFA. While other reviews have suggested that improvements to word finding might have "knock-on effects for grammatical integrity and macrostructure" (Dipper et al., 2021), any evidence for discourse generalization was notably absent in this study. The central takeaway across these studies is that there is no evidence to support the claim that SFA has meaningful or reliable generalization to general monologue discourse production.

4.2 Challenges in Capturing Change in Monologue Discourse in Aphasia Treatment

Given these null effects, it is crucial to acknowledge the ever-present challenges when studies examine changes in monologue discourse production in aphasia. General challenges in reliably measuring discourse-level communication in aphasia treatments have been well-described and include underspecified or inadequate psychometric properties, the general complexity of discourse measurement, day-to-day and minute-to-minute variability in performance, degrees of freedom in elicitation, transcription, and coding methodology, and small sample sizes (e.g., Boyle, 2020; Bryant et al., 2016; Dietz & Boyle, 2018; Dipper et al., 2021; Kintz & Wright, 2018; Stark et al., 2020; Wallace et al., 2018). These challenges are undoubtedly relevant to the current study, but additional study-specific considerations exist.

One explanation for this study's lack of treatment effect is the (limited) extent to which the Nicholas & Brookshire stimuli protocol elicit the types of improvements expected after SFA or provide participants with opportunities to apply gains in word finding. SFA's classic treatment effects are typically considered to be item- and category- specific, improving performance on trained words and untrained words with close semantic relationships to those trained words (Quique et al., 2019). While some treated semantic categories in SFA-1 and SFA-2 are relevant to some discourse stimuli (e.g., a kitchen utensils category and the cookie theft picture), this is true for only a small subset of treatment lists and discourse stimuli. Moreover, the extent to which stimuli are likely to elicit words in related categories is unclear. For example, relatively few kitchen utensils are depicted in the cookie theft picture, and the salience of the image focuses on the children reaching for the cookie jar and the sink overflowing. Without establishing whether discourse stimuli elicit the trained words and features practiced during SFA, it is possible that any item-specific gains which might be useful in the right context go unnoticed. Additionally, assuming that the discourse stimuli are unlikely to elicit trained words and features, improvements at the discourse level under a restorative mechanism would require substantial response generalization to related, untrained words *and* stimulus generalization to an untrained task (monologue discourse production). The substantial magnitude of response generalization required for this possibility is unlikely, even in the picture naming context (Quique et al., 2019) and the findings of this study suggest that any stimulus generalization, if present, is insufficient to produce such an effect at the discourse level.

An alternative for future treatment studies (which anticipate item-specific effects) is to a-priori establish that discourse elicitation tasks are likely to elicit expected improvements, should they occur. Examples of this approach include the previously described Rider et al. (2008) study, which evaluated the extent to which SFA might improve contextual discourse when treating items frequently produced by individuals without aphasia in response to short videos and procedural discourse prompts. Similarly, Conroy et al. (2018) developed a list of potential treatment targets from a set of discourse elicitation stimuli (Where's Waldo/Wally pictures) from words often produced by individuals without aphasia. Individuals with aphasia then received a retrieval-practice treatment paradigm using words from this "core lexicon." Critically, the authors examined the extent to which these individuals increased how often they produced treated targets on the discourse stimuli from which those targets were derived. In the most comprehensive approach to date, Dipper, Cruice and colleagues (Dipper et al., 2022) developed personalized stimuli from personal narratives and treated those stimuli using a multilevel approach that targeted the use of those targets at the word, sentence, and discourse level, and then examined the extent to which the production of personal narratives improved. Regardless of the approach, applying theoretically motivated discourse outcome measures to language samples elicited by stimuli related to treatment targets is likely to create a clearer test of generalization to connected speech.

While the restorative mechanisms discussed in this study predict item- and category-specific improvements, a compensatory SFA mechanism does not rely on such constraints. Instead, it may be reasonable to hypothesize that individuals with aphasia who internalize and then apply self-cueing strategies learned during SFA do so in any instance of anomia regardless of the semantic context. Thus, the generalization effects from a compensatory mechanism might be more plausible even with the Nicholas & Brookshire stimuli. Unfortunately, this study did not analyze discourse samples for the frequency or success of circumlocution or self-cueing attempts, so there is no direct evidence to indicate whether self-cueing strategies changed due to treatment. A number of small-N studies have examined the role and use of self-cueing in SFA

(Antonucci, 2009; Falconer & Antonucci, 2012; Peach & Reuter, 2010; Tilton-Bolowsky et al., 2022), but larger-scale studies are needed to establish population-level effects.

A second challenge specific to evaluating multilevel discourse outcomes in aphasia is the heterogeneity in discourse profiles at baseline. Not all participants will need to improve on every measure, nor will some participants be capable of making substantial improvements on some measures (e.g., participants with severe and agrammatic aphasia may have difficulty increasing utterance length even if lexical retrieval improves). Additionally, not all measures are bipolar such that "more" is always "better;" changes in the opposite direction of the majority might constitute improvements. For example, individuals with fluent, empty discourse might benefit from decreasing utterance length. Thus, null findings for the *average* treatment effect for a heterogeneous cohort might obscure individual improvements on different measures, depending on participants' baseline profile. Consider two hypothetical examples:

Participants with relatively minimal verbal output (e.g., less than 10-15 tokens/sample) in 1-2 word utterances might be more likely to improve in their lexical diversity or discourse informativeness but may not show any change in the rate of producing semantic errors (particularly if they rarely produce semantic errors, to begin with) or their utterance length. For these participants, their "next adjacent possible" improvement may be to generate a more diverse set of 1-2 word utterances, not to expand the complexity of their utterances. On the other hand, a participant with more fluent output and frequent semantic errors with more typical utterance length might be more likely to reduce their rate of producing semantic errors because there are far more opportunities to make them. While reducing the frequency of semantic errors might improve communication, this effect is unlikely to translate to utterance length, which is already near ceiling.

The current participant sample is highly diverse, as demonstrated by both the scores on the discourse outcome measures and standardized assessments. While this larger sample size is important for statistical precision, it also has the potential to obscure effects that are specific to different subgroups of individuals with aphasia. This problem is less important in clinical settings, where treatments are more easily tailored to the needs of individual patients. A solution in research settings is to conduct research on subpopulations based on stricter eligibility criteria or with sample sizes large enough for subgroup analyses stratified by baseline discourse-level language characteristics. Although this sample is not nearly large enough for subgroup analyses, a post-hoc exploratory analysis tested interactions between aphasia severity and time point for each measure (Appendix E). This post-hoc analysis replicated an analysis conducted with only discourse informativeness on the preliminary data with the full sample (n = 60) and all four discourse outcomes. Across all four models, there were no statistically reliable interactions between aphasia severity and treatment effects, which suggests that the problem posed by heterogeneity may not be so easily captured by overall aphasia severity. Instead, considering baseline discourse profiles, such as those described by Gordon (2020) may be more fruitful.

Discourse outcome measure selection and reliability constitute a third challenge to capturing treatment-induced changes in monologue discourse production. This study’s initial proposal included alternative measures of grammatical complexity (predicate argument structure) as well as a macrostructural measure (global coherence), but these additional measures were abandoned due to concerns about their poor psychometric properties in conjunction with their additional time and labor costs. The measures in this study were selected based on theoretical and pragmatic constraints after considering the wide variety of discourse measures used in the aphasia literature to date (Bryant et al., 2016). While the present argument is that the current outcome measures are sufficiently theoretically tied to the two candidate mechanisms of SFA, it is possible that alternative outcome measures might be more sensitive or better descriptors of SFA’s effects on discourse production and thus more likely to capture restorative or compensatory improvements that result from the treatment.

Scoring reliability is another well-described challenge to capturing changes in discourse production for many discourse measures. This study used a rigorous approach for improving the accuracy, reliability, and consistency of discourse scoring by using two independent raters to score all samples and then resolve all differences - in contrast to examining the reliability of a small fraction of scores to a second rater (Bryant et al., 2016). Even still, there were challenges with scoring discourse samples. For example, while a scoring manual was created a-priori for each outcome measure, it was not possible to anticipate all scoring ambiguities (e.g., rules around utterance boundaries or defining semantic error). The scoring manual Appendix D results from an iterative process documenting additional rules for accounting for ambiguity in the initial coding guidelines.

Thoughtful measure selection and careful scoring cannot resolve the high task-to-task and day-to-day variability in discourse production typical of individuals with aphasia. This variability is expected in everyday communication contexts as well as the discourse monologue context, where researchers at least have some experimental control over the stimulus and task length. It makes cross-sectional and longitudinal measurement challenging. There are two study-specific features that have the potential to impact this variability as well.

First, discourse stimuli were chosen from an established list of 10 stimuli, creating two smaller sets of five stimuli equated for discourse genre. These lists were administered in a pseudorandom order across study time points (e.g., ABAB or ABBA). In retrospect, the inconsistent order of stimuli is an important limitation to any strong conclusions about the lack of item-specific generalization effects of the restorative mechanism because there is no experimental control over the content elicited at each time point. From another viewpoint, analyzing archival discourse data elicited from this implementation of Nicholas and Brookshire protocol is perhaps best viewed as a strict test of *far generalization* to general monologue discourse performance.

While this experimental design decision may help to reduce any inadvertent impact of stimulus-specific repeated practice unrelated to treatment, it also likely introduces additional tasks-specific variability between time points. Stimuli had varying degrees of salience for participants, which likely affected performance on specific lists. Typically, discourse scores across stimuli are averaged before analysis, or analyses are run separately by stimuli. To account for this experimental feature and the variability inherent to both discourse genre and specific discourse stimuli in this study, multilevel models in this study were run at the stimuli level, using the discourse stimuli as a group level (i.e., random effect). This approach explicitly incorporates known stimuli-level variability (Stark, 2019) in the statistical model. In theory, this approach also makes the estimates generated by the population-level effects more generalizable to other, similar discourse stimuli. In future studies, using consistent stimuli across time points is recommended if item-specific effects are possible. There is scant evidence to suggest that there are strong practice effects at the discourse level (Cameron et al., 2010). It’s likely that the limited threat of practice effects is far outweighed by the cost of further increasing performance variability.

Second, measurement certainty and error become particularly challenging as discourse sample lengths decrease, as in the case of individuals with more severe expressive deficits.

Because the inclusion criteria were minimally restrictive, this study include several participants with limited verbal output in connected speech settings. Many discourse measures (including all four measures used in this study) are appropriately adjusted in some way for sample length, which is highly variable and rarely related to the construct at hand. Still, measurement precision typically improves as sample length increases. For example, estimates of an individual's discourse informativeness in a single task are more likely to be closer to their "true" ability when they produce longer samples. This is because the proportion of CIUs to words stabilizes as sample length increases. Accordingly, estimates of informativeness for an individual who produced 10 CIUs out of 20 words should be less certain than if the same individual produced 100 CIUs out of 200 words. Because individuals with more impoverished verbal output typically produce short samples regardless of stimuli, the uncertainty around their scores should be greater. Two of the four measures in this study (lexical-semantic processing and informativeness) explicitly adjust certainty as a function of sample length in the statistical approach (using beta-binomial generalized linear models) but adjusting for sample length was not feasible for MLU and lexical diversity. Moreover, adjusting for sample length statistically ultimately doesn't resolve the challenges created by short samples. Devising future assessments and elicitation strategies that facilitate longer sample lengths is of utmost importance for future research on discourse in aphasia. Future studies may need to consider stricter eligibility criteria for sample length minimums. Alternatively, the clear benefits of tailoring outcome measures to the baseline characteristics underscores the need to pursue small-N and case-series designs to complement group-level trials.

4.3 Discourse Generalization in SFA: Moving Beyond "train and hope"

In addition to the measurement-related challenges in capturing treatment-related changes in monologue discourse production after SFA, there is an important theoretical limitation to the premise that improvements in word-finding after SFA should generalize to monologue discourse. A core concept underlying SFA is that repeated target naming and feature generation will improve lexical access to trained words and semantically related, untrained words. If true, SFA can focus on a limited subset of words to improve lexical access to a broad vocabulary across different semantic categories. It follows that such improvements might be useful in a variety of discourse contexts that bear at least some semantic relationship to the trained semantic categories, provided that individuals with aphasia are able to access those improvements in new contexts. Furthermore, the potential benefits of SFA's compensatory mechanism are not limited to word-finding difficulties within specific semantic categories. They are theoretically useful when individuals with aphasia experience *any* instances of anomia in connected speech. Both candidate mechanisms, however, assume that generalization to connected speech is a passive by-product of treatment; SFA (at least in the classic paradigm) lacks any ingredients theoretically motivated to facilitate generalization to connected speech.

This tenuous assumption of passive generalization to more ecologically valid contexts has been termed "train and hope" in the behavior modification literature. It was heavily criticized in two seminal papers by Stokes and Baer (Stokes & Baer, 1977; Stokes & Osnes, 1989) as the major contributor to poor generalization outcomes in the behavior modification evidence base. "Train and hope" is the idea that generalization is a passive phenomenon that occurs spontaneously after effective training and does not require intentional treatment ingredients focused on generalization. Aphasia treatment studies often use this passive "train and hope'" approach to discourse generalization, invoking theoretical mechanisms for improving the language system at a word or subcomponent level without specifying the mechanisms for changes in connected speech. These studies often examine discourse measures as “secondary outcomes” and assume that generating improvements in the language system (i.e., as measured by primary outcomes like picture naming) will result in improvements at the discourse level. Examples of studies that fit this “train and hope” approach include the present SFA studies of interest, Silkes et al. (2021) for both Phonomotor Treatment and SFA, Ballard and Thompson (1999) in the context of Treatment for Underlying Forms, or Edmonds (2014) in the context of Verb Network Strength Training.

In the standard SFA protocol (Boyle, 2010), there are no explicit treatment ingredients geared towards scaffolding performance during treatment into a connected speech context. In the context of a restorative mechanism, it may not be reasonable to expect the word-level, confrontation-style practice in SFA to transfer passively to lexical retrieval in a connected speech context. Indeed, SFA likely needs to make substantial improvements to the underlying language system to passively generalize to an unrelated monologue-discourse context such the Nicholas and Brookshire protocol. However, this extent of “system restoration” has been evasive in aphasia from both behavioral and neuroscience perspectives (Brady et al., 2016; Schevenels et al., 2020).

Nor is it clear that individuals with aphasia will transfer success with feature generation as a method of self-cueing to connected speech contexts without explicit meta-cognitive or actual practice beyond word-level tasks. A recent pilot study by Tilton-Bolowsky et al. (2022) combined metacognitive strategy training and SFA in order to increase the use of circumlocution strategies across different contexts. They found that participants improved in their naming of trained and untrained words (regardless of semantic relationship) and “gained explicit strategy knowledge and produced more feature words (i.e., engaged in circumlocution) during naming attempts.” Tilton-Bolowsky et al., (2022) provide preliminary evidence that explicit attention to treatment ingredients that focus on generalization, in this case, a metacognitive component, has the potential to improve generalization in SFA.

While there were no meta-cognitive components in SFA-1 and SFA-2, there was an additional treatment component that asked participants to use the target word in a sentence with additional semantically rich context, with cues and scaffolding from a clinician. This component was intended to help foster generalization to connected speech, though it may be more reasonable to assume that benefits may be limited to structured sentence generation tasks. This added ingredient still takes a “train and hope” approach to discourse generalization – relying on passive transfer of treatment effects from the sentence-level to monologue discourse tasks.

Both the behavior modification literature (e.g., P. C. Kendall, 1989; Stokes & Baer, 1977, 1977) and knowledge transfer literature (e.g., Gick & Holyoak, 1987; Patrick, 1992) offer potential strategies and treatment ingredients for facilitating generalization to discourse-level communication. Aphasia treatment studies should consider implementing mediational or meta-cognitive strategies focused on generalization, for example as described by Tilton-Bolowski et al. (2022) or Evans et al. (Evans, Cavanaugh, Quique, et al., 2021). Aphasia studies might increase both stimulus and response variability, for example, varying the stimulus or task demands to better match the variable demands of discourse-level communication. Varying training environments and communication partners, which have been done on a small scale (e.g., Doyle et al., 1989), might help facilitate generalization. Treatments should provide and reinforce opportunities for generalization to the intended context (e.g., Falconer & Antonucci, 2012; Peach

& Reuter, 2010). Treatments might take an eclectic approach to include these strategies or incorporate them systematically by combining elements of increasing linguistic complexity from existing evidence-based interventions. Many of these strategies are likely employed by clinicians informally, but their potential contributions to increasing generalization are poorly understood empirically. These potential avenues for improving treatment efficacy for discourse-level outcomes warrant an extensive program of comparative effectiveness research.

4.4 Potential Moderators of Discourse Generalization in SFA

A common method of examining heterogeneity in a clinical trial is to evaluate theoretically motivated moderators of treatment response, which was the focus of the exploratory Aim 2. Null main effects of treatment do not preclude the existence of potential treatment moderators – it is plausible that the average treatment effect might not be reliably different from zero, but certain individuals are more likely to respond to treatment. To examine potential moderators of treatment response, interactions were added between non-language cognitive variables identified as key, theory-based factors in discourse production (e.g., Sherratt, 2007) and the population-level effect of time point in the statistical model. As an example, assuming a positive treatment effect (even if not statistically reliable), a positive interaction between the TEA distraction score and time point (from entry to exit) would indicate that the immediate effect of treatment was greater for individuals with higher scores on the TEA distraction subtest. In other words, real treatment effects may only be observed for individuals with some measured characteristics. Also key to interpreting the interaction effects is to be aware of any simple effects of the moderating variable, which describes the extent to which differences already exist at the reference level of time point: treatment entry.

By and large, there were few statistically reliable interaction effects, either in the binary interpretation of whether a 95% credible interval excluded zero or in a more probabilistic interpretation of the probability of direction, which indicates the percentage of the posterior distribution that falls on the majority side of zero. The overarching interpretation of the aim 2 statistical models suggests that there was little evidence for any moderating relationships between semantic memory, divided attention, and executive function scores, and treatment outcomes. Furthermore, the effects that were statistically reliable tended to be inconsistent across time points or similar in magnitude to the amount of change between treatment enrollment and entry, precluding any strong conclusions in this exploratory analysis.

For example, the statistically reliable interaction between the TEA distraction subtest score and time point (entry to exit) in the lexical-semantic processing model was similar in magnitude to the interaction for enrollment to entry and appeared to be similar in direction. Visualization of the interaction (Figure 11, Panel B) indicates that differences in performance at entry for different TEA scores are seen neither at enrollment or treatment exit. This pattern is also apparent for the interaction between semantic memory and change in lexical-semantic processing (Figure 11, Panel A). These results suggest that, although the interactions are statistically reliable, the effects are indistinguishable from performance variability between time points.

The pattern of change in lexical-semantic processing as a function of executive function scores from the WCST is more consistent, with a relatively stable relationship between the WCST category and performance from enrollment to treatment exit and increasing separation at treatment follow-up. However, the direction of this effect suggests that individuals who achieved higher categories on the WCST *increased* their rate of semantic error production at 1-month treatment follow-up. A speculative explanation for this pattern is that individuals with greater executive function abilities might be more likely to generalize circumlocution strategies even without metacognitive attention to this strategy, and this could lead to higher rates of semantic "errors" in the process of circumlocution. However, this interpretation is undercut given that this relationship is only apparent at 1-month follow-up but not treatment exit.

Additionally, the opposite pattern was seen for discourse informativeness: participants who achieved higher categories on the WCST demonstrated marginally greater improvement in discourse informativeness from entry to exit and, more reliably, from entry to follow-up. This finding suggests that individuals with more preserved executive function abilities may be more likely to generalize treatment effects from SFA to discourse informativeness. This finding is logical in the context of informativeness - a hybrid micro- and macro-linguistic discourse measure that relies on both successful lexical retrieval and successful awareness and self-monitoring to ensure topicality and relevance to the stimulus. It may be reasonable to think that executive function abilities play an important role in generalization, especially in the maintenance period after treatment.

Additionally, there were almost no reliable relationships between non-language cognitive scores and any of the four discourse measures at treatment entry (the lone exception was weak evidence for a relationship between TEA scores and the lexical-semantic processing measure). This finding runs contrary to multilevel discourse theory as described by Sherratt (2007) and others, which hypothesizes that non-language cognitive factors such as attention, semantic memory, and executive function play an important role in discourse production in general.

However, this finding is consistent with the patterns of results reported by Dutta (2020), who found minimal correlations between verbal executive function scores and micro- and macro-linguistic discourse measures. These effects were adjusted for aphasia severity, so one possible explanation is that the sample size was not sufficient to detect relationships between the non-language cognitive factors and performance at treatment entry over and above the strong relationships between aphasia severity and discourse measures. Still, the absence of any strong simple effects warrants further examination of the relationships between non-language cognitive factors and discourse production. While larger sample sizes might be necessary to detect theoretically important effects, the modest and unreliable effects with the present sample size does call into question whether cognitive moderates of discourse outcomes are large enough to be of clinical interest.

In summary, the inconsistency of interactions across measures and time point comparisons limits any strong conclusions about moderating effects. For this reason, these results should be interpreted with considerable caution. Future studies should seek to examine to moderators of response to treatment at the discourse level in treatments that are more likely to produce greater treatment response on average (thereby increasing statistical power and precision).

4.5 Clinical Implications

Anomia treatments are among the most used approaches in clinical services provided to individuals with aphasia, and SFA is one of the most widely used anomia treatments with a robust research evidence base (Raymer & Roitsch, 2022). This study does not detract from the robust word-level improvements established in anomia treatments including SFA. However, it does warrant careful consideration about how word-level treatments like SFA engender improvements to monologue discourse or other connected speech and everyday communication contexts and how those changes are measured. Clinical researchers, instructors, and clinicians should not assume that word-level treatment effects will passively generalize to connected speech, even with high treatment dosage or additional practice of treatment targets at the sentence level. A “train and hope” approach to generalization is unlikely to be effective in aphasia rehabilitation, as demonstrated by the current findings.

Clinicians often use eclectic treatment approaches synthesized from the treatment literature and adapted to the individual circumstances of their clients. It is imperative that clinicians make an intentional effort to consider how their treatment approaches facilitate generalization of treatment effects to contexts that are prioritized by their clients. There is some guidance (discussed above) on strategies that may be useful in facilitating generalization (Coppens & Patterson, 2017; Thompson, 1989; Webster et al., 2015), though more research is needed in this area. There are a number of treatment studies that have made intentional efforts to improve generalization to and communication at the discourse level for SFA and other treatment approaches, with promising findings (e.g., Cruice et al., 2022; Doyle et al., 1989; Falconer & Antonucci, 2012; Obermeyer et al., 2019; Tilton-Bolowsky et al., 2022; Wambaugh & Martinez, 2000; Whitworth et al., 2015).

The present findings also underscore the need for thoughtful outcome measurement beyond the word level when anomia treatments are used. There is increasing guidance for clinicians who wish to incorporate discourse outcome measurement into their clinical routine (Boyle, 2020; Dalton et al., 2020; Leaman & Archer, 2023). Briefly, clinicians should aim to map treatment ingredients to the choice of discourse outcome measure and elicitation context to ensure that outcome measurement is likely to capture desired changes in connected speech. As noted in this discussion, eliciting discourse samples using the Nicholas and Brookshire protocol is likely incongruent with the potential effects of treatments like SFA, unless the content of these stimuli is intentionally trained. Instead, for treatments with item-specific effects, clinicians should lean towards using personally relevant elicitation tasks intentionally tied to personalized treatment targets to increase the likelihood that treatment effects are captured by an outcome measure and meaningful for clients. Discourse outcome measurement is also challenging in today's clinical practice settings (e.g., Cruice et al., 2020). For situations where discourse measurement is not feasible, clinicians might consider alternative outcome measures that reflect communication in more ecologically valid contexts when evaluating anomia treatment success (e.g., Babbitt et al., 2011; Doedens & Meteyard, 2020; Hula et al., 2015; Lomas et al., 1989).

4.6 Additional Limitations of the Study

Several limitations have been noted throughout the discussion, but two additional limitations specific to this study are worth highlighting. First, this was a secondary analysis of existing data across two treatment studies without a control condition. While examining changes from study enrollment to treatment entry (which typically occurred roughly 1-month apart) provided a glimpse at performance variability in the absence of the SFA intervention, this is not a substitute for randomized treatment designs. For example, randomizing individuals with aphasia into treatment and no-treatment groups or into immediate and delayed treatment groups would improve our ability to draw causal relationships from the present data. On the other hand, given the null effects resulting from this study, the lack of randomized experimental design has little impact on the primary (null) conclusions.

Similarly, discourse performance was measured across multiple stimuli but only once at each time point before and after the intervention. Given the variability in monologue discourse performance, additional probes at each time point, alternative sampling designs (e.g., an interrupted time series design), or additional discourse stimuli would improve our ability to precisely estimate performance at each time point and establish a more robust estimate of change. It is also worth noting that such additional assessments were not feasible in either SFA-1 or SFA-2 given the overall aims of the studies and may not be feasible in many studies where assessment batteries are already time-consuming and burdensome for participants.

5.0 Conclusions

Multiple stakeholder groups, most notably including individuals with aphasia (Wallace, Worrall, Rose, & Le Dorze, 2017; Wallace, Worrall, Rose, Le Dorze, et al., 2017), have stated priorities for improving connected speech abilities and everyday communication. The present study examined improvements in theoretically motivated outcome measures for 60 individuals with aphasia after Semantic Feature Analysis, one of the most frequently used treatments for aphasia in clinical practice settings. There was no evidence for meaningful or statistically reliable improvements in general monologue discourse performance after intensive SFA. Moreover, while there was some statistical evidence for relationships between non-language cognitive skills and treatment outcomes, neither the size nor the consistency of the effects are sufficient to inform treatment theory or clinical decision-making. These findings exist in the context of a dosage that far exceeds most clinical settings (Cavanaugh et al., 2021) and an SFA protocol that included an additional sentence-level treatment component intended to facilitate generalization beyond the word level.

The lack of reliable and meaningful generalization of treatment effects to discourse-level communication are an important replication of the existing literature with a large, well-powered sample, with implications for clinical instruction and practice. Item-specific improvements at the word level cannot be assumed to generalize to discourse-level communication. In other words, SFA doesn’t appear to meaningfully change discourse production the way it has typically been measured in anomia treatments. Additionally, there is a clear need to pair theoretically informed treatments designed to facilitate generalization to discourse with intentional measurement paradigms designed to capture it. Ultimately, improving the impact of aphasia interventions on discourse-level communication will require the integration of these ideas: synergistically improving the rigor of discourse outcome measurement, incorporating principles of generalization into established treatment approaches, and examining variation in individual-level treatment response in the context of larger populations.

Appendix A Full Model Results, Aim 2

Appendix Table 1. Full model results for lexical-semantic processing (Aim 2)

Effects	Parameter	Component	Median	95% CI	pd
Population	Intercept	conditional	-3.95	[-4.66, -3.23]	100%
	Time point [Enrollment - Entry]	conditional	-0.10	[-0.62, 0.43]	65.36%
	Time point [Exit - Entry]	conditional	2.44e-03	[-0.45, 0.42]	50.51%
	Time point [Follow-up - Entry]	conditional	-0.40	[-1.00, 0.09]	94.11%
	TEA (z)	conditional	-0.23	[-0.45, -0.01]	98.00%
	PPT	conditional	0.27	[-0.28, 0.78]	83.99%
	CAT T-Score (z)	conditional	-0.13	[-0.33, 0.07]	90.39%
	Time point [Enrollment - Entry] x TEA (z)	conditional	0.16	[-0.14, 0.47]	85.56%
	Time point [Exit - Entry] x TEA (z)	conditional	0.25	[0.00, 0.50]	97.70%
	Time point [Follow-up - Entry] x TEA (z)	conditional	-0.15	[-0.43, 0.12]	86.98%
	Time point [Enrollment - Entry] x PPT	conditional	-0.24	[-0.97, 0.44]	75.31%
	Time point [Exit - Entry] x PPT	conditional	-0.49	[-1.08, 0.10]	94.80%
	Time point [Follow-up - Entry] x PPT	conditional	-0.85	[-1.56, -0.14]	99.00%
	WCST	conditional	2.33e-03	[-0.18, 0.19]	50.96%
	Time point [Enrollment - Entry] x WCST	conditional	-0.08	[-0.35, 0.17]	74.45%
	Time point [Exit - Entry] x WCST	conditional	0.03	[-0.18, 0.25]	61.48%
	Time point [Follow-up - Entry] x WCST	conditional	0.28	[0.05, 0.55]	98.95%
Group	SD (Intercept: participant)	conditional	0.43	[0.23, 0.68]	100%
	SD (Time point [Enrollment - Entry] : participant)	conditional	0.50	[0.08, 0.98]	100%
	SD (Time point [Exit - Entry] :participant)	conditional	0.21	[0.01, 0.58]	100%
	SD (Time point [Follow-up - Entry] : participant)	conditional	0.16	[0.01, 0.51]	100%

Effects	Parameter	Component	Median	95% CI	pd
	cor (Intercept ~ Time point [Enrollment - Entry]:participant)	conditional	0.49	[-0.24, 0.91]	91.77%
	cor (Intercept ~ Time point [Exit - Entry]:participant)	conditional	-0.42	[-0.91, 0.58]	79.90%
	cor (Time point [Enrollment - Entry] ~ Time point [Exit - Entry]: participant)	conditional	-3.40e-03	[-0.79, 0.77]	50.19%
	cor (Intercept ~ Time point [Follow-up - Entry]:participant)	conditional	0.03	[-0.79, 0.81]	51.82%
	cor (Time point [Enrollment - Entry] ~ Time point [Follow-up - Entry]:participant)	conditional	0.15	[-0.77, 0.85]	60.98%
	cor (Time point [Exit - Entry] ~ Time point [Follow-up - Entry]: participant)	conditional	0.07	[-0.79, 0.84]	55.05%
	SD (Intercept: stimuli)	conditional	0.88	[0.55, 1.58]	100%
Population	phi	distributional	68.47	[48.16, 100.94]	100%
Monotonic	WCST[1]	simplex	0.21	[0.01, 0.70]	100%
	WCST[2]	simplex	0.18	[0.01, 0.66]	100%
	WCST[3]	simplex	0.20	[0.01, 0.69]	100%
	Time point [Enrollment - Entry] x WCST[1]	simplex	0.24	[0.01, 0.73]	100%
	Time point [Enrollment - Entry] x WCST[2]	simplex	0.22	[0.01, 0.69]	100%
	Time point [Enrollment - Entry] x WCST[3]	simplex	0.19	[0.01, 0.67]	100%
	Time point [Exit - Entry] x WCST[1]	simplex	0.19	[0.01, 0.67]	100%
	Time point [Exit - Entry] x WCST[2]	simplex	0.24	[0.01, 0.72]	100%
	Time point [Exit - Entry] x WCST[3]	simplex	0.21	[0.01, 0.70]	100%
	Time point [Follow-up - Entry] x WCST[1]	simplex	0.17	[0.01, 0.66]	100%
	Time point [Follow-up - Entry] x WCST[2]	simplex	0.20	[0.01, 0.69]	100%
	Time point [Follow-up - Entry] x WCST[3]	simplex	0.24	[0.01, 0.73]	100%

Note: CI: Credible Interval, pd: probability of direction, SD: standard deviation, cor: correlation.

Appendix Table 2. Full model results for lexical diversity (Aim 2)

Effects	Parameter	Component	Median	95% CI	pd
Population	Intercept	conditional	1.76	[1.51, 2.01]	100%
	Time point [Enrollment - Entry]	conditional	-0.12	[-0.32, 0.07]	89.96%
	Time point [Exit - Entry]	conditional	0.05	[-0.12, 0.25]	70.23%
	Time point [Follow-up - Entry]	conditional	-0.12	[-0.28, 0.04]	93.89%
	TEA (z)	conditional	0.04	[-0.10, 0.18]	72.91%
	PPT	conditional	-0.03	[-0.38, 0.32]	57.04%
	CAT T-Score (z)	conditional	0.14	[0.00, 0.27]	97.81%
	Time point [Enrollment - Entry] x TEA (z)	conditional	-0.03	[-0.14, 0.09]	68.23%
	Time point [Exit - Entry] x TEA (z)	conditional	0.03	[-0.06, 0.12]	72.39%
	Time point [Follow-up - Entry] x TEA (z)	conditional	0.04	[-0.06, 0.13]	76.39%
	Time point [Enrollment - Entry] x PPT	conditional	-0.08	[-0.34, 0.18]	73.00%
	Time point [Exit - Entry] x PPT	conditional	0.01	[-0.20, 0.22]	53.94%
	Time point [Follow-up - Entry] x PPT	conditional	0.08	[-0.16, 0.32]	74.34%
	WCST	conditional	-9.80e-03	[-0.14, 0.10]	56.71%
	Time point [Enrollment - Entry] x WCST	conditional	0.03	[-0.07, 0.13]	75.30%
	Time point [Exit - Entry] x WCST	conditional	-0.06	[-0.16, 0.02]	92.36%
	Time point [Follow-up - Entry] x WCST	conditional	0.06	[-0.04, 0.20]	88.61%
Group	SD (Intercept: participant)	conditional	0.40	[0.31, 0.53]	100%
	SD (Time point [Enrollment - Entry]: participant)	conditional	0.24	[0.05, 0.40]	100%
	SD (Time point [Exit - Entry]: participant)	conditional	0.08	[0.00, 0.22]	100%
	SD (Time point [Follow-up - Entry]: participant)	conditional	0.11	[0.01, 0.25]	100%
	cor (Intercept ~ Time point [Enrollment - Entry]: participant)	conditional	-0.50	[-0.80, 0.05]	96.70%
	cor (Intercept ~ Time point [Exit - Entry]: participant)	conditional	-0.42	[-0.89, 0.55]	82.14%
	cor (Time point [Enrollment - Entry] ~ Time point [Exit - Entry] x participant)	conditional	0.49	[-0.56, 0.92]	82.31%
	cor (Intercept ~ Time point [Follow-up - Entry]: participant)	conditional	0.07	[-0.67, 0.76]	56.60%

Effects	Parameter	Component	Median	95% CI	pd
	cor (Time point [Enrollment - Entry] ~ Time point [Follow-up - Entry]: participant)	conditional	-0.20	[-0.84, 0.66]	66.55%
	cor (Time point [Exit - Entry] ~ Time point [Follow-up - Entry]: participant)	conditional	-0.05	[-0.82, 0.77]	53.80%
	SD (Intercept: stimuli)	conditional	0.06	[0.01, 0.13]	100%
Population	phi	distributional	39.15	[34.99, 43.58]	100%
Monotonic	WCST1[1]	simplex	0.21	[0.01, 0.69]	100%
	WCST1[2]	simplex	0.17	[0.01, 0.68]	100%
	WCST1[3]	simplex	0.21	[0.01, 0.70]	100%
	Time point [Enrollment - Entry] x WCST1[1]	simplex	0.25	[0.01, 0.73]	100%
	Time point [Enrollment - Entry] x WCST1[2]	simplex	0.22	[0.01, 0.69]	100%
	Time point [Enrollment - Entry] x WCST1[3]	simplex	0.17	[0.01, 0.65]	100%
	Time point [Exit - Entry] x WCST1[1]	simplex	0.21	[0.01, 0.70]	100%
	Time point [Exit - Entry] x WCST1[2]	simplex	0.24	[0.01, 0.72]	100%
	Time point [Exit - Entry] x WCST1[3]	simplex	0.37	[0.02, 0.79]	100%
	Time point [Follow-up - Entry] x WCST1[1]	simplex	0.13	[0.01, 0.57]	100%
	Time point [Follow-up - Entry] x WCST1[2]	simplex	0.11	[0.00, 0.55]	100%
	Time point [Follow-up - Entry] x WCST1[3]	simplex	0.26	[0.01, 0.71]	100%

Note: CI: Credible Interval, pd: probability of direction, SD: standard deviation, cor: correlation

Appendix Table 3. Full model results for mean length of utterance (Aim 2)

Effects	Parameter	Component	Median	95% CI	pd
Population	Intercept	conditional	1.76	[1.53, 2.00]	100%
	Time point [Enrollment - Entry]	conditional	-0.01	[-0.12, 0.09]	58.14%
	Time point [Exit - Entry]	conditional	-5.80e-03	[-0.12, 0.07]	55.95%
	Time point [Follow-up - Entry]	conditional	0.02	[-0.06, 0.10]	70.89%
	TEA (z)	conditional	-2.29e-03	[-0.14, 0.14]	51.28%
	PPT	conditional	0.10	[-0.27, 0.45]	70.60%
	CAT T-Score (z)	conditional	0.26	[0.11, 0.40]	99.89%
	Time point [Enrollment - Entry] x TEA (z)	conditional	5.97e-03	[-0.05, 0.06]	58.80%
	Time point [Exit - Entry] x TEA (z)	conditional	0.02	[-0.03, 0.07]	82.11%
	Time point [Follow-up - Entry] x TEA (z)	conditional	0.04	[-0.01, 0.08]	94.55%
	Time point [Enrollment - Entry] x PPT	conditional	0.02	[-0.11, 0.16]	62.10%
	Time point [Exit - Entry] x PPT	conditional	-0.03	[-0.15, 0.09]	69.55%
	Time point [Follow-up - Entry] x PPT	conditional	-0.03	[-0.14, 0.09]	67.06%
	WCST	conditional	-0.05	[-0.18, 0.06]	83.30%
	Time point [Enrollment - Entry] x WCST	conditional	-0.03	[-0.07, 0.03]	85.96%
	Time point [Exit - Entry] x WCST	conditional	-0.02	[-0.11, 0.04]	71.81%
	Time point [Follow-up - Entry] x WCST	conditional	-0.02	[-0.08, 0.02]	85.89%
Group	SD (Intercept: participant)	conditional	0.41	[0.33, 0.53]	100%
	SD (Time point [Enrollment - Entry]: participant)	conditional	0.11	[0.04, 0.17]	100%
	SD (Time point [Exit - Entry]: participant)	conditional	0.07	[0.00, 0.14]	100%
	SD (Time point [Follow-up - Entry]: participant)	conditional	0.02	[0.00, 0.07]	100%
	cor (Intercept ~ Time point [Enrollment - Entry]: participant)	conditional	3.43e-03	[-0.46, 0.49]	50.52%
	cor (Intercept ~ Time point [Exit - Entry]: participant)	conditional	0.08	[-0.59, 0.71]	60.05%
	cor (Time point [Enrollment - Entry] ~ Time point [Exit - Entry]: participant)	conditional	-0.38	[-0.90, 0.48]	81.62%

Effects	Parameter	Component	Median	95% CI	pd
	cor (Intercept ~ Time point [Follow-up - Entry]: participant)	conditional	-0.05	[-0.82, 0.78]	54.51%
	cor (Time point [Enrollment - Entry] ~ Time point [Follow-up - Entry]: participant)	conditional	-7.45e-03	[-0.81, 0.80]	50.71%
	cor (Time point [Exit - Entry] ~ Time point [Follow-up - Entry]: participant)	conditional	0.20	[-0.75, 0.88]	63.91%
	SD (Intercept: stimuli)	conditional	0.09	[0.05, 0.17]	100%
Sigma	sigma	sigma	0.23	[0.22, 0.24]	100%
Monotonic	WCST[1]	simplex	0.17	[0.01, 0.66]	100%
	WCST[2]	simplex	0.23	[0.01, 0.71]	100%
	WCST[3]	simplex	0.21	[0.01, 0.68]	100%
	Time point [Enrollment - Entry] x WCST[1]	simplex	0.23	[0.01, 0.72]	100%
	Time point [Enrollment - Entry] x WCST[2]	simplex	0.21	[0.01, 0.67]	100%
	Time point [Enrollment - Entry] x WCST[3]	simplex	0.18	[0.01, 0.67]	100%
	Time point [Exit - Entry] x WCST[1]	simplex	0.28	[0.01, 0.78]	100%
	Time point [Exit - Entry] x WCST[2]	simplex	0.17	[0.01, 0.68]	100%
	Time point [Exit - Entry] x WCST[3]	simplex	0.10	[0.00, 0.71]	100%
	Time point [Follow-up - Entry] x WCST[1]	simplex	0.10	[0.00, 0.61]	100%
	Time point [Follow-up - Entry] x WCST[2]	simplex	0.14	[0.01, 0.63]	100%
	Time point [Follow-up - Entry] x WCST[3]	simplex	0.49	[0.02, 0.92]	100%

Note: CI: Credible Interval, pd: probability of direction, SD: standard deviation, cor: correlation

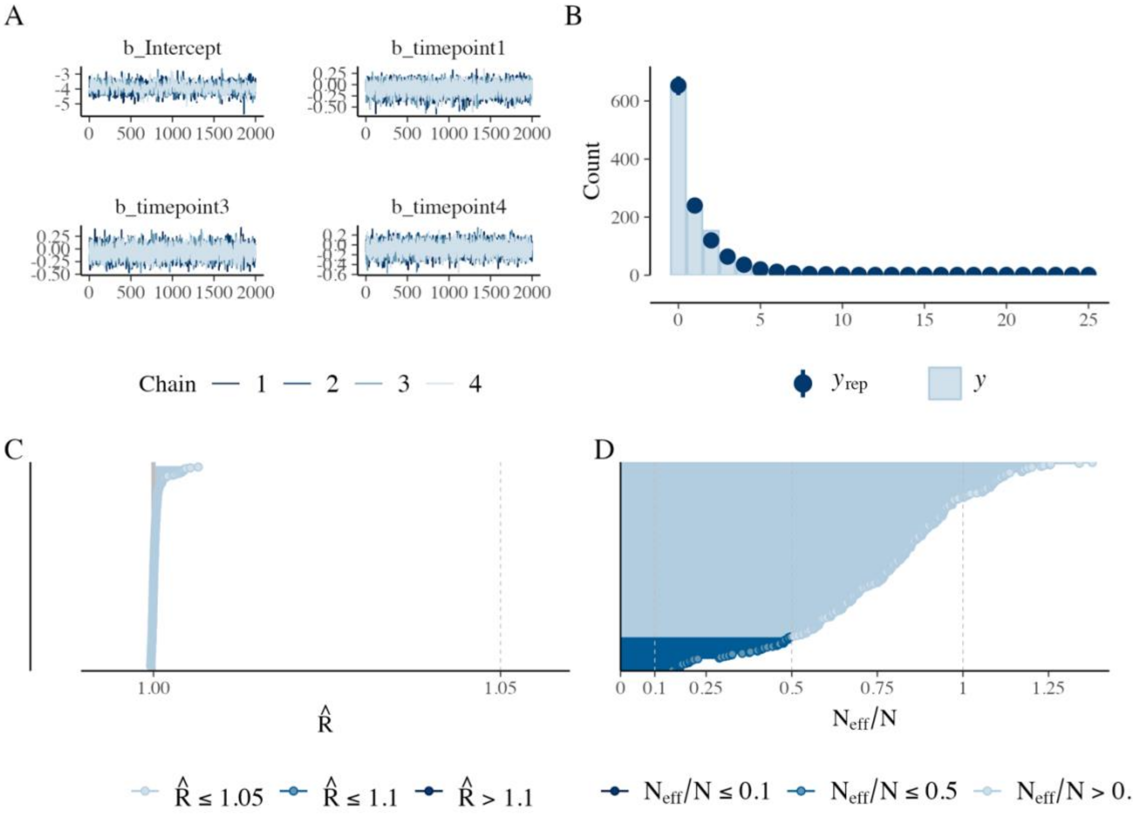
Appendix Table 4 Full model results for mean length of utterance (Aim 2)

Effects	Parameter	Component	Median	95% CI	pd
Population	Intercept	conditional	-0.23	[-0.59, 0.13]	90.56%
	phi_Intercept	conditional	3.22	[2.66, 3.79]	100%
	Time point [Enrollment - Entry]	conditional	0.04	[-0.14, 0.22]	64.88%
	Time point [Exit - Entry]	conditional	0.02	[-0.15, 0.18]	60.85%
	Time point [Follow-up - Entry]	conditional	0.05	[-0.14, 0.23]	70.54%
	TEA (z)	conditional	0.09	[-0.09, 0.28]	82.84%
	PPT	conditional	0.02	[-0.48, 0.48]	52.81%
	CAT T-Score (z)	conditional	0.49	[0.30, 0.68]	100%
	Time point [Enrollment - Entry] x TEA (z)	conditional	-0.08	[-0.18, 0.03]	91.49%
	Time point [Exit - Entry] x TEA (z)	conditional	-0.05	[-0.14, 0.04]	87.49%
	Time point [Follow-up - Entry] x TEA (z)	conditional	-2.20e-03	[-0.10, 0.10]	51.95%
	Time point [Enrollment - Entry] x PPT	conditional	-2.91e-03	[-0.29, 0.28]	50.98%
	Time point [Exit - Entry] x PPT	conditional	-0.02	[-0.27, 0.23]	55.95%
	Time point [Follow-up - Entry] x PPT	conditional	-0.14	[-0.41, 0.14]	83.75%
	phi_Time point [Enrollment - Entry]	conditional	-0.26	[-0.67, 0.14]	89.76%
	phi_Time point [Exit - Entry]	conditional	0.10	[-0.33, 0.53]	68.42%
	phi_Time point [Follow-up - Entry]	conditional	0.14	[-0.33, 0.59]	71.74%
	WCST	conditional	1.02e-03	[-0.17, 0.17]	50.62%
Population (ZI)	Time point [Enrollment - Entry] x WCST	conditional	-9.10e-04	[-0.12, 0.09]	50.69%
	Time point [Exit - Entry] x WCST	conditional	0.04	[-0.04, 0.13]	86.15%
	Time point [Follow-up - Entry] x WCST	conditional	0.14	[0.01, 0.37]	98.20%
Group	SD (Intercept: participant)	conditional	0.56	[0.44, 0.74]	100%
	SD (Time point [Enrollment - Entry]: participant)	conditional	0.14	[0.01, 0.31]	100%
	SD (Time point [Exit - Entry]: participant)	conditional	0.08	[0.00, 0.22]	100%
	SD (Time point [Follow-up - Entry]: participant)	conditional	0.11	[0.01, 0.26]	100%

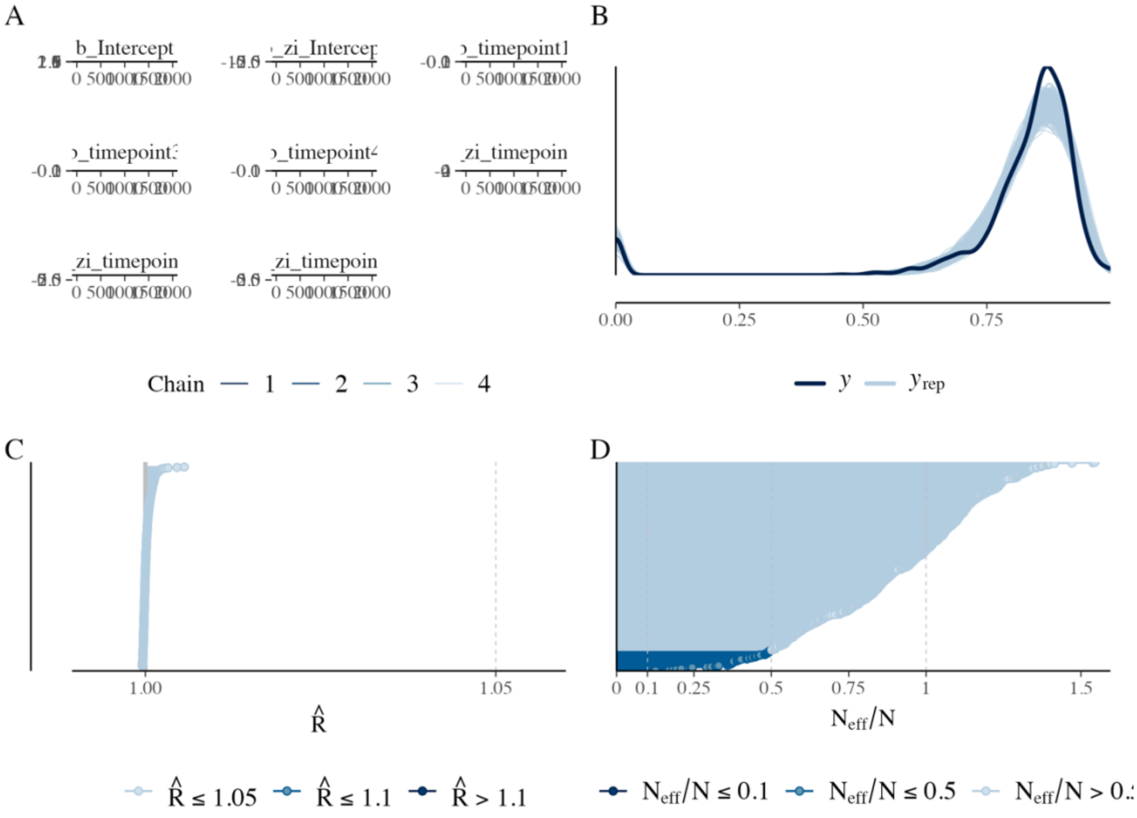
Effects	Parameter	Component	Median	95% CI	pd
	SD (phi_Intercept: participant)	conditional	0.72	[0.46, 1.05]	100%
	cor (Intercept ~ Time point [Enrollment - Entry]: participant)	conditional	-0.35	[-0.82, 0.44]	83.74%
	cor (Intercept ~ Time point [Exit - Entry]: participant)	conditional	0.22	[-0.62, 0.84]	69.53%
	cor (Time point [Enrollment - Entry] ~ Time point [Exit - Entry]: participant)	conditional	0.03	[-0.80, 0.79]	52.09%
	cor (Intercept ~ Time point [Follow-up - Entry]: participant)	conditional	0.41	[-0.47, 0.89]	84.91%
	cor (Time point [Enrollment - Entry] ~ Time point [Follow-up - Entry]: participant)	conditional	-0.12	[-0.83, 0.71]	59.79%
	cor (Time point [Exit - Entry] ~ Time point [Follow-up - Entry]: participant)	conditional	0.22	[-0.70, 0.87]	66.53%
	SD (Intercept: stimuli)	conditional	0.19	[0.11, 0.36]	100%
	SD (phi_Intercept: stimuli)	conditional	0.60	[0.34, 1.16]	100%
Monotonic	WCST1[1]	simplex	0.21	[0.01, 0.71]	100%
	WCST1[2]	simplex	0.18	[0.01, 0.67]	100%
	WCST1[3]	simplex	0.19	[0.01, 0.68]	100%
	Time point [Enrollment - Entry] x WCST1[1]	simplex	0.24	[0.01, 0.74]	100%
	Time point [Enrollment - Entry] x WCST1[2]	simplex	0.20	[0.01, 0.66]	100%
	Time point [Enrollment - Entry] x WCST1[3]	simplex	0.18	[0.01, 0.68]	100%
	Time point [Exit - Entry] x WCST1[1]	simplex	0.19	[0.01, 0.69]	100%
	Time point [Exit - Entry] x WCST1[2]	simplex	0.26	[0.01, 0.77]	100%
	Time point [Exit - Entry] x WCST1[3]	simplex	0.21	[0.01, 0.67]	100%
	Time point [Follow-up - Entry] x WCST1[1]	simplex	0.20	[0.01, 0.69]	100%
	Time point [Follow-up - Entry] x WCST1[2]	simplex	0.22	[0.01, 0.67]	100%
	Time point [Follow-up - Entry] x WCST1[3]	simplex	0.22	[0.01, 0.69]	100%

Note: CI: Credible Interval, pd: probability of direction, SD: standard deviation, cor: correlat

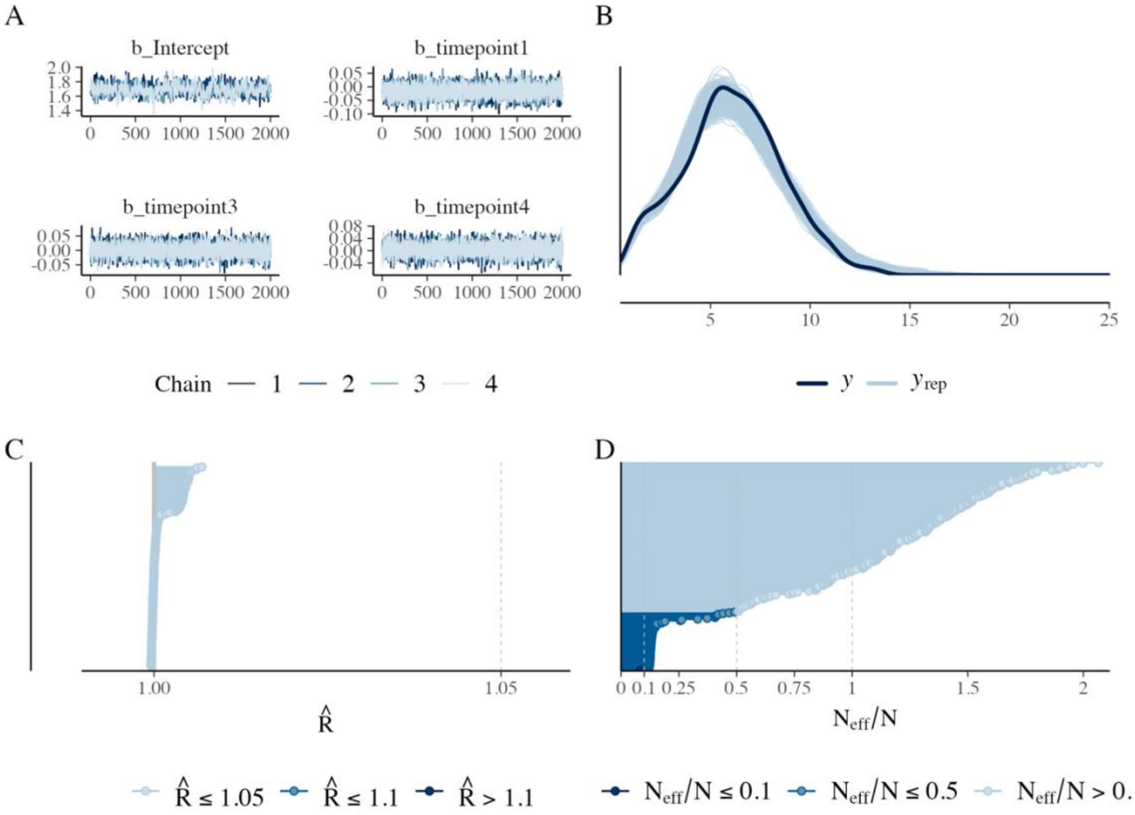
Appendix B Model Diagnostic Plots



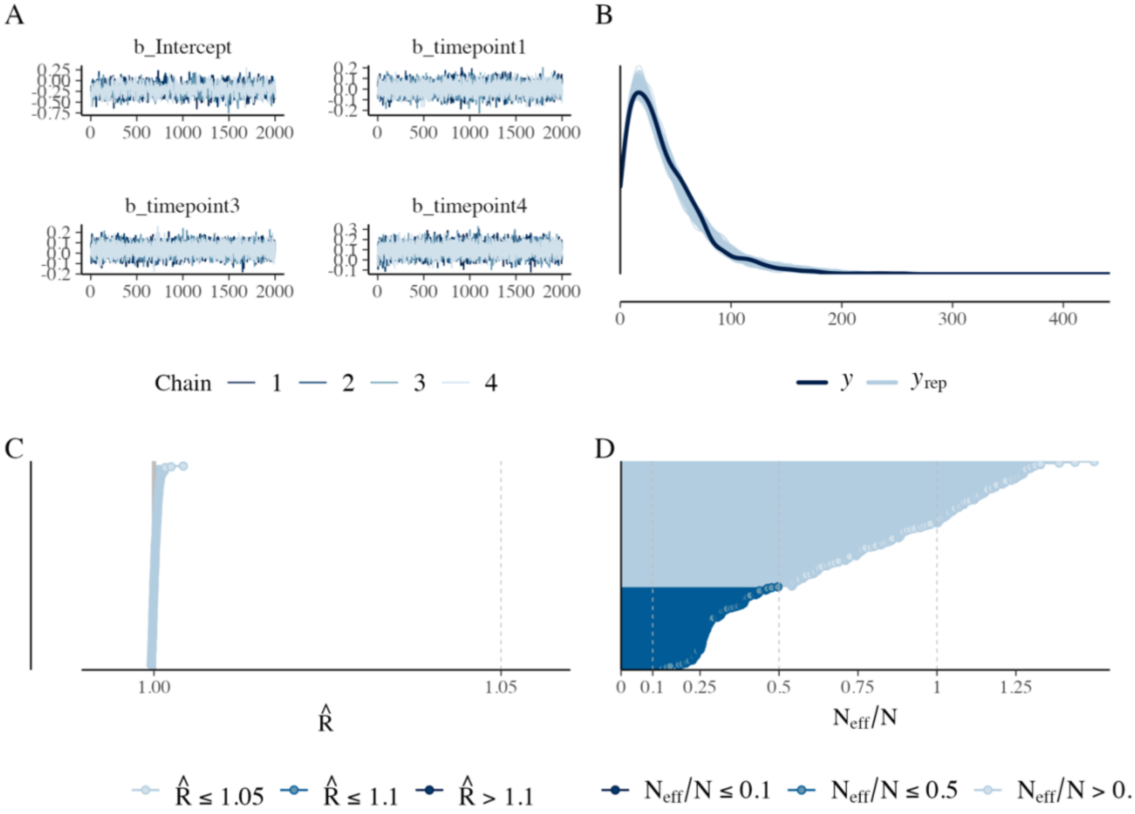
Appendix Figure 1. Diagnostic plots for lexical-semantic processing model Aim 1



Appendix Figure 2. Diagnostic plots for lexical diversity model Aim 1



Appendix Figure 3. Diagnostic plots for mean length of utterance model Aim 1



Appendix Figure 4. Diagnostic plots for discourse informativeness model Aim 1

Appendix C Initial Bayesian Sample Size Estimation

Preliminary data were used to determine the appropriate sample size and statistical power in the Bayesian context, using an accuracy in parameter estimate approach [AIPE; Maxwell et al. (2008) and Bayes Factor Design Analysis (Schönbrodt & Wagenmakers, 2018; Stefan et al., 2019). In this case, fixed and random effect estimates from the preliminary model were used to simulate a 5-variable multivariate model as described above using brms. 500 simulations were run on samples sizes of 44, 50, 55, 60, 65, and 70 to estimate the model convergence rate (as indicated by the split-half potential scale reduction factor <1.01 and absence of divergent transitions) and width of the credible interval around the fixed effect of time for each sample size. At n = 60, model convergence rates exceeded 95% and the credible interval for the fixed effect of time fell below 0.06, indicating that this sample size is sufficient to robustly estimate change in discourse outcomes over time. To ensure that this sample size is adequate for correlation estimation, I conducted a Bayes Factor Design Analysis using the R package BFDA (Schönbrodt, 2018). This approach provides an estimate of the strength of evidence given an expected effect size (similar in principle to the frequentist power analysis) and the rate of misleading evidence when the effect size is zero (i.e., type 1 error rate). Using weakly informative prior distributions over 10,000 simulations, BFDA revealed that 87.8% of simulations provided at least moderate evidence for an effect when r = 0.4 or greater while only 0.4% of simulations provided evidence for a null effect when r = 0.4 (type 2 error). Similarly, BFDA revealed that 86.9% of simulations provided at least moderate evidence for a null effect when r = 0, while only 0.9% of simulations provided evidence for an effect > 0 (type 1 error).

These analyses suggest that n=60 is sufficient to evaluate treatment effects and subsequent correlations.

While this initial analysis plan included multivariate models permitting estimation of correlations of treatment effects across outcome measures, which would describe the extent to which discourse outcome measures changed together. However, this approach was not possible due to the need to use different probability distributions for each outcome measure to best capture the data generation process. Therefore, the final analysis plan was simplified to include multiple univariate models, one for each outcome measure.

Appendix D Discourse Scoring Codebook

Samples were orthographically transcribed in Microsoft Word due to institutional software constraints. Transcription markers were used to annotate orthographic transcriptions in accordance with the following rules:

{} - Neologism, non-linguistic filler, part-word rep with <50% correct morphemes

Examples:
- {uh uh um}
- {s}{s} six
- {oh} if it's a linguistic filler (but not interjection)
- {yeah yeah yeah um yeah} (note: does not include linguistic fillers)

() - Metalinguistic/commentary on the task.

Examples:
- (oh I've seen this before) – commenting on the task that they're being asked to do is familiar
- (I guess) –unsure if sample is good enough or correct
- (I don't know, oh my goodness) – commenting on their inability to produce a target
- (What a mess) – i.e., I'm not doing this very well
- (yeah) – confirming yes that's what I meant to say
- (Oh I'm supposed to be talking about a letter right)
- (Well let me think about this)
- (Start from the beginning) i.e., Let me start the task over*;* does not apply if indicating a single utterance repair
- (what do you call it)
- (amen, done) – i.e., signaling they are done with the task.

Don't mark the following as metalinguistic:
- "I don't know what they did" [to cause them to be arguing]. This reflects the participant communicating ambiguity about the image and is not commentary on the task.
- "I don't know where they're going" in response to the directions picture – Similarly communicates something implied and is not commentary on the task
- "Oh my goodness" [the candle fell over] reflecting the consequences of a candle falling over –
- "What a mess" [car crashes]

- "I can see…[participant refers to something in the picture]". Because it affects the grammatical complexity of the utterance if excluded, keep it in. Example: "I can see a man flying a kite" – the entire utterance should be kept in and not put in parentheses. There is an exception if every utterance begins with the same carrier phrase.

[] - Repetition or repaired utterance
Examples:

- He went to the [fair] fair.
- Exception: when a repeated word is used for emphasis. "He said no no no!"
- [To go get the tree_cat] (or) to get the cat from the tree
- Well [the dog] the dog found out
- and the firemen [came from the] came from [the] {um} the fire truck
- [And the muh_man] and the man just [didn't] didn't calf_care
- they'd be dried with a {uh uh} [(what's that)] [that's a {uh} {t} trowel_towel] [trowel_towel] towel
- To {uh} wash the dishes [put them] get them in the kitchen

_ - Marks target utterance for errors when known or surmised from context (underscore)

Examples:

- She was washing fishes_dishes

:s - Marks semantic error when target known or surmised from context.

All real word errors that have a possible semantic relationship w/ the target as semantic. Follow PNT rules for semantic errors as a general rule. When in doubt, lean towards marking as semantic, to ensure that the item is discussed if marked inconsistently across raters.

Examples:

- There was a dog_cat:s in the tree
- He_she:s is washing dishes

Note: mark mixed errors (per PNT rules) as regular errors (fishes_dishes, as above).

Segmenting Utterances

Follow QPA rules on segmenting (Saffran, Berndt, & Schwartz, 1989, p. 468-473; excerpts adapted below). Compound sentences (with two grammatically complete utterances) should be split

- Priority 1: "Unless there are strong indicators otherwise, a grammatically complete sentence should be an utterance (see note on splitting compound utterances)"
- Priority 2: "Falling intonation suggests end of utterance"
- Priority 3: "Pauses are can be helpful, but their reliability depends on the person. In some cases (frequent pauses, non-fluent aphasia), pauses are likely to be unhelpful."

Making decisions in the grey areas: "The overall pattern of a patient's productions (e.g., pausal patterns, semantic paraphasias) must be considered when bracketing utterances. In all cases, however, utterance boundaries should be drawn conservatively: when in doubt, place boundaries to create shorter rather than longer utterances."

Appendix Table 5. Segmenting examples for discoures transcription

Guideline	Text	Split (if applicable)
Split two grammatically complete ideas connected by "and"	She was washing dishes and he was mowing the lawn	She was washing dishes / and he was mowing the lawn
Keep together if two verb-objects use the same subject	She was washing dishes and looking outside	She was washing dishes and looking outside
Split if *repeated* grammatically complete ideas connected by a conjunction and share a subject	There is a cat and a man and a tree and a tricycle and a bird on a branch	There is a cat / and a man / and a tree / and a tricycle / and a bird on a branch
Second example of above	I write a letter and put a stamp on it and bring it to the post office and send it and hope it gets there and finished	I write a letter / and put a stamp on it / and bring it to the post office / and send it / and hope it gets there / and (finished).
Negations can be considered metacommentary when in repairs	I drove the truck no the car	I drove [the truck_car:s] (no) the car.
Negations not considered metacommentary when in part of an utterance	I drove the …not car…that's the truck	I drove the not car_truck:s / that's the truck.
Metacommentary doesn't require its own utterance (but can be segmented with prior or following utterances, or on its own, because it just gets removed).	And we go into the the the the um p p um just a minute its called um washing the dishels	and we go into [the] [the] [the] the {um} {p} {p} {um} /(just a minute) it's called {um} washing the dishels_dishes

Appendix E Interactions Between Treatment Time Point and Aphasia Severity

Outcome Measure	Parameter	Median	95% CI	pd
Lexical-semantic Processing	Time point [Enrollment - Entry] x CAT T-Score (z)	0.23	[-0.02, 0.49]	96.47%
	Time point [Exit - Entry] x CAT T-Score (z)	0.03	[-0.22, 0.28]	59.86%
	Time point [Follow-up - Entry] x CAT T-Score (z)	0.11	[-0.15, 0.37]	79.92%
Lexical Diversity	Time point [Enrollment - Entry] x CAT T-Score (z)	0.01	[-0.08, 0.12]	61.99%
	Time point [Exit - Entry] x CAT T-Score (z)	0.02	[-0.06, 0.10]	70.73%
	Time point [Follow-up - Entry] x CAT T-Score (z)	-0.01	[-0.10, 0.08]	61.27%
Mean Length Utterance	Time point [Enrollment - Entry] x CAT T-Score (z)	-1.11e-03	[-0.05, 0.04]	52.04%
	Time point [Exit - Entry] x CAT T-Score (z)	-0.02	[-0.07, 0.02]	83.85%
	Time point [Follow-up - Entry] x CAT T-Score (z)	-0.03	[-0.07, 0.01]	92.65%
Informativeness	Time point [Enrollment - Entry] x CAT T-Score (z)	-0.01	[-0.11, 0.09]	60.11%
	Time point [Exit - Entry] x CAT T-Score (z)	-5.75e-03	[-0.10, 0.09]	54.25%
	Time point [Follow-up - Entry] x CAT T-Score (z)	0.08	[-0.03, 0.18]	92.86%

Appendix F Model Syntax and R Session Information

The following appendix provides a general record of statistical model code.

Setup

These are the main packages used in the analysis.

```
# Load packages
library(tidyverse)
library(brms)
library(easystats)
library(here)
library(tidybayes)
library(ggdist)
library(modelr)
library(targets)
```

Aim 1

Preview the data structure for context using fake data generated from the dataset.

```
# Read data
df <- read.csv(here("data", "example-data.csv"))
kable(df)
```

participant	timepoint	stimuli	cius	words	sem_errors	total_content_words	mattr	mlu
1	1	cookie	44	137	5	70	0.87	10.00
2	1	letter	30	217	3	105	0.80	6.64
3	1	dishes	97	134	4	76	0.84	7.94
4	1	argument	71	132	7	84	0.85	4.78
5	1	cat	6	20	4	19	0.64	1.18
6	1	argument	15	26	6	15	0.71	4.80

The same timepoint contrasts were used for all models, setting entry as the reference.

```
# Set contrasts
contrasts(df$timepoint) = contr.treatment(4, base = 2)
```

Lexical-semantic Processing (% Semantic Errors)

```
# Prior
bprior <- c(prior(normal(0, 1), class = "b"))

# Model
model.semE =
  brm(
    bf(
      sem_errors | trials(total_content_words) ~ timepoint +
        (timepoint|participant) + (1|stimuli)
    ),
    family = beta_binomial(),
    data = df,
    backend = "cmdstan",
    cores = 4,
    chains = 4,
    warmup = 1000,
    iter = 3000,
    prior = bprior,
    control = list(adapt_delta = 0.9)
  )
```

Lexical Diversity (MATTR-10)

```
# Priors
priors = c(
  c(prior(normal(0, 1), class = b),
    prior(normal(0, 3), class = b, dpar = zi)))

# Model
model.ld =
  brm(
    bf(
      mattr.z ~ timepoint + (timepoint|participant) + (1|stimuli),
      zi ~ timepoint + (timepoint|participant) + (1|stimuli)
    ),
    family = zero_inflated_beta(),
    data = df,
    backend = "cmdstan",
    cores = 4,
    chains = 4,
    warmup = 1000,
    iter = 3000,
    prior = priors,
    control = list(adapt_delta = 0.9)
  )
```

Grammatical Complexity (Mean Length of Utterance)

```
# Priors
bprior <- c(prior(normal(0, 1), class = "b"),
            prior(normal(2, 2), class = "Intercept"))

# Model
model.mlu =
  brm(
    bf(
      mlu ~ timepoint + (timepoint|participant) + (1|stimuli)
      ),
    family = lognormal(),
    data = df,
    backend = "cmdstan",
    cores = 4,
    chains = 4,
    warmup = 1000,
    iter = 3000,
    prior = bprior
  )
```

Discourse Informativeness (% CIUs)

```
# Priors
priors = c(prior(normal(0,1), class = b))

# Model
mod.nb =
  brm(
   bf(
    cius | trials(words) ~ timepoint + (timepoint|participant) + (1|stimuli)
   ),
   data = df,
   family = beta_binomial(),
   cores = 4,
   warmup = 1000,
   iter = 3000,
   prior = priors,
   chains = 4,
   control = list(adapt_delta = 0.9),
   backend = "cmdstan"
  )
```

Aim 2

Preview the data structure for context using fake data generated from the dataset.

```
# read in data
df <- read.csv(here("data", "example-data-mods.csv"), colClasses = chr)
kable(df %>% rename(" ... " = elipsis))
```

participant	timepoint	stimuli	…	cat_t.z	tea_distraction_scaled.z	ppt	wcst_cat
1	1	directions		0.77	1.11	0.93	1
2	1	letter		1.95	0.18	0.9	2
3	1	letter		-0.04	-0.17	0.9	4
4	1	argument		0.99	-0.62	0.88	0
5	1	directions		-0.78	1.56	0.96	2
6	1	picnic		1.02	-0.24	0.89	1

- The same timepoint contrasts were used for all models, setting entry as the reference.
- WCST categories 3 and 4 were binned into 3 and converted to an ordered factor.
- PPT was categorized as impaired (< 0.9) or not.
- TEA is centered and scaled

```
# set contrasts
contrasts(df$timepoint) = contr.treatment(4, base = 2)

df$wcst_cat <- ifelse(df$wcst_cat > 3, 3, mods$wcst_cat)
df$wcst_cat = factor(df$wcst_cat, ordered = TRUE)
df$ppt_bin = as.factor(ifelse(df$ppt < .9, 0.5, -0.5))
```

Lexical-semantic Processing (% Semantic Errors)

```
# Priors
bprior <- c(prior(normal(0, 1), class = "b"))
# Model
mod.semE.2 =
  brm(
    bf(
      sem_errors | trials(total_content_words) ~
        timepoint*mo(wcst_cat) +
        timepoint*tea_distraction_scaled.z +
        timepoint*ppt_bin +
        cat_t.z +
        (timepoint|participant) + (1|stimuli)
    ),
    family = beta_binomial(),
    data = df,
    backend = "cmdstan",
    cores = 4, chains = 4,
    warmup = 1000, iter = 3000,
    prior = bprior,
    control = list(adapt_delta = 0.9)
  )
```

Lexical Diversity (MATTR-10)

```
# Priors
bprior <- c(prior(normal(0, 1), class = "b"))
# Model
mod.ld.2 =
  brm(
    bf(
      mattr.z ~
        timepoint*mo(wcst_cat) +
        timepoint*tea_distraction_scaled.z +
        timepoint*ppt_bin +
        cat_t.z +
        (timepoint|participant) + (1|stimuli)
    ),
    family = zero_inflated_beta(),
    data = df,
    backend = "cmdstan",
    cores = 4, chains = 4,
    warmup = 1000, iter = 3000,
    prior = bprior,
    seed = 42,
    control = list(adapt_delta = 0.9)
  )
```

Grammatical Complexity (Mean Length of Utterance)

```r
# Priors
bprior <- c(prior(normal(0, 1), class = b))
# Model
mod.mlu.2 =
  brm(
    bf(
      mlu ~
        timepoint*mo(wcst_cat) +
        timepoint*tea_distraction_scaled.z +
        timepoint*ppt_bin +
        cat_t.z +
        (timepoint|participant) + (1|stimuli)
    ),
    family = shifted_lognormal(),
    data = df,
    backend = "cmdstan",
    cores = 4, chains = 4,
    warmup = 1000, iter = 3000,
    prior = bprior,
    control = list(adapt_delta = 0.93)
  )
```

Discourse Informativeness (% CIUs)

```r
# Priors
bprior <- c(prior(normal(0, 1), class = b))
# Model
mod.nb.2 =
  brm(
    bf(
      cius | trials(words) ~
        timepoint*mo(wcst_cat) +
        timepoint*tea_distraction_scaled.z +
        timepoint*ppt_bin +
        cat_t.z +
        (timepoint|participant) + (1|stimuli),
      phi ~ timepoint + (1|participant) + (1|stimuli)
    ),
    data = df,
    family = beta_binomial(),
    cores = 4, chains = 4,
    warmup = 1000, iter = 3000,
    prior = bprior,
    control = list(adapt_delta = 0.9),
    backend = "cmdstan"
  )
```

Post-hoc Models

Post-hoc models estimating interactions between time point and aphasia severity used the same structure as Aim 1, with an addition population-level effect of aphasia severity, and interaction between aphasia severity and time point. A generalized form of the model structure was:

```
dependent_var ~ timepoint*cat_t.z + (timepoint|participant) + (1|stimuli)
```

Session Information

```
## ─ Session info ───────────────────────────────────────────────────────────────
##  setting  value
##  version  R version 4.2.2 (2022-10-31)
##  os       macOS Monterey 12.6
##  system   aarch64, darwin20
##  ui       X11
##  language (EN)
##  collate  en_US.UTF-8
##  ctype    en_US.UTF-8
##  tz       America/New_York
##  date     2023-03-16
##  pandoc   2.19.2
##
## ─ Packages ───────────────────────────────────────────────────────────────────
##  package     * version     date (UTC) lib source
##  bayestestR  * 0.13.0.2    2022-11-30 [1] easystats.r-universe.dev (R 4.2.2)
##  brms        * 2.18.0      2022-09-19 [1] CRAN (R 4.2.0)
##  correlation * 0.8.3       2022-10-09 [1] CRAN (R 4.2.0)
##  datawizard  * 0.6.5       2022-12-14 [1] CRAN (R 4.2.2)
##  dplyr       * 1.0.10      2022-09-01 [1] CRAN (R 4.2.0)
##  easystats   * 0.6.0       2022-11-29 [1] CRAN (R 4.2.0)
##  effectsize  * 0.8.2.00004 2022-12-05 [1] easystats.r-universe.dev (R 4.2.2)
##  flextable   * 0.8.3       2022-11-06 [1] CRAN (R 4.2.0)
##  forcats     * 0.5.2       2022-08-19 [1] CRAN (R 4.2.0)
##  ggdist      * 3.2.0       2022-07-19 [1] CRAN (R 4.2.0)
##  ggplot2     * 3.4.0       2022-11-04 [1] CRAN (R 4.2.0)
##  here        * 1.0.1       2020-12-13 [1] CRAN (R 4.2.0)
##  insight     * 0.18.8.2    2022-11-26 [1] easystats.r-universe.dev (R 4.2.2)
##  knitr       * 1.41        2022-11-18 [1] CRAN (R 4.2.0)
##  modelbased  * 0.8.6       2023-01-13 [1] CRAN (R 4.2.0)
##  modelr      * 0.1.10      2022-11-11 [1] CRAN (R 4.2.0)
##  officer     * 0.4.4.006   2022-09-07 [1] Github (davidgohel/officer@66a360b)
##  parameters  * 0.20.1      2023-01-11 [1] CRAN (R 4.2.0)
##  performance * 0.10.2      2023-01-12 [1] CRAN (R 4.2.0)
##  purrr       * 1.0.1       2023-01-10 [1] CRAN (R 4.2.0)
##  Rcpp        * 1.0.9       2022-07-08 [1] CRAN (R 4.2.0)
##  readr       * 2.1.3       2022-10-01 [1] CRAN (R 4.2.0)
##  report      * 0.5.5.3     2022-12-01 [1] easystats.r-universe.dev (R 4.2.2)
##  see         * 0.7.4.1     2022-11-27 [1] easystats.r-universe.dev (R 4.2.2)
##  stringr     * 1.5.0       2022-12-02 [1] CRAN (R 4.2.0)
##  targets     * 0.14.1      2022-11-29 [1] CRAN (R 4.2.0)
##  tibble      * 3.1.8       2022-07-22 [1] CRAN (R 4.2.0)
##  tidybayes   * 3.0.2       2022-01-05 [1] CRAN (R 4.2.0)
##  tidyr       * 1.2.1       2022-09-08 [1] CRAN (R 4.2.0)
##  tidyverse   * 1.3.2       2022-07-18 [1] CRAN (R 4.2.0)
##
##  [1] /Library/Frameworks/R.framework/Versions/4.2-arm64/Resources/library
##
## ──────────────────────────────────────────────────────────────────────────────
```

References

Allen, C. M., Martin, R. C., & Martin, N. (2012). Relations between Short-term Memory Deficits, Semantic Processing, and Executive Function. *Aphasiology, 26*(3–4), 428–461. https://doi.org/10.1080/02687038.2011.617436

Andreetta, S., Cantagallo, A., & Marini, A. (2012). Narrative discourse in anomic aphasia. *Neuropsychologia*, *50*(8), 1787–1793.

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

Babbitt, E. M., Heinemann, A. W., Semik, P., & Cherney, L. R. (2011). Psychometric properties of the Communication Confidence Rating Scale for Aphasia (CCRSA): Phase 2. *Aphasiology*, *25*(6–7), 727–735. https://doi.org/10.1080/02687038.2010.537347

Ballard, K. J., & Thompson, C. K. (1999). Treatment and generalization of complex sentence production in agrammatism. *Journal of Speech, Language, and Hearing Research*, *42*(3), 690–707.

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

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

Boyle, M. (2011). Discourse treatment for word retrieval impairment in aphasia: The story so far. *Aphasiology*, *25*(11), 1308–1326.

Boyle, M. (2020). Choosing discourse outcome measures to assess clinical change. *Seminars in Speech and Language*, *41*(01), 001–009.

Boyle, M., Akers, C. M., Cavanaugh, R., Hula, W. D., Swiderski, A. M., & Elman, R. J. (2022). Changes in discourse informativeness and efficiency following communication-based group treatment for chronic aphasia. *Aphasiology*, 1–35.

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.

Brady, M. C., Godwin, J., Enderby, P., Kelly, H., & Campbell, P. (2016). Speech and language therapy for aphasia after stroke. *Stroke*, *47*(10), e236–e237. https://doi.org/10.1161/STROKEAHA.116.014439

Brogan, E., Godecke, E., & Ciccone, N. (2020). Behind the therapy door: What is “usual care” aphasia therapy in acute stroke management? *Aphasiology*, *0*(0), 1–23. https://doi.org/10.1080/02687038.2020.1759268

Brookshire, R. H., & Nicholas, L. E. (1994). Test-retest stability of measures of connected speech in aphasia. *Clinical Aphasiology*, *22*, 119–133.

Bryant, L., Ferguson, A., & Spencer, E. (2016). Linguistic analysis of discourse in aphasia: A review of the literature. In *Clinical Linguistics and Phonetics*. https://doi.org/10.3109/02699206.2016.1145740

Bunker, L. D., Wright, S., & Wambaugh, J. L. (2018). Language changes following combined aphasia and apraxia of speech treatment. *American Journal of Speech-Language Pathology*, *27*(1S), 323–335. https://doi.org/10.1044/2018_AJSLP-16-0193

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101.

Cahana-Amitay, D., & Jenkins, T. (2018). Working memory and discourse production in people with aphasia. *Journal of Neurolinguistics*, *48*, 90–103. https://doi.org/10.1016/j.jneuroling.2018.04.007

Cameron, R. M., Wambaugh, J. L., & Mauszycki, S. C. (2010). Individual variability on discourse measures over repeated sampling times in persons with aphasia. *Aphasiology*, *24*(6–8), 671–684. https://doi.org/10.1080/02687030903443813

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1).

Cavanaugh, R., Kravetz, C., Jarold, L., Quique, Y., Turner, R., & Evans, W. S. (2021). Is There a Research–Practice Dosage Gap in Aphasia Rehabilitation? *American Journal of Speech-Language Pathology*, *30*(5), 2115–2129. https://doi.org/10.1044/2021_AJSLP-20-00257

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

Conroy, P., Sotiropoulou Drosopoulou, C., Humphreys, G. F., Halai, A. D., & Lambon Ralph, M. A. (2018). Time for a quick word? The striking benefits of training speed and accuracy of word retrieval in post-stroke aphasia. *Brain*, *141*(6), 1815–1827. https://doi.org/10.1093/brain/awy087

Coppens, P., & Patterson, J. (2017). Generalization in aphasiology: What are the best strategies. *Aphasia Rehabilitation: Clinical Challenges*, 205–248.

Covington, M. A. (2007). *MATTR user manual*. University of Georgia Artificial Intelligence Center.

Cruice, M., Aujla, S., Bannister, J., Botting, N., Boyle, M., Charles, N., Dhaliwal, V., Grobler, S., Hersh, D., Marshall, J., & others. (2022). Creating a novel approach to discourse treatment through coproduction with people with aphasia and speech and language therapists. *Aphasiology*, *36*(10), 1159–1181.

Cruice, M., Botting, N., Marshall, J., Boyle, M., Hersh, D., Pritchard, M., & Dipper, L. (2020). UK speech and language therapists' views and reported practices of discourse analysis in aphasia rehabilitation. *International Journal of Language & Communication Disorders*, *55*(3), 417–442.

Cunningham, K. T., & Haley, K. L. (2020). Measuring lexical diversity for discourse analysis in aphasia: Moving-average type–token ratio and word information measure. *Journal of Speech, Language, and Hearing Research*, *63*(3), 710–721.

Dalton, S. G. H., Hubbard, H. I., & Richardson, J. D. (2020). Moving toward non-transcription based discourse analysis in stable and progressive aphasia. *Seminars in Speech and Language*, *41*, Article 01.

Dean, M. P., Della Sala, S., Beschin, N., & Cocchini, G. (2017). Anosognosia and self-correction of naming errors in aphasia. *Aphasiology*, *31*(7), 725–740.

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

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

Dietz, A., & Boyle, M. (2018). Discourse measurement in aphasia: Consensus and caveats. *Aphasiology*, *32*(4), 487–492.

Dipper, L., Botting, N., Boyle, M., Hersh, D., Marshall, J., & Cruice, M. (2022). *Reporting on LUNA, a novel discourse intervention for people with mild and moderate aphasia – primary efficacy outcomes*. International Aphasia Rehabilitation Conference, Philadelphia, PA.

Dipper, L., Marshall, J., Boyle, M., Botting, N., Hersh, D., Pritchard, M., & Cruice, M. (2020). Treatment for improving discourse in aphasia: A systematic review and synthesis of the evidence base. *Aphasiology*, 1–43.

Dipper, L., Marshall, J., Boyle, M., Hersh, D., Botting, N., & Cruice, M. (2021). Creating a Theoretical Framework to Underpin Discourse Assessment and Intervention in Aphasia. *Brain Sciences*, *11*(2), 183. https://doi.org/10.3390/brainsci11020183

Doedens, W., & Meteyard, L. (2020). Measures of functional, real-world communication for aphasia: A critical review. *Aphasiology*, *34*(4), 492–514.

Doyle, P. J., Goldstein, H., Bourgeois, M. S., & Nakles, K. O. (1989). Facilitating generalized requesting behavior in Broca's aphasia: An experimental analysis of a generalization training procedure. *Journal of Applied Behavior Analysis*, *22*(2), 157–170.

Duffy, J. R. (2013). *Motor Speech Disorders—E-Book: Substrates, Differential Diagnosis, and Management*. Elsevier Health Sciences. https://books.google.com/books?id=ATARAAAAQBAJ

Dutta, M. (2020). *Evaluating the relationship between executive functioning, spoken discourse, and life participation in aphasia [Unpublished Dissertation]*. Indiana University.

Edmonds, L. A., Mammino, K., & Ojeda, J. (2014). Effect of Verb Network Strengthening Treatment (VNeST) in Persons With Aphasia: Extension and Replication of Previous Findings. *American Journal of Speech-Language Pathology*, *23*(May). https://doi.org/10.1044/2014

Efstratiadou, E. A., Papathanasiou, I., Holland, R., Archonti, A., & Hilari, K. (2018). A systematic review of semantic feature analysis therapy studies for aphasia. In *Journal of Speech, Language, and Hearing Research*. https://doi.org/10.1044/2018_JSLHR-L-16-0330

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

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

Falconer, C., & Antonucci, S. M. (2012). Use of semantic feature analysis in group discourse treatment for aphasia: Extension and expansion. *Aphasiology*, *26*(1), 64–82.

Fergadiotis, G., Kellough, S., & Hula, W. D. (2015). Item response theory modeling of the Philadelphia Naming Test. *Journal of Speech, Language, and Hearing Research*, *58*(3), 865–877.

Fergadiotis, G., & Wright, H. H. (2011). Lexical diversity for adults with and without aphasia across discourse elicitation tasks. *Aphasiology*, *25*(11), 1414–1430.

Fergadiotis, G., Wright, H. H., & West, T. M. (2013). Measuring lexical diversity in narrative discourse of people with aphasia. *American Journal of Speech-Language Pathology*.

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

Frederiksen, C. H., Bracewell, R. J., Breuleux, A., & Renaud, A. (1990). The cognitive representation and processing of discourse: Function and dysfunction. In *Discourse ability and brain damage* (pp. 69–110). Springer.

Gabry, J., & Češnovar, R. (2020). *cmdstanr: R interface to "CmdStan"* (0.3.0).

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

Gick, M. L., & Holyoak, K. J. (1987). The Cognitive Basis of Knowledge Transfer. In S. M. Cormier & J. D. Hagman (Eds.), *Transfer of Learning* (pp. 9–46). Academic Press. https://doi.org/10.1016/B978-0-12-188950-0.50008-4

Gilmore, N., Meier, E. L., Johnson, J. P., & Kiran, S. (2019). Nonlinguistic Cognitive Factors Predict Treatment-Induced Recovery in Chronic Poststroke Aphasia. *Archives of Physical Medicine and Rehabilitation, 100*(7), 1251–1258. https://doi.org/10.1016/j.apmr.2018.12.024

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

Gordon, J. K. (2020). Factor analysis of spontaneous speech in aphasia. *Journal of Speech, Language, and Hearing Research*, *63*(12), 4127–4147.

Grant, D. A., & Berg, E. A. (1948). The Wisconsin Card Sort Test: Directions for administration and scoring. *Odessa: Psychological Assessment*.

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

Harris Wright, H., & Capilouto, G. J. (2012). Considering a multi-level approach to understanding maintenance of global coherence in adults with aphasia. *Aphasiology, 26*(5), 656–672.

Hilari, K. (2011). The impact of stroke: Are people with aphasia different to those without? *Disability and Rehabilitation*, *33*(3), 211–218. https://doi.org/10.3109/09638288.2010.508829

Howard, D., & Patterson, K. E. (1992). *The Pyramids and Palm Trees Test: A test of semantic access from words and pictures*. Thames Valley Test Company.

Hula, W. D., Doyle, P. J., Stone, C. A., Austermann Hula, S. N., Kellough, S., Wambaugh, J. L., Ross, K. B., Schumacher, J. G., & St Jacque, A. (2015). The aphasia communication outcome measure (ACOM): Dimensionality, item bank calibration, and initial validation. *Journal of Speech, Language, and Hearing Research*, *58*(3), 906–919. https://doi.org/10.1044/2015_JSLHR-L-14-0235

Hula, W. D., Fergadiotis, G., & Martin, N. (2012). *Model choice and sample size in item response theory analysis of aphasia tests*.

Kendall, D. L., Moldestad, M. O., Allen, W., Torrence, J., & Nadeau, S. E. (2019). Phonomotor versus semantic feature analysis treatment for anomia in 58 persons with aphasia: A randomized controlled trial. *Journal of Speech, Language, and Hearing Research*, *62*(12), 4464–4482. https://doi.org/10.1044/2019_JSLHR-L-18-0257

Kendall, P. C. (1989). The generalization and maintenance of behavior change: Comments, considerations, and the “no-cure” criticism. *Behavior Therapy*, *20*(3), 357–364.

Kintz, S., & Wright, H. H. (2018). Discourse measurement in aphasia research. *Aphasiology*, *32*(4), 472–474.

Lambon Ralph, M. A., Snell, C., Fillingham, J. K., Conroy, P., & Sage, K. (2010). Predicting the outcome of anomia therapy for people with aphasia post CVA: both language and cognitive status are key predictors. *Neuropsychological Rehabilitation*, *20*(2), 289–305. https://doi.org/10.1080/09602010903237875

Larfeuil, C., & Dorze, G. L. (1997). An analysis of the word-finding difficulties and of the content of the content of the discourse of recent and chronic aphasic speakers. *Aphasiology*, *11*(8), 783–811.

Leaman, M. C., & Archer, B. (2023). Choosing discourse types that align with person-centered goals in aphasia rehabilitation: A clinical tutorial. *Perspectives of the ASHA Special Interest Groups*, 1–20.

Lomas, J., Pickard, L., Bester, S., Elbard, H., Finlayson, a, & Zoghaib, C. (1989). The communicative effectiveness index: Development and psychometric evaluation of a functional communication measure for adult aphasia. *The Journal of Speech and Hearing Disorders*, *54*(1), 113–124.

Marini, A., Andreetta, S., Del Tin, S., & Carlomagno, S. (2011). A multi-level approach to the analysis of narrative language in aphasia. *Aphasiology*, *25*(11), 1372–1392.

Martin, N., Schwartz, M. F., & Kohen, F. P. (2006). Assessment of the ability to process semantic and phonological aspects of words in aphasia: A multi-measurement approach. *Aphasiology*, *20*(02–04), 154–166.

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

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

Meyer, A., Wheeldon, L., & Krott, A. (2007). *Automaticity and control in language processing* (Vol. 1). Psychology Press.

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

Morey, R. D., & Rouder, J. N. (2022). *BayesFactor: Computation of bayes factors for common designs* [Manual]. https://CRAN.R-project.org/package=BayesFactor

Nicholas, L. E., & Brookshire, R. H. (1993). A System for Quantifying the Informativeness and Efficiency of the Connected Speech of Adults With Aphasia. *Journal of Speech and Hearing Research*, *36*(April), 338–350.

Nickels, L. (2002). Therapy for naming disorders: Revisiting, revising, and reviewing. *Aphasiology*, *16*(10–11), 935–979. https://doi.org/10.1080/02687030244000563

Obermeyer, J. A., Rogalski, Y., & Edmonds, L. A. (2019). Attentive Reading with Constrained Summarization-Written, a multi-modality discourse-level treatment for mild aphasia. *Aphasiology*, 1–26.

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.

Olsson, C., Arvidsson, P., & Blom Johansson, M. (2019). Relations between executive function, language, and functional communication in severe aphasia. *Aphasiology*, *33*(7), 821–845.

Patrick, J. (1992). *Training: Research and practice.* (pp. xx, 561). Academic Press.

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

Pedersen, P. M., Vinter, K., & Olsen, T. S. (2004). Aphasia after stroke: Type, severity and prognosis. The Copenhagen aphasia study. *Cerebrovascular Diseases*, *17*(1), 35–43. https://doi.org/10.1159/000073896

Pritchard, M., Hilari, K., Cocks, N., & Dipper, L. (2018). Psychometric properties of discourse measures in aphasia: Acceptability, reliability, and validity. *International Journal of Language & Communication Disorders*, *53*(6), 1078–1093.

Purdy, M., & Koch, A. (2006). Prediction of strategy usage by adults with aphasia. *Aphasiology*, *20*(02–04), 337–348.

Quique, Y., 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), 1. https://doi.org/10.1044/2018_AJSLP-17-0155

R Core Team. (2020). *R: A language and environment for statistical computing* (4.0.3). R Foundation for Statistical Computing. https://www.r-project.org/

Ramsberger, G. (2005). Achieving conversational success in aphasia by focusing on non-linguistic cognitive skills: A potentially promising new approach. *Aphasiology*, *19*(10–11), 1066–1073.

Raymer, A. M., & Roitsch, J. (2022). Word retrieval treatments in aphasia: A survey of professional practice. *Aphasiology*, 1–26.

Rider, J. D., Wright, H. H., Marshall, R. C., & Page, J. L. (2008). Using semantic feature analysis to improve contextual discourse in adults with aphasia. *American Journal of Speech-Language Pathology / American Speech-Language-Hearing Association*, *17*(2), 161–172. https://doi.org/10.1044/1058-0360(2008/016)

Robertson, I. H., Ward, T., Ridgeway, V., Nimmo-Smith, I., & others. (1994). The test of everyday attention (TEA). *Bury St. Edmunds, UK: Thames Valley Test Company*, 197–221.

Saffran, E. M., Berndt, R. S., & Schwartz, M. F. (1989). The quantitative analysis of agrammatic production: Procedure and data. *Brain and Language*, *37*(3), 440–479.

Schevenels, K., Price, C. J., Zink, I., De Smedt, B., & Vandermosten, M. (2020). A review on treatment-related brain changes in aphasia. *Neurobiology of Language*, *1*(4), 402–433.

Sherratt, S. (2007). Multi-level discourse analysis: A feasible approach. *Aphasiology*, *21*(3–4), 375–393.

Silkes, J. P., Fergadiotis, G., Graue, K., & Kendall, D. L. (2021). Effects of phonomotor therapy and semantic feature analysis on discourse production. *American Journal of Speech-Language Pathology*, *30*(1S), 441–454.

Simic, T., Chambers, C., Bitan, T., Stewart, S., Goldberg, D., Laird, L., Leonard, C., & Rochon, E. (2020). Mechanisms underlying anomia treatment outcomes. *Journal of Communication Disorders*, *88*, 106048.

Simmons-Mackie, N. (2018). Aphasia in North America. *Aphasia Access*.

Stark, B. C. (2019). A comparison of three discourse elicitation methods in aphasia and age-matched adults: Implications for language assessment and outcome. *American Journal of Speech-Language Pathology*, *28*(3), 1067–1083.

Stark, B. C., Dutta, M., Murray, L. L., Bryant, L., Fromm, D., MacWhinney, B., Ramage, A. E., Roberts, A., den Ouden, D. B., Brock, K., & others. (2020). Standardizing assessment of spoken discourse in aphasia: A working group with deliverables. *American Journal of Speech-Language Pathology*, 1–12.

Stokes, T. F., & Baer, D. M. (1977). An implicit technology of generalization 1. *Journal of Applied Behavior Analysis*, *10*(2), 349–367.

Stokes, T. F., & Osnes, P. G. (1989). An operant pursuit of generalization. *Behavior Therapy*, *20*(3), 337–355.

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

Thompson, C. K. (1989). Generalization research in aphasia: A review of the literature. *Clinical Aphasiology*, *18*, 195–222.

Thompson, C. K. (2006). Single subject controlled experiments in aphasia: The science and the state of the science. *Journal of Communication Disorders*, *39*(4), 266–291.

Tilton-Bolowsky, V., Hoffman, La., Evans, W. S., & Vallila-Rohter, S. (2022). *Incorporating metacognitive strategy training into semantic treatment promotes generalization in naming and improves functional communication in chronic aphasia*. 61st Annual Academy of Aphasia, Philadelphia, PA.

Villard, S., & Kiran, S. (2017). To what extent does attention underlie language in aphasia? *Aphasiology*, *31*(10), 1226–1245. https://doi.org/10.1080/02687038.2016.1242711

Wallace, S. J., Worrall, L. E., Rose, T., & Le Dorze, G. (2018). Discourse measurement in aphasia research: Have we reached the tipping point? A core outcome set… or greater standardisation of discourse measures? *Aphasiology*, *32*(4), 479–482.

Wallace, S. J., Worrall, L., Rose, T., & Le Dorze, G. (2017). Which treatment outcomes are most important to aphasia clinicians and managers? An international e-Delphi consensus study. *Aphasiology*, *31*(6), 643–673. https://doi.org/10.1080/02687038.2016.1186265

Wallace, S. J., Worrall, L., Rose, T., Le Dorze, G., Cruice, M., Isaksen, J., Kong, A. P. H., Simmons-Mackie, N., Scarinci, N., & Gauvreau, C. A. (2017). Which outcomes are most important to people with aphasia and their families? An international nominal group technique study framed within the ICF. *Disability and Rehabilitation*, *39*(14), 1364–1379. https://doi.org/10.1080/09638288.2016.1194899

Wambaugh, J. L., & Martinez, A. L. (2000). Effects of modified response elaboration training with apraxic and aphasic speakers. *Aphasiology*, *14*(5–6), 603–617.

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), 334–370. https://doi.org/10.1044/1058-0360(2013/12-0070)

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

Whitworth, A., Leitao, S., Cartwright, J., Webster, J., Hankey, G., Zach, J., Howard, D., & Wolz, V. (2015). NARNIA: a new twist to an old tale. A pilot RCT to evaluate a multilevel approach to improving discourse in aphasia. *Aphasiology*, *29*(11), 1345–1382.