

Determinants of Multilevel Discourse Outcomes in Anomia Treatment for Aphasia

Robert Cavanaugh, Michael Walsh Dickey, William D. Hula, Davida Fromm, Jennifer Golovin, Julie Wambaugh, Gerasimos Fergadiotis, William S. Evans

Affiliations:

Robert Cavanaugh: Northeastern University, Portland, ME; University of Pittsburgh, VA Pittsburgh Healthcare System, Pittsburgh PA

Jennifer Golovin: VA Pittsburgh Healthcare System, Pittsburgh PA

Davida Fromm: Carnegie Mellon University, Pittsburgh PA

Julie Wambaugh: University of Utah, Salt Lake City, UT

Gerasimos Fergadiotis: Portland State University, Portland, OR

William D. Hula: VA Pittsburgh Healthcare System, University of Pittsburgh, Pittsburgh PA

Michael Walsh Dickey: University of Pittsburgh, VA Pittsburgh Healthcare System, Pittsburgh PA

William S. Evans: University of Pittsburgh, VA Pittsburgh Healthcare System, Pittsburgh PA

Funding: This research was supported by the following awards: 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 views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs, the National Institutes of Health, or the United States government.

Disclosures: The authors have no conflicts of interest to disclose

Purpose: 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 among 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. Aim 1 evaluated changes in theoretically motivated discourse outcomes representing lexical-semantic processing, lexical diversity, grammatical complexity, and discourse informativeness. Aim 2 explored the potential moderating role of non-language cognitive factors (semantic memory, divided attention, and executive function) on discourse outcomes.

Methods: This study was a retrospective analysis of archival monologic discourse outcomes after intensive SFA for $n = 60$ (Aim 1) and a subset $n = 44$ (Aim 2). Outcome measures included lexical-semantic processing (% semantic errors), lexical diversity (moving-average type-token ratio), grammatical complexity (mean utterance length), and discourse informativeness (% correct information units). Bayesian generalized mixed-effects models were used to examine changes across four study timepoints: enrollment, entry, exit, and 1-month follow-up.

Results: The present study found no evidence for meaningful or statistically reliable improvements in monologue discourse performance after Semantic Feature Analysis when measured using standard, general topic discourse stimuli. There was weak and inconsistent evidence that non-language cognitive factors may play a role in moderating treatment response.

Discussion: These findings indicate a clear need to pair theoretically informed treatments designed to facilitate generalization to discourse with intentional measurement paradigms designed to capture it. 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.

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). Impairment-focused clinical interventions for aphasia aim to ameliorate underlying linguistic deficits, with the intention that improvements will generalize to discourse, or language use above the sentence-level (Thompson, 2006). However, discourse outcomes have been modest and highly variable to date, which may be due in part to the lack of a clear theoretical basis for such generalization (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.

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 a common anomia treatment in clinical research (e.g., Boyle, 2010; Efstratiadou et al., 2018; Quique et al., 2019) and clinical practice (Raymer & Roitsch, 2022). SFA comprises interleaved, effortful retrieval of target words and semantic features in several semantic categories, typically superordinate, use, action, properties, location, and personal association (Boyle & Coelho, 1995).

SFA's foundational mechanism of action is based on a spreading activation model of semantic processing (Collins & Loftus, 1975). Elicitation and production of semantic features is thought to spread activation to associated concepts within the semantic network 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. In the context of an alternative model, the two-step interactive activation model of lexical access (Foygel & Dell, 2000), repeated production of target words and semantically related features might strengthen the connections between conceptual and lexical representations. Reinforcement of these connections could subsequently spread activation between semantically related items in the lexical-semantic network. Or perhaps repeated feature generation and naming of target items improves the resting activation for both the target item and other items within a semantic category. Regardless, the core prediction of these restorative hypotheses is that SFA will improve lexical access to both trained and semantically related, untrained words. If true, SFA need only focus on a limited subset of words in order to improve lexical access to a broader range of semantically related words, potentially leading to widespread improvements in word finding.

Reviews and meta-analyses of the effects of SFA on naming of both treated and semantically related, untreated words have mostly substantiated this pattern of improvements. 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. 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 individuals with milder aphasia were more likely to improve after SFA. Group-level studies have largely corroborated these findings (Evans, Cavanaugh, Gravier, et al., 2021; Kendall et al., 2019).

Recent work on learning and retrieval practice in aphasia provides a potential complementary restorative mechanism of action in SFA. Repeated retrieval practice, thought to strengthen both lexical-semantic and semantic-phonological connections, can improve naming performance on trained words (Middleton et al., 2016), and may contribute to improvements in naming treated words in SFA. This mechanism would also suggest that repeated feature generation (another form of retrieval practice) should improve an individual's ability to retrieve semantic features (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.

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 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.

While this mechanism of SFA has been proposed numerous times, it is worth noting that many SFA protocols do not address circumlocution or self-cueing strategies explicitly with participants (e.g., by providing metacognitive training around strategic communication) and, similar to the hypothesized restorative mechanism, assume such effects may occur passively. Examples of studies which have included direct attention to strategic communication include Wambaugh et al. (2013), who found improvements to trained words after incorporating an explicit self-cueing mediating strategy training phase during SFA. Tilton-Bolowsky et al. (2022) found that four out of four participants increased their awareness and use of strategic communication after SFA with a metacognitive component focused on self-cueing and circumlocution. In both studies however, generalization of treatment effects to untrained items was equivocal.

Effects of Semantic Feature Analysis Treatment on Discourse Production

The candidate restorative and compensatory mechanisms have clear implications for improvements to word finding deficits in the context of confrontation picture naming. But how

might these mechanisms engender improvements during discourse production? 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 enough trained and untrained (related) words across multiple semantic categories that these improvements could be relevant for discourse production even when the semantic content (i.e., words elicited by) of discourse stimuli 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. Alternatively, the compensatory account of SFA suggests that discourse on any topic could improve if individuals learn and then generalize communication strategies to use when they experience anomia.

The primary measurement paradigm for examining discourse outcomes in SFA is “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. In general, the evidence for generalization of SFA’s treatment effects to discourse production under this paradigm is inconsistent. Most recently, Silkes and colleagues (2021) did not find evidence that SFA improved discourse informativeness (a positive but non-significant effect at treatment exit) on a general-topic story retell task using stimuli from the Discourse Comprehension Test (Brookshire & Nicholas 1993). Across 30 individuals who received SFA in the study, there was substantial variability in outcomes, with only a subset of participants demonstrating positive change after treatment, though the authors did not identify any individual factors associated with treatment response, such as aphasia severity. These findings are consistent with single subject and small-N studies, which have reported highly variable changes in discourse performance after SFA (e.g., Boyle, 2011; Oh et al., 2016; Rider et al., 2008).

Preliminary Work

Unpublished, preliminary analyses of a recently completed SFA clinical trial (n = 44; Evans, Cavanaugh, Gravier, et al., 2021; Gravier et al., 2018) motivated the current study. They showed that individuals with aphasia demonstrated modest improvements in discourse informativeness (percentage of accurate, relevant words to total words produced) on general-topic, monologue-based discourse tasks (Nicholas & Brookshire, 1993) after intensive SFA (~60 hours of treatment). This SFA clinical trial may have been more likely to engender improvements in discourse production because the treatment paradigm also included a sentence generation component, requiring participants to generate a sentence using the target word and two semantically rich elements in every treatment trial. The improvements corresponded to an estimated average gain of 5 percentage points in correct information units (CIUs; a measure of discourse informativeness), 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 across participants.

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. 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 mechanism responsible for such gains, which is important both for understanding theories of language processing (i.e., Nickels, 2002) and facilitating generalization from often-used anomia treatments to valued connected speech contexts.

Multilevel Discourse Theory and Treatment for Aphasia

Multilevel discourse theory provides a 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 showed that a reduction in paraphasias was associated with resolution of grammatical deficits and improvements in discourse informativeness. Marini and colleagues interpreted this as evidence 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 of discourse production as a bidirectional, interactive process, from activating conceptual representations to formulating language, supported by necessary cognitive functions (Sherratt, 2007). Multi-level discourse models are supported by established relationships between discourse micro- and macrostructure in aphasia (e.g., Gordon, 2020; Harris Wright & Capilouto, 2012). Additional studies have also applied multilevel discourse approaches for characterizing discourse outcomes in aphasia treatment (Dipper et al., 2022; Whitworth et al., 2015).

Aim 1: Evaluating Monologue Discourse Outcomes in SFA

This study sought to evaluate four components of discourse production which are interrelated, can be conceptualized within multilevel discourse theory (Sherratt, 2007), and tie to the hypothesized mechanisms in SFA. The present claim is that multilevel analysis of discourse outcomes in a large SFA trial provides a unique opportunity and tractable basis for characterizing the linguistic context surrounding gains in discourse informativeness in preliminary data and examining the role of potential mechanisms supporting discourse generalization in SFA. The four components (measures) are lexical-semantic processing (the percentage of semantic paraphasias), lexical diversity (moving average type-token ratio), grammatical complexity (mean utterance length), and discourse informativeness (%CIUs). These measures were chosen to capture potential cascading treatment effects stemming from improvements in anomia, while balancing known psychometric properties and feasibility (see methods). 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.

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 more informative responses to discourse stimuli. Under this premise, if the preliminary findings of improved discourse informativeness are driven by the restorative mechanism, SFA should produce correlated improvements across lexical-semantic processing, lexical diversity, grammatical complexity, and informativeness. 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, and (c) the presence of an additional sentence-level component designed to promote generalization beyond the word level.

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). This pattern is better aligned with the deployment of compensatory strategies, but minimal resolution of the underlying impairment.

On the other hand, unrelated improvements across measures would not support the underlying restorative motivation for anomia treatments: that resolution of word-finding deficits will engender more effective discourse-level communication. Instead, this finding might indicate that the mechanism supporting discourse generalization in SFA largely depends on individual characteristics, which is the focus of Aim 2.

Aim 2: Examining Cognitive-linguistic Moderators of Monologue Discourse Outcomes in SFA

Key cognitive processes may be necessary for generalization to discourse-level communication in aphasia. Multilevel processing models suggest that semantic memory supports the process of

discourse production from frame/schema generation to linguistic formulation, whereas judgment processes such as attention and executive function support macrostructural-type processes such as selection of information, chunking of propositions, and linguistic formulation (Dipper et al., 2021; Pritchard et al., 2018; Sherratt, 2007). Although theoretical discourse frameworks have long hypothesized such relationships, there is limited empirical evidence for how these factors influence discourse production in aphasia. Still, these cognitive skills are associated with both direct training and stimulus generalization effects in anomia treatments (Lambon Ralph et al., 2010; Simic et al., 2020). Accordingly, deficits in semantic memory, attention, and executive function may moderate discourse outcomes in language-based interventions like SFA. For example, 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. Therefore, the purpose of Aim 2 is to explore the relationship between cognitive factors and discourse outcomes after SFA.

Methods

Description of study data sources

This study leveraged de-identified archival behavioral testing data from the aphasia research program at VA Pittsburgh (VAPHS) using the VAPHS Research Registry (Institutional Review Board #1617477). It included data from 60 individuals with chronic aphasia due to unilateral left-hemisphere stroke who completed an intensive SFA treatment across two highly similar, sequential clinical trials: 44 individuals with aphasia from *Dosage and predictors of naming treatment response in aphasia* (SFA-1; VA RR&D 5I01RX000832-05) 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 R01DC017475-01A1). 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).

Eligibility Criteria and Cohort characteristics

For both studies, participants were recruited from regional and national sources both within and outside the VA system. Inclusion criteria were (a) at least 6 months post-onset of left-hemisphere stroke, (b) no progressive neurological disorder or severe motor speech disorder, (c) scores 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), and (d) no concurrent speech-language treatment outside of study activities for the

duration of the study. Participants were required to be fluent English speakers prior to their stroke. Participants were not required to be veterans. Demographic characteristics of all participants are reported in Table 1.

	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]

Table 1. Demographic information for participants in SFA-1, SFA-2, and overall. SD: standard deviation, Min: Minimum, Max: Maximum, CAT: Comprehensive Aphasia Test

Assessment

A speech-language pathologist (SLP) administered and scored an extensive assessment battery at study enrollment, treatment entry, exit, and 1-month follow-up. Monologue-based discourse production was elicited using the Nicholas and Brookshire protocol for all participants at all time points (Brookshire & Nicholas, 1994; Nicholas & Brookshire, 1993). This protocol comprises two matched sets of 5 structured discourse stimuli across multiple genres: 2 picture descriptions, 1 narrative, 1 personal story, and 1 procedural discourse sample. Set order was pseudorandomized across timepoints to minimize ordering effects.

Only individuals with aphasia in SFA-1 ($n = 44$) completed the cognitive testing battery at treatment entry. The battery included (1) Pyramids and Palm Trees (PPT; Howard & Patterson, 1992) for semantic memory, (2) Test of Everyday Attention, elevator counting with and without distraction (TEA; Robertson et al., 1994), and (3) Wisconsin Card Sorting Task (WCST; Grant & Berg, 1948) for executive function (cognitive flexibility and problem solving). These assessments were originally chosen for SFA-1 because of known associations with aphasia treatment outcomes (Lambon Ralph et al., 2010).

Semantic Feature Analysis Treatment

Participants received SFA four to five days per week over four weeks for a total of roughly 60 hours. 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. For both studies, treatment was administered to each list sequentially, progressing between lists based on criterion (SFA-1) or after four days of treatment (SFA-2). For SFA-2, the last 3 days of treatment were sampled randomly from the 3 lists until all items were treated again. For both studies, treated and untreated words were selected for each participant based on their baseline naming performance. For SFA-1 words were selected based on performance on an extensive naming battery (see Gravier et al., 2018) and for SFA-2, words were selected to have a 30% probability of correct response as estimated by performance on the Philadelphia Naming Test (Fergadiotis, Swiderski & Hula, 2019). Treatment fidelity was monitored in both SFA-1 and SFA-2 (see Gravier et al., 2018). The SFA paradigm for both studies included these four components:

1. Naming: participants were presented with a picture and asked to name it, without feedback.
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 semantically rich elements (verb, subject, or object), typically 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.

Discourse Transcription and Coding

Discourse samples from each time point were orthographically transcribed and manually scored for correct information units and words by a VAPHS research SLP blinded to time point, which formed the basis for the preliminary analysis. Ten percent of samples were examined by a separate research SLP for transcription reliability using percent agreement on a word-by-word basis, which exceeded 90%. When frank transcription errors were found during the coding process (described below) and a correction was agreed upon by two independent raters, transcripts were amended. For these discourse samples, the following measures were selected based on team expertise, psychometric data (Boyle, 2020; Marini et al., 2011; Pritchard et al., 2018), established relationships between measures (Andreetta et al., 2012; Fergadiotis et al., 2013; Harris Wright & Capilouto, 2012; Marini et al., 2011), and theoretical ties to SFA.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). We hypothesized that a restorative SFA mechanism should reduce the frequency of semantic and verbal errors given improvements to the underlying semantic system while a compensatory SFA mechanism would have little impact on error frequency. 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) and real-word errors without a clear target were excluded from this measure. Semantic relatedness was based on potential relationships with the target word given the relationships identified in the SFA protocol (i.e., a contextual, functional, description, or categorical relationship to the target). 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 semantic. This definition was aligned with the intended construct: the incidence of semantic errors, given the opportunities to make semantic errors.

Lexical Diversity

Lexical Diversity is “the range of vocabulary deployed in a text” reflecting the speaker’s

¹ For the sake of full transparency, we note that in the grant proposal which funded this work, outcome measures included a) global coherence, which was dropped from the analysis plan, and b) predicate argument structure, which was replaced by mean utterance length. These changes were made due to a combination of psychometric limitations reported by colleagues (i.e., difficulty obtaining adequate rater agreement for the middle scores on the global coherence scales) and overall study scope/feasibility.

“capacity to access and retrieve target words from the lexicon” (Fergadiotis & Wright, 2011, p.145). We hypothesized that a restorative mechanism which produces improvements to the underlying semantic system or a compensatory mechanism which facilitated successful retrieval would increase lexical diversity by improving individuals access to a wider range of words. Lexical diversity was measured using the moving average type-token ratio (MATTR; Covington, 2007), a length-invariant approach which calculates a type-token ratio (the ratio of unique lexical items divided by the total number of words) for sequential, overlapping segments of a sample. MATTR has been shown to be an unbiased estimate of lexical diversity in individuals with aphasia (Fergadiotis et al., 2013) and used previously to evaluate 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).

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). 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). We hypothesized that MLU would increase as participants were engaged in producing grammatically and semantically rich utterances on each trial of SFA, resulting from knock-on improvements in lexical-semantic processing and lexical diversity.

Discourse Informativeness

Informativeness was measured by calculating the percentage of correct information units to total words (Brookshire & Nicholas, 1994) as utilized in the preliminary analysis and following the protocols outlined by Brookshire & Nicholas (1994). As demonstrated in Marini et al. (2011) and noted above, informativeness may be sensitive to changes in underlying linguistic structures (restorative) or solely improvements in strategic communication use (compensatory).

Reliability

All samples were independently coded by two certified SLPs blinded to time point. A discourse scoring codebook was established prior to coding and iteratively revised throughout discourse coding to account for edge-cases not addressed by initial coding rules. The codebook, including coding rules, rationales, and examples is reported in supplemental materials S3. 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 SLPs met to resolve all disagreements. A third expert

rater blinded to time point (4th author) resolved instances when the first two raters could not come to an agreement. Because all samples were independently rated by two raters and reviewed together, we did not record inter-rater reliability or intra-rater reliability. 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.

Statistical Analysis

Changes in discourse outcome measures and moderators of such change were analyzed using Bayesian generalized mixed-effects models. 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). The statistical procedures are described in detail in Supplemental S1, including prior distributions, prior predictive checks, model fitting, and convergence. Although data sharing is not permitted by VA Research Registry IRB restrictions, R code for statistical models is reported in Supplemental S4.

Aim 1 Analysis

Separate models were implemented for each discourse outcome measure with performance on the outcome measure as the dependent variable. A 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) random effect structures established a-priori. Random effects included stimuli as a random intercept, to account for the pseudorandomized order of stimuli presentation across participants, and a group-level intercept for participant and a by-participant slope for time point. This random effect structure permits participants not only to begin at their own level (intercept) but also change at different rates. Separate probability distributions were used for each outcome measure: beta-binomial mixture (lexical-semantic processing, informativeness), zero-inflated beta (lexical diversity), and shifted lognormal (grammatical complexity).

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 ultimately not possible due to the need to use different probability distributions for each outcome measure to best capture the data generation process. As a feasible alternative, 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 (i.e., measurement error at each time point), and interpretation of these correlations should therefore be considered exploratory and treated with caution.

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 fixed 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.

Results

Aim 1: Evaluating Monologue Discourse Outcomes in SFA

Fixed-effect parameters are reported in Table 2, with full model results for each outcome measure reported in the supplemental materials S2. For lexical semantic processing, 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 rare across all treatment time points, with no evidence for meaningful reduction.

For lexical diversity, 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 due to 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.

For grammatical complexity, 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.

For discourse informativeness, the intercept estimate was -0.21 (95% CI: [-0.45, 0.04]) indicating that %CIUs was approximately 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.

Correlations between individual change scores are reported in supplemental materials S2. None of the correlations were credibly different from zero with no discernable patterns, indicating that there was no set of individuals who were more likely to improve across multiple measures. These findings should be interpreted with caution given that these correlation estimates are influenced by multiple sources of measurement error omitted from the model.

Outcome	Parameter	Median	95% CI	pd
Lexical-Semantic Processing	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%
Lexical Diversity	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%
Grammatical Complexity	Intercept	1.7	[1.56, 1.83]	100%
	Time point [Enrollment - Entry]	-0.02	[-0.06, 0.03]	76.80%
	Time point [Exit - Entry]	0.00	[-0.04, 0.04]	52.30%
	Time point [Follow-up - Entry]	0.00	[-0.03, 0.04]	58.81%

Informativeness	Intercept	-0.21	[-0.45, 0.04]	95.14%
	Time point [Enrollment - Entry]	0.00	[-0.10, 0.11]	54.36%
	Time point [Exit - Entry]	0.04	[-0.07, 0.14]	75.15%
	Time point [Follow-up - Entry]	0.1	[-0.01, 0.21]	96.49%

Table 2. Fixed effects paramters for Aim 1. CI: Credible Interval, pd: probability of direction.
Note: Models use differing probability distributions.

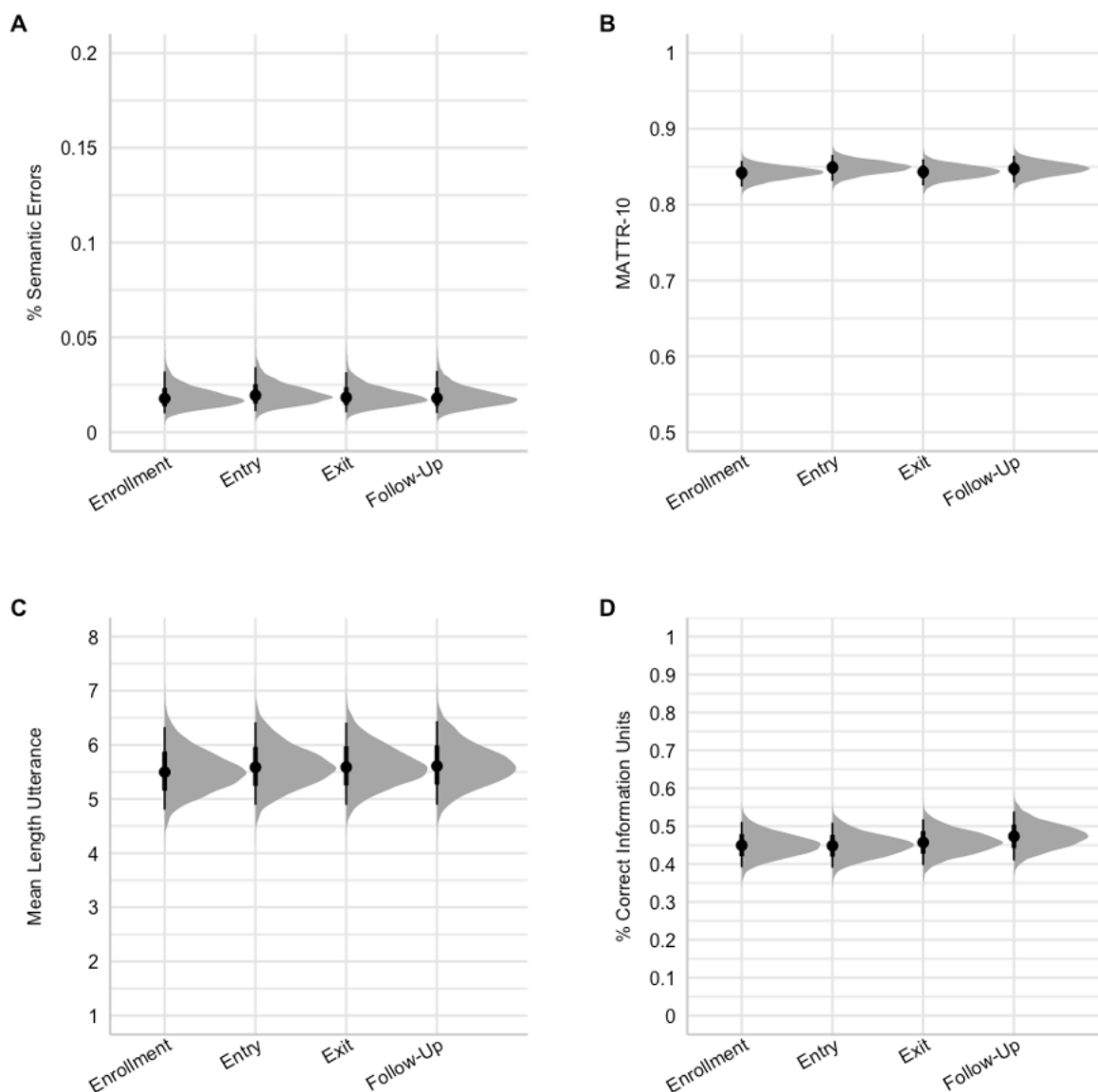


Figure 1. Conditional effects of timepoint for all four outcome measures. Panel A: Lexical-semantic processing. Panel B: Moving Average Type Token Ratio. Panel C: Mean Length Utterance. Panel D: Discourse Informativeness.

Aim 2: Examining Cognitive-linguistic Moderators of Monologue Discourse Outcomes in SFA

Lexical Semantic Processing

Full Aim 2 model results and supporting visualizations are reported in S2. The interaction between TEA performance and change in percentage of semantic errors 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 moderator of any potential treatment effect on the production of semantic errors.

The interaction between PPT performance and change in percentage of semantic errors 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.

The interaction between WCST performance and change in percentage of semantic errors between enrollment and entry was negative, but not reliably different from zero (-0.08, 95% CI: [-0.35, 0.17]). The interaction between WCST 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.

Lexical Diversity

None of the three non-language cognitive variables appeared to have a relationship with change in MATTR-10 across time points. 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 PPT 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]). 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.

Grammatical Complexity

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 PPT 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 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]).

Discourse Informativeness

Interactions between TEA performance and change were 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 PPT 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]).

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.

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 similar clinical trials of intensive SFA (total $n = 60$). 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 SFA treatment for aphasia through multilevel discourse analysis. The hypothesis was that patterns of improvement across discourse 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. Thus, Aim 1 provided little evidence for either mechanism, thereby warranting thoughtful consideration of how well SFA and other word-level aphasia treatments generalize to general topic discourse performance.

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 ($n = 44$). 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.

No Evidence for Change in Monologue Discourse Outcomes

The lack of strong evidence for improvements in *any* of the outcome measures for the full sample of 60 participants suggests that, when using the predominant discourse measurement paradigm (e.g., monologue discourse samples from the Nicholas & Brookshire protocol), SFA has minimal impact on general-topic discourse-level communication. This appears to be the case despite a high treatment intensity and additional sentence-level treatment component added to the standard SFA treatment protocol.

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 at baseline, 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 translate to large changes on the log-odds scale). 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. 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. In retrospect, focusing solely on semantic errors in discourse samples (i.e., excluding mixed and unrelated errors) provided cleaner data, but may have somewhat limited our ability to detect changes in lexical-semantic

processing.

Similarly, there was no evidence for a post-treatment 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 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 (i.e., affecting only trained and closely semantically related untrained words), 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. In other words, while SFA did not show these improvements in general-topic discourse tasks, it may produce stimulus generalization to discourse in more closely related elicitation contexts, although results to date in this area have been generally modest and inconsistent (Webster et al., 2015).

Without 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. The lack of such change is perhaps even more striking given the additional sentence-production step added to this protocol.

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. The most likely explanation for this change from preliminary data is modest regression to the mean (i.e., no effect) with a larger sample, though it is also possible that minor changes to the SFA paradigm in SFA-2 may have also played a role.

Taken together, the lack of changes across outcome measures does not support the role of either candidate mechanism (restorative or compensatory) on unrelated general monologue discourse outcomes in SFA. Instead, these results replicate recently published discourse outcomes, where Silkes and colleagues (2021) found little evidence that SFA improved discourse informativeness using a story retell procedure. A central takeaway across both studies is that there is consistently no evidence that SFA produces meaningful or reliable generalization to general monologue discourse production.

Potential Moderators of Discourse Outcomes in SFA

A common method of examining heterogeneity in clinical trials is to evaluate theoretically motivated moderators of treatment response, which was the focus of the exploratory Aim 2. In

general, there was scant evidence that better semantic memory, divided attention, and executive function scores are associated with greater treatment outcomes. Furthermore, the effects that were statistically reliable tended to be inconsistent in direction across time-points, 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. This pattern is also apparent for the interaction between semantic memory and change in lexical-semantic processing. 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.

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. It is also worth noting that the WCST is focused on set-shifting as one component of executive function, and it is possible that other aspects of executive function may be important for discourse generalization or maintenance of treatment effects in aphasia.

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. 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. The absence of any strong simple effects warrants further examination of the relationships between non-language cognitive factors and discourse production and calls into question the extent to which such relationships are large enough to be of clinical interest.

While cognitive-linguistic moderators of treatment outcomes are examined less often, aphasia severity is often proposed as a moderator of treatment outcomes. For example, individuals with mild to moderate aphasia might be more like to improve naming performance relative to individuals with severe aphasia after SFA (e.g., Quique et al., 2019). Therefore, as a secondary analysis, we also evaluated aphasia severity as a moderator of outcomes across each measure (using the same methods as above), but did not find any reliable evidence that severity moderated changes from treatment entry to exit or follow-up (see supplemental table S.5) The lack of any consistent moderators of treatment effects in this study add to the web of evidence that, in this study, SFA had little systematic impact on general monologic discourse outcomes for any single participant or specific group of participants.

Discourse Outcome Measurement Challenges in Aphasia Treatment

Given these largely null effects, it is crucial to acknowledge the challenges posed by seeking to examine treatment-related changes in monologue discourse production in aphasia. General challenges in reliably measuring discourse-level communication in aphasia 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; Dipper et al., 2021; Stark et al., 2020). For example, our a-priori choices discourse outcome measures were well-intentioned and selected based on both theoretical and practical grounds, but have important limitations, as noted throughout this discussion. Additional study-specific considerations also exist.

As noted briefly above, 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 employ 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. 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 monologue discourse production. An alternative for future treatment studies is to select discourse elicitation tasks more directly aligned with expected treatment-related changes, such as contexts which are known to elicit targeted vocabulary (e.g., Conroy et al., 2018; Rider et al., 2008).

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 (though these could manifest in longer MLU), 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; Tilton-Bolowsky et al., 2022), but larger-scale studies are needed to establish any effects of this mechanism.

Another challenge specific to evaluating multilevel discourse outcomes in aphasia is the large degree of heterogeneity in discourse profiles at baseline. Not all participants will need to improve on every measure, nor will all participants be capable of making substantial improvements on each measure (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 some participants. 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. Similarly, 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 for both SFA-1 and SFA-2 were minimally restrictive, this study included several participants with limited verbal output in connected speech settings, limiting measurement precision.

Discourse Generalization in SFA: Moving beyond “Train and Hope”

There is also a key theoretical limitation to the premise that improvements in word-finding after SFA should generalize to monologue discourse. Each of the hypothesized mechanisms of action for SFA discussed in the introduction 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” (Stokes & Baer, 1977) and put forward as the major contributor to poor generalization outcomes in other fields.

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. 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 would likely need to make enormous improvements to the underlying language system to passively generalize to an unrelated monologue-discourse. However, realizing this level of “system restoration” has been elusive in aphasia rehabilitation from both behavioral and neuroscience perspectives (Brady et al., 2016; Schevenels et al., 2020). Given our null effects, it also appears unlikely that individuals with aphasia 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. Recent work by Tilton-Bolowsky et al., (2022) provides preliminary evidence that intentionally incorporating treatment ingredients that focus on generalization may have the potential to improve generalization in anomia treatments like SFA.

Both the behavior modification literature (Stokes & Baer, 1977) and knowledge transfer literature (e.g., Gick & Holyoak, 1987) offer potential strategies and treatment ingredients for facilitating generalization. Future treatment studies could examine mediational or meta-cognitive strategies (e.g., Tilton-Bolowski et al., 2022, Evans, Cavanaugh, Quique, et al., 2021), increase stimulus or response variability, varying training environments and communication partners, (Doyle et al., 1989), and provide and reinforce opportunities for generalization to the intended context (e.g., Falconer & Antonucci, 2012). Alternatively, researchers interested in improving discourse outcomes might be best served by selecting discourse-focused treatments, rather than continuing to use anomia treatments.

Clinical Implications

Anomia treatments are among the most used approaches in clinical services provided to individuals with aphasia, and SFA is one of the more common 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, these findings should elicit careful consideration about how word-level treatments like SFA engender improvements to monologue discourse or other connected speech contexts. Clinical researchers 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. In other words, the results of the present study suggest that SFA does not have any clinically-meaningful effects on general-topic discourse monologue in aphasia. Given this finding, we recommend that clinicians do not employ traditional SFA to improve discourse monologue. Instead, it is likely more effective to use discourse treatment approaches (e.g., Dipper et al., 2022, or Leaman & Edwards 2024) to improve discourse production and everyday communication.

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 for clinicians on strategies that may be useful in facilitating generalization (Coppens &

Patterson, 2017; Mayer et al., 2024; Thompson, 1989; Webster 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; 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. 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. 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).

Conclusion

Multiple stakeholder groups, most notably including individuals with aphasia (Wallace, Worrall, Rose, & Le Dorze, 2017; Wallace, Worrall, Rose, Le Dorze, et al., 2017), prioritize 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 was sufficient to inform treatment theory or clinical decision-making. These findings indicate 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.

Acknowledgements

This work was supported by many researchers, clinicians, and students at the University of Pittsburgh and VA Pittsburgh Healthcare Center and would have been impossible without the contributions of more than 60 participants with aphasia and their families. It was supported by the following awards: 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 and

National Institute on Deafness and Other Communication Disorders NRSA F31DC019853-01 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. Finally, this work was supported by the Audrey Holland Endowed Student Research Award at the University of Pittsburgh. Dr. Holland had an immeasurable impact on the lives of countless individuals with aphasia, clinicians, and researchers, and she will be profoundly missed.

Data Availability Statement

The datasets analyzed during the current study are not publicly available due to data restrictions under the relevant IRB protocols. Analytical code is available in the supplemental materials.

Table & Figure Captions

Table 1. Demographic information for participants in SFA-1, SFA-2, and overall. SD: standard deviation, Min: Minimum, Max: Maximum, CAT: Comprehensive Aphasia Test

Table 2. Primary fixed effects for Aim 1 Models. CI: Credible Interval, pd: probability of direction. Note: Models use differing probability distributions.

Figure 1. Conditional effects of timepoint for all four outcome measures. Panel A: Lexical-semantic processing. Panel B: Moving Average Type Token Ratio. Panel C: Mean Length Utterance. Panel D: Discourse Informativeness.

References

- 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>
- 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., & 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. (1993). Discourse Comprehension Test. Communication Skill Builders.
- 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.
- 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).
- 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.
- 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.
- 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.
- Dutta, M. (2020). *Evaluating the relationship between executive functioning, spoken discourse, and life participation in aphasia [Unpublished Dissertation]*. Indiana University.
- 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., Swiderski, A., & Hula, W. D. (2019). Predicting confrontation naming item difficulty. *Aphasiology*, 33(6), 689–709.
- 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) [Computer software].
- 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>
- 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.
- 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 Jacques, 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
- 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
- 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>
- 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.
- Leaman, M. C., & Edmonds, L. A. (2024). Pilot Results for ECoLoGiC-Tx: A New Conversation-Level Intervention Improving Language in People With Moderate to Severe Aphasia. *American Journal of Speech-Language Pathology*, 33(1), 153-172.
- 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.
- 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>
- Mayer, J. F., Madden, E. B., Mozeiko, J., Murray, L. L., Patterson, J. P., Purdy, M., ... & Wallace, S. E. (2024). Generalization in Aphasia Treatment: A Tutorial for Speech-Language Pathologists. *American journal of speech-language pathology*, 33(1), 57-73.
- McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC 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.
- 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.
- 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. (2023). 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\(2008/016\)](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.
- 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-Rohrer, 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.
- 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., & Nessler, C. (2013). Semantic feature analysis: Incorporating typicality treatment and mediating strategy training to promote generalization.

- 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., Leitaio, 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.

Supplemental Materials

- S1: Complete description of statistical analysis methods
- S2: Full model results including tables and figures
- S3: Discourse scoring codebook
- S4: Model syntax and R session information
- S5: Post-hoc models: Interactions between aphasia severity (CAT mean modality T-score) and timepoint across outcome measures

Supplemental Materials S1. Statistical Analysis

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.

Aim 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 $\text{lkj} = 1$ was utilized for the correlation parameter. With a value of 1, extreme correlation values become less likely.

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 (z_i parameter). Provided that at least 10 tokens were produced, it then estimates lexical diversity as measured by the MATTR-10. Both the z_i 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 z_i 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 $\text{lkj} = 1$ was utilized for the correlation parameter, such that extreme correlation values are thought to be less likely.

Grammatical Complexity (Mean Length 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 $\text{lkj} = 1$ was utilized for the correlation parameter. With a value of 1, extreme correlation values become less likely.

Discourse Informativeness (% Correct Information Units to Content Words)

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 $\text{lkj} = 1$ was utilized for the correlation parameter.

Aim 2

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.

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 below.

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.

Model Diagnostic Plots

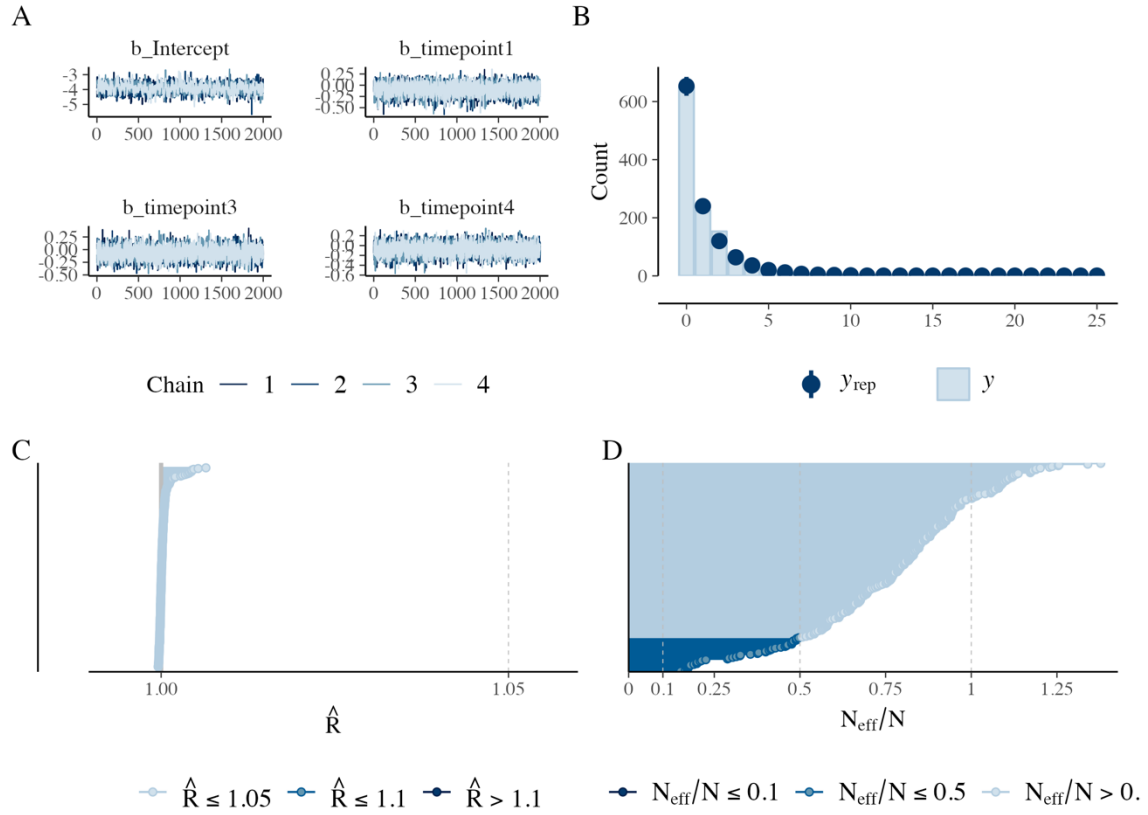


Figure S1.1 Diagnostic plots for lexical-semantic processing model Aim 1

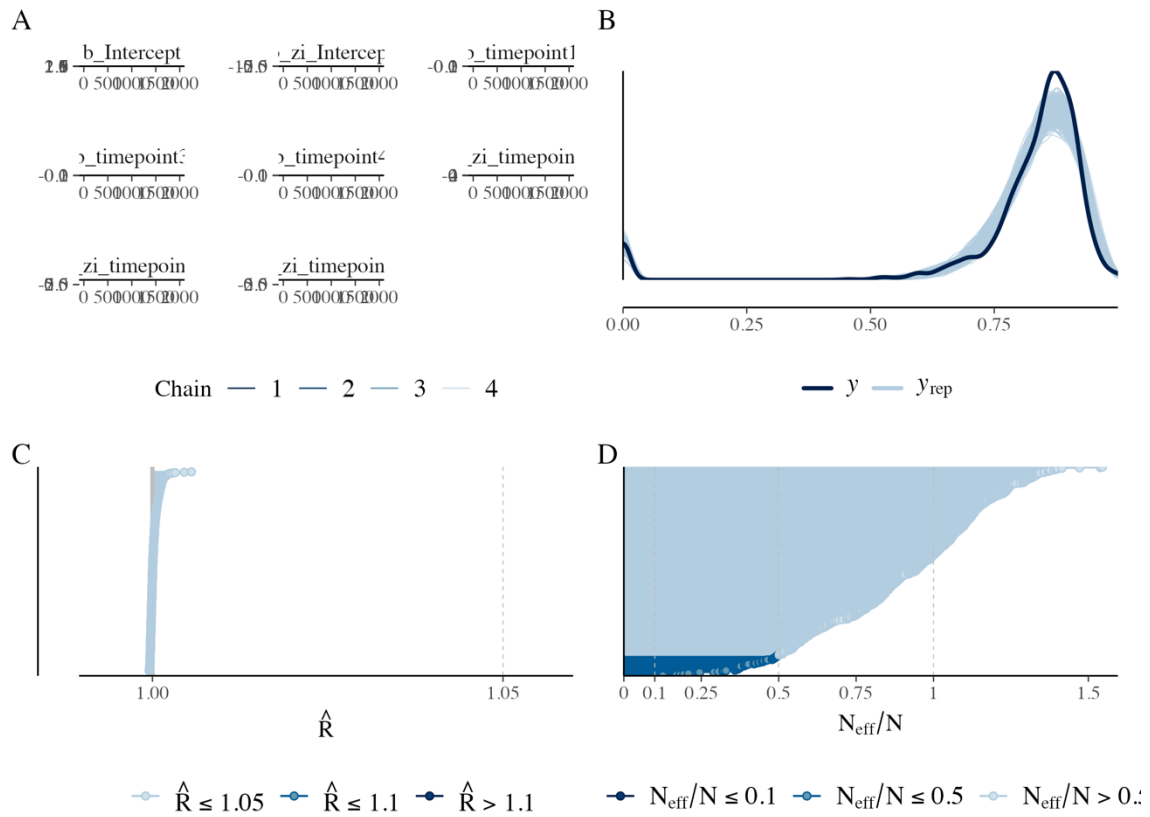


Figure S1.2 Diagnostic plots for lexical diversity model Aim 1

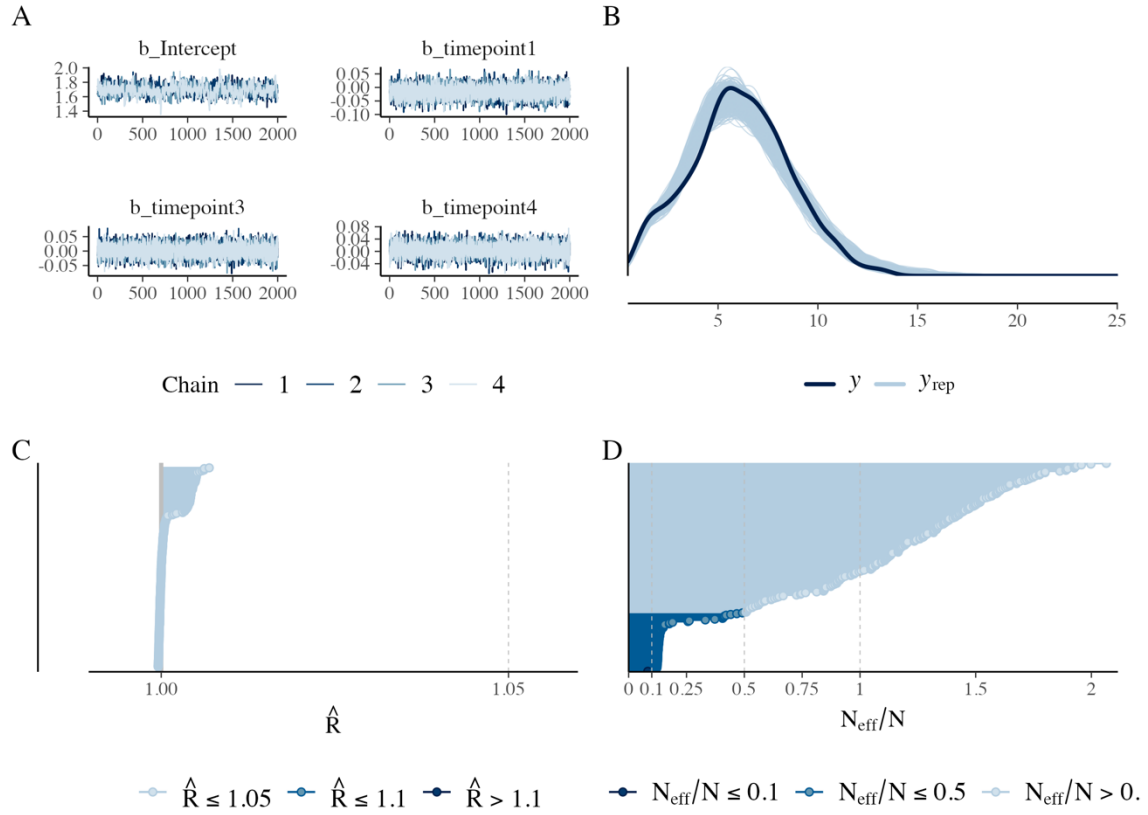


Figure S1.3 Diagnostic plots for mean length of utterance model Aim 1

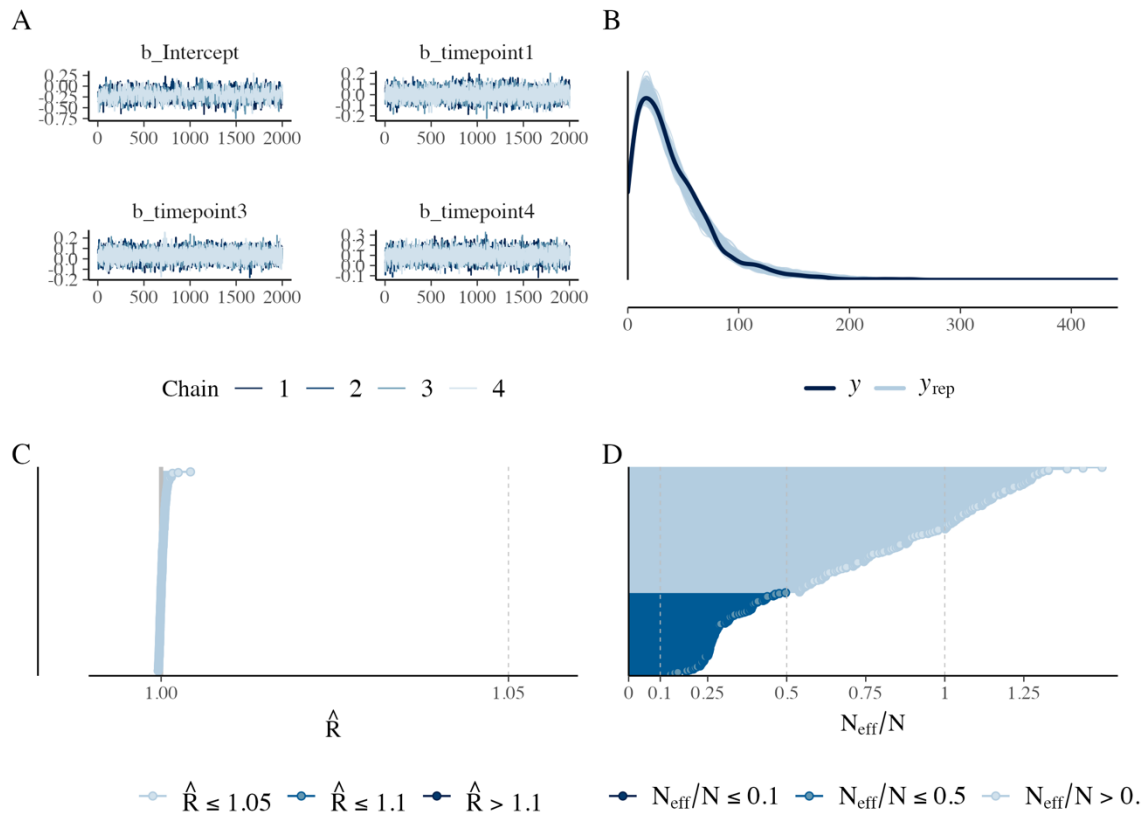


Figure S1.4 Diagnostic plots for discourse informativeness model Aim 1

S2. Supplemental: Full model results

Aim 1

Table S2.1 Model results for lexical-semantic processing

Effects	Parameter	Median	95% CI	pd
Fixed	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%
Random	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%
Fixed	phi	73.86	[54.55, 104.75]	100%

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

Table S2.2 Model results for lexical diversity

Effects	Parameter	Median	95% CI	pd
Fixed	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%
Random	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%
	phi	40.31	[36.76, 44.27]	100%
Fixed: 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%
Random: ZI	SD (Intercept: participant)	3.53	[2.26, 5.48]	100%

Effects	Parameter	Median	95% CI	pd
	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%
	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

Table S2.3 Model results for mean length of utterance

Effects	Parameter	Median	95% CI	pd
Fixed	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%
Random	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%
Fixed	sigma	0.23	[0.22, 0.24]	100%

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

Table S2.4 Model results for discourses informativeness

Effects	Parameter	Median	95% CI	pd
Fixed	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%
Random	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%
	phi	18.90	[16.73, 21.30]	100%

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

Correlations between Outcome measures

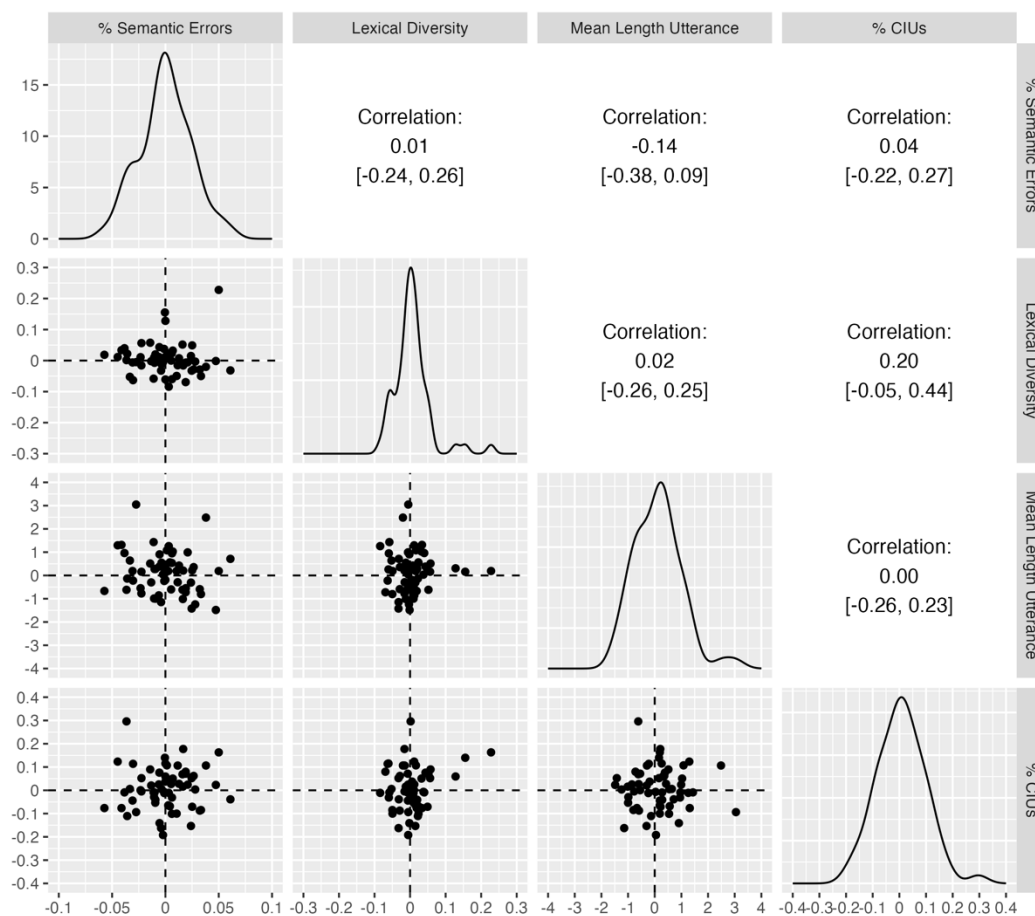


Figure S2.1 Correlations between change scores from treatment enrollment to treatment entry

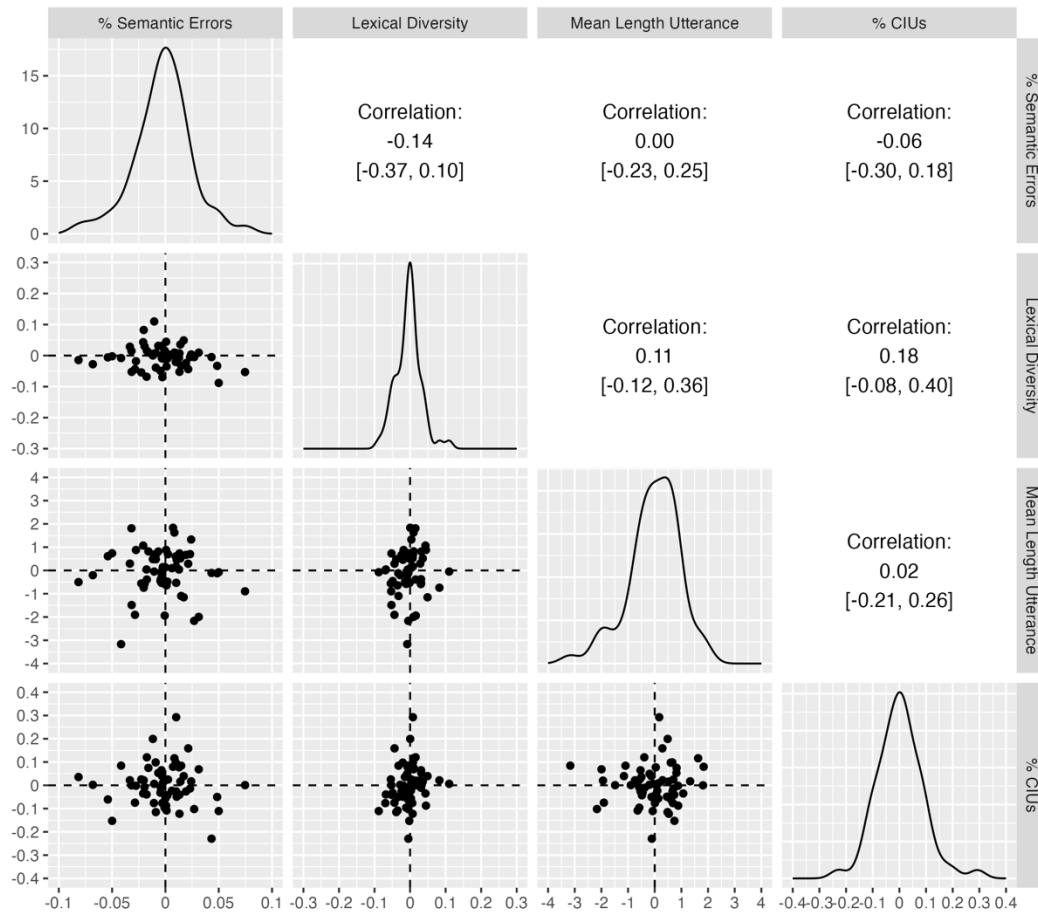


Figure S2.2. Correlations between change scores from treatment entry to treatment exit

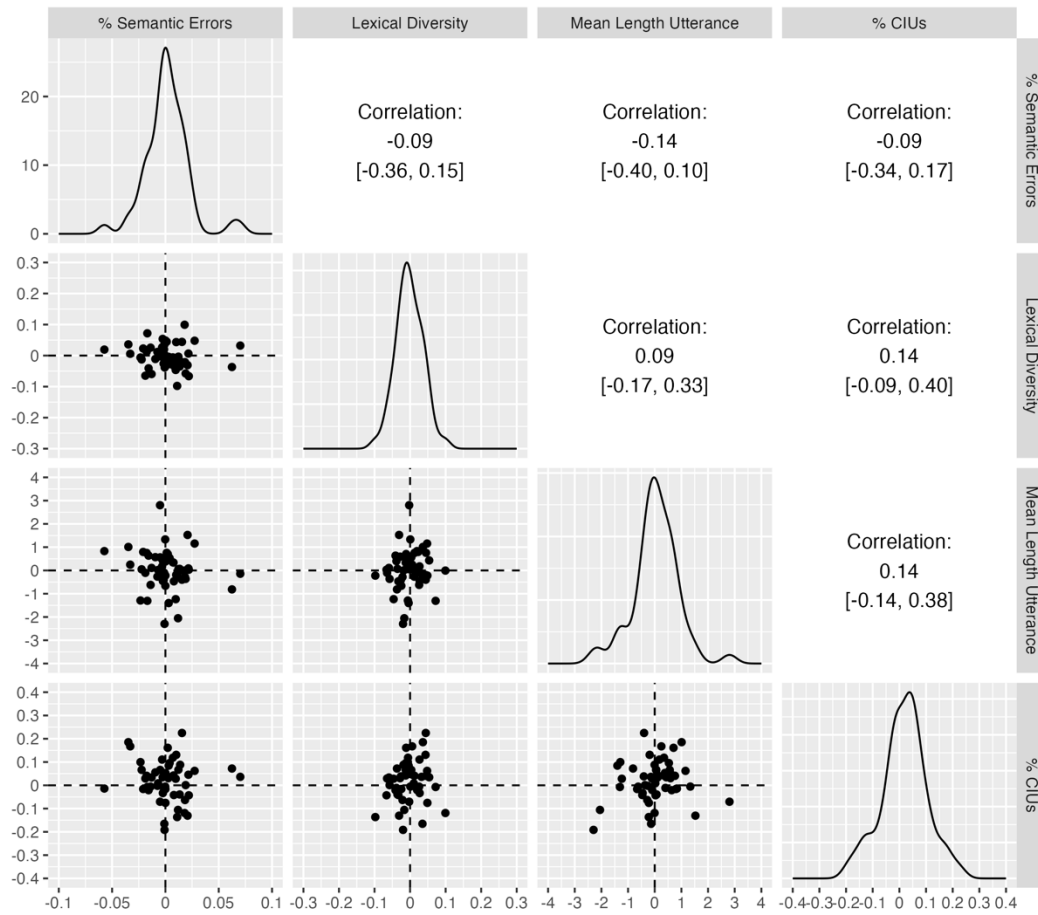


Figure S2.3 Correlations between change scores from treatment entry to 1-month Follow-up

Aim 2

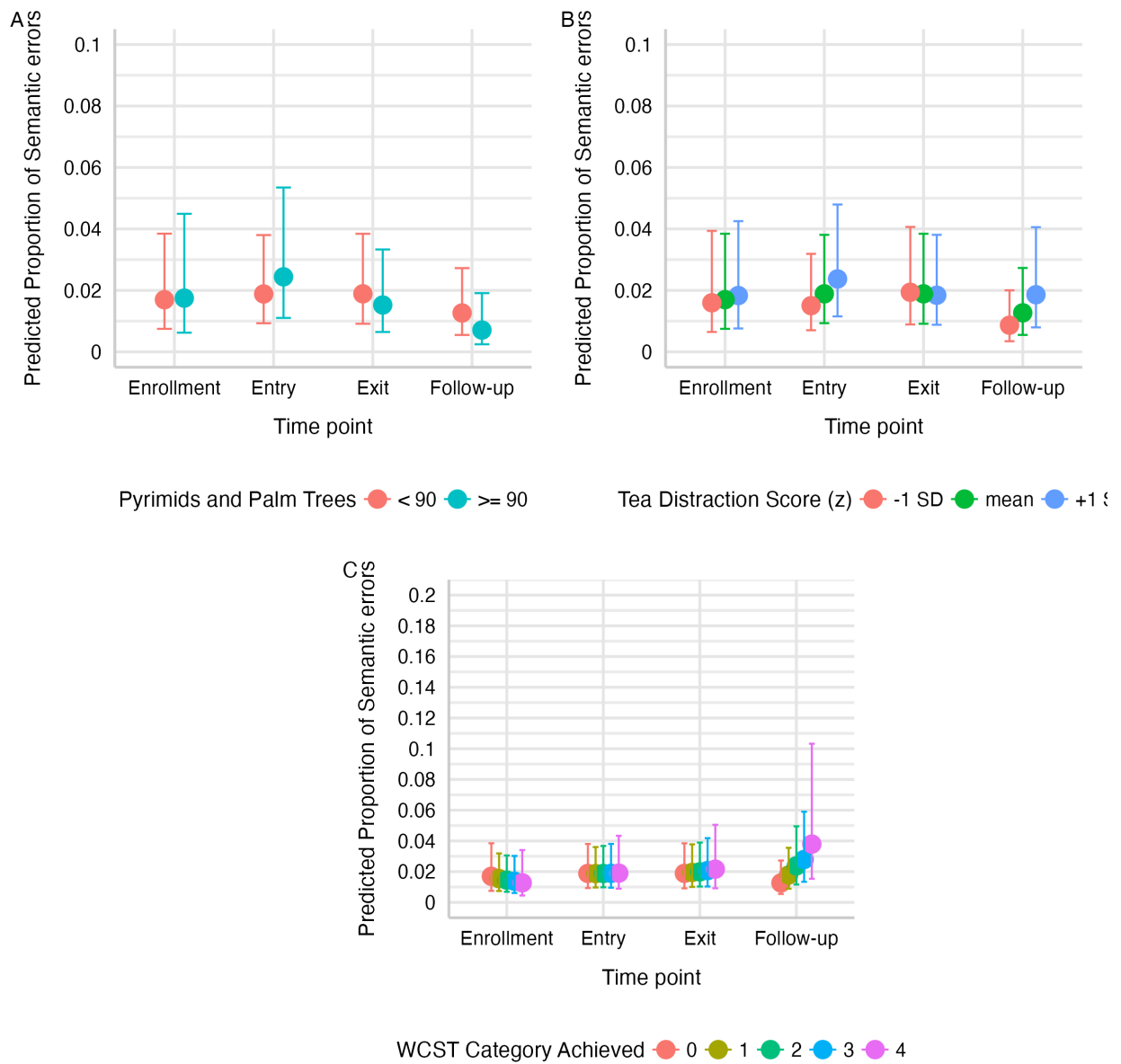


Figure S2.4. Relationships between lexical-semantic processing outcomes and non-language cognitive scores.

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

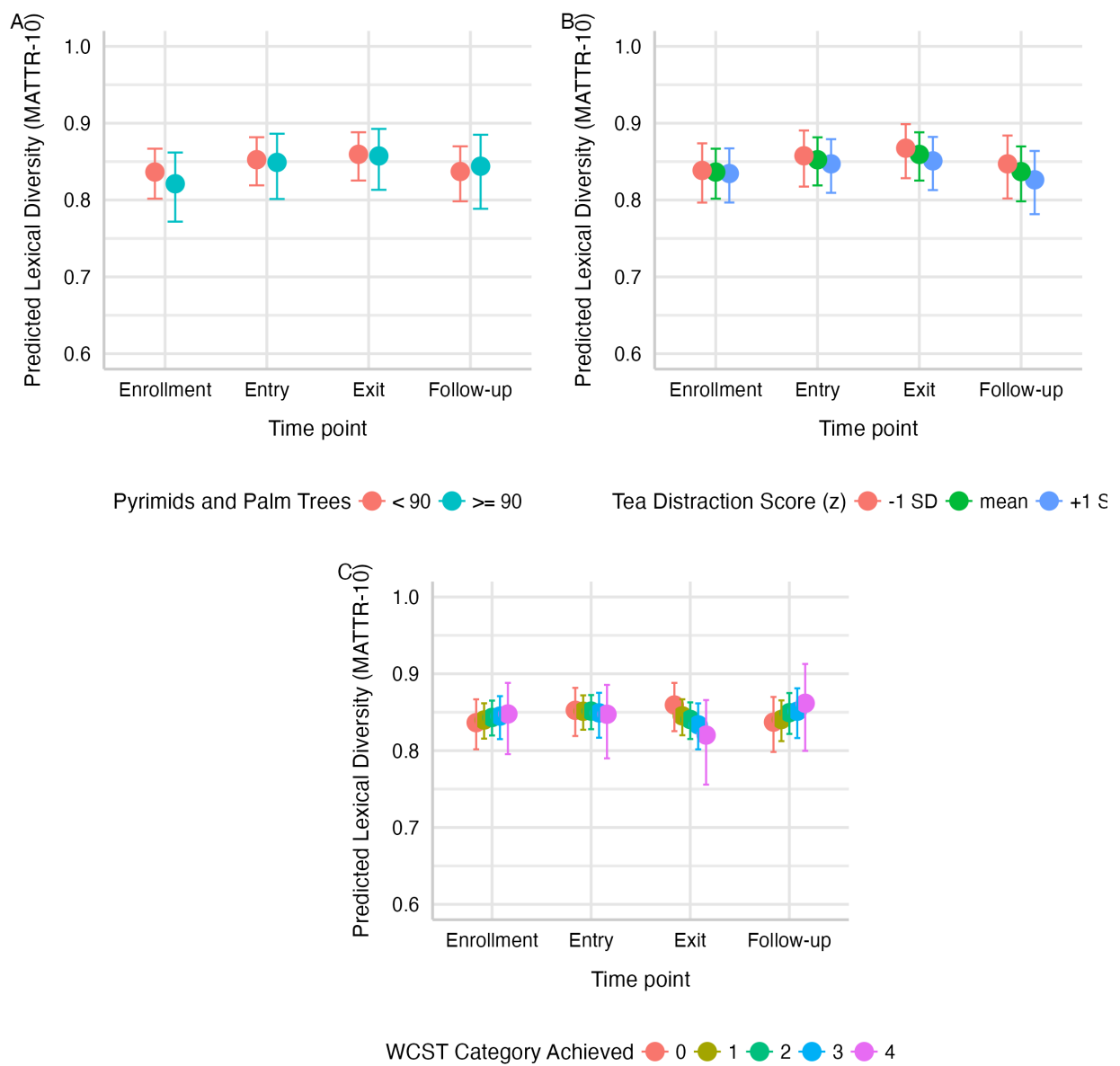


Figure S2.5 Relationships between lexical diversity outcomes and non-language cognitive scores

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

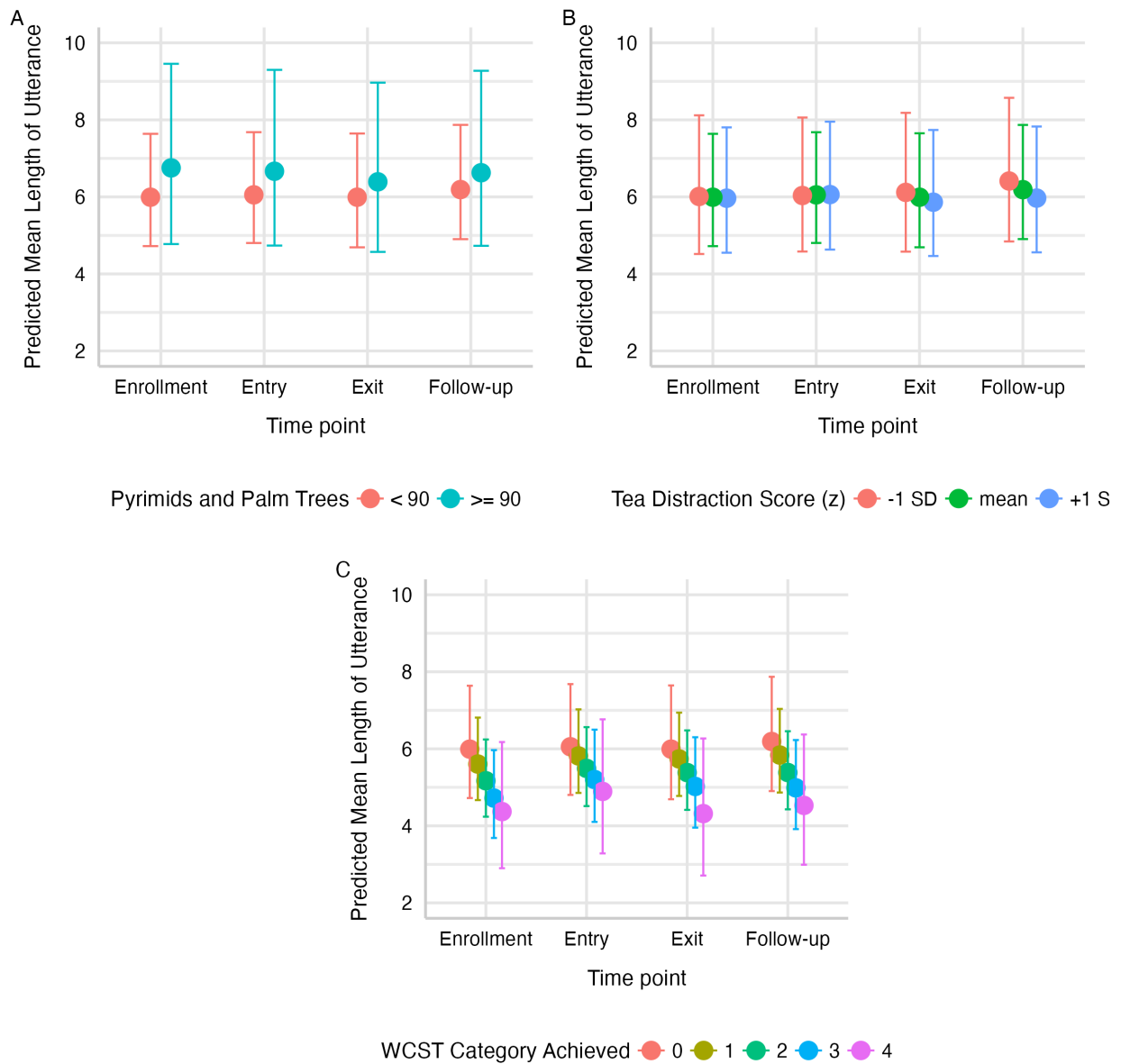


Figure S2.6 Relationships between mean length of utterance outcomes and non-language cognitive scores
Panel A: Semantic Memory; Panel B: Divided Attention; Panel C: Executive Function

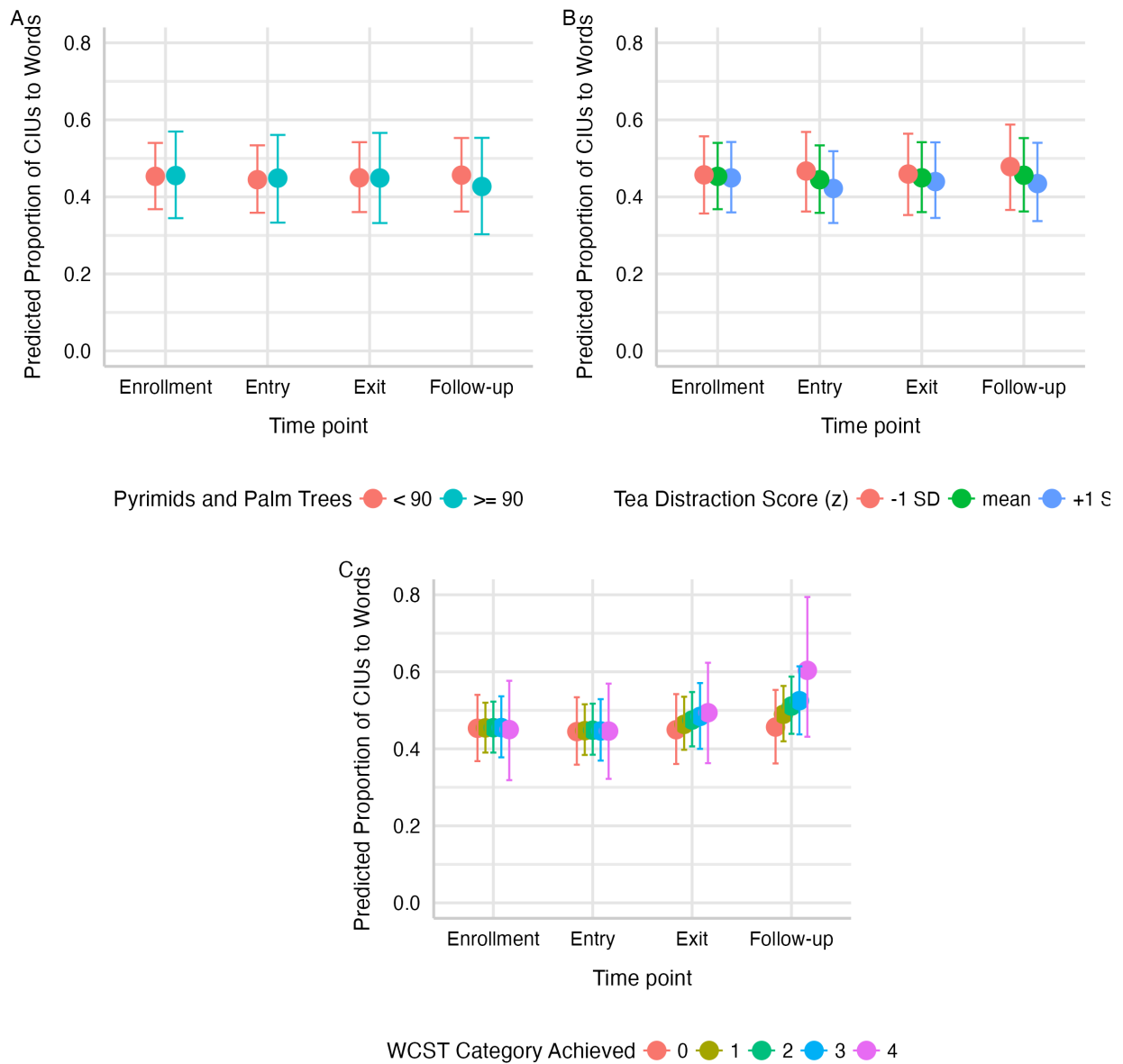


Figure S2.7 Relationships between discourse informativeness outcomes and non-language cognitive scores
Panel A: Semantic Memory; Panel B: Divided Attention; Panel C: Executive Function

Table S2.5 Full model results for lexical-semantic processing (Aim 2)

Effects	Parameter	Component	Median	95% CI	pd
Fixed	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%
Random	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%
	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%

Effects	Parameter	Component	Median	95% CI	pd
Fixed Monoton ic	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%
	phi	distributional	68.47	[48.16, 100.94]	100%
	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 S2.6 Full model results for lexical diversity (Aim 2)

Effects	Parameter	Component	Median	95% CI	pd
Fixed	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%
Random	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%

Effects	Parameter	Component	Median	95% CI	pd
Fixed Monoton ic	cor (Intercept ~ Time point [Follow-up - Entry]: participant)	conditional	0.07	[-0.67, 0.76]	56.60%
	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%
	phi	distributional	39.15	[34.99, 43.58]	100%
	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 S2.7 Full model results for mean length of utterance (Aim 2)

Effects	Parameter	Component	Median	95% CI	pd
Fixed	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%
Random	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%

Effects	Parameter	Component	Median	95% CI	pd
Sigma Monotonic	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%
	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	0.23	[0.22, 0.24]	100%
	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 S2.8 Full model results for discourse informativeness (Aim 2)

Effects	Parameter	Component	Median	95% CI	pd
Fixed	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%
Fixed (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%
Random	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%

Effects	Parameter	Component	Median	95% CI	pd
	SD (Time point [Follow-up - Entry]: participant)	conditional	0.11	[0.01, 0.26]	100%
	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%
Monoton ic	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%

Effects	Parameter	Component	Median	95% CI	pd
	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: correlation

S3. Supplemental 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.”

Segmenting examples for discourses 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 <i>repeated</i> 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
<i>Second example of above</i>	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

S4. 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)

```
# 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)

```
# 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"
  )
```

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
##
## -----
```

S5. Post-hoc models: Interactions between aphasia severity (CAT mean modality T-score) and timepoint across outcome measures

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