Published in final edited form as:

Aphasiology. 2020 ; 34(2): 214–234. doi:10.1080/02687038.2019.1643002.

The relationship between trained ratings and untrained listeners' judgments of global coherence in extended monologues

Yvonne Rogalski^a, Sarah Key-DeLyria^b, Sarah Mucci^c, Jonathan Wilson^d, Lori J. P. Altmann^e

^aIthaca College

^bPortland State University

^cNew York University Steinhardt

^dMidwestern University

^eUniversity of Florida

Abstract

Background: Global coherence rating scales have been used by a number of researchers to examine spoken discourse in populations with and without acquired neurogenic communication disorders. The 4-point global coherence (GC) scale in the current study has demonstrated reliability and convergent validity. However, we have not yet established how a global coherence rating corresponds to functional communication.

Aims: The current study explores the clinical meaningfulness of the 4-point GC rating scale. Survey questions and ratings were developed to examine discourse quality and functional coherence as perceived by untrained listeners. Ratings by untrained listeners were compared to trained discourse ratings using the established 4-point GC scale.

Methods: Twelve discourse samples, scored by a trained rater, were selected for the current study from a previously collected set of discourse transcripts. Transcripts were extended monologues in response to one of four possible open-ended questions that were re-recorded by the trained rater to remove any distracting features of the original recordings, such as articulatory errors. Twenty-four untrained listeners rated the discourse samples using a short questionnaire that included questions for each sample about: topic maintenance, inclusion of unnecessary information, and the listener level of interest and attention. Questions for untrained listeners were based on operational definitions of global coherence and discourse quality respectively. These untrained ratings were compared to trained ratings of global coherence. Outcome measures were compared using non-parametric tests and a Spearman Rank Order test was also used to examine relationships among variables.

Outcomes and Results: Untrained listeners' ratings for topic maintenance and inclusion of unnecessary information were significantly different between trained low, medium, and high global coherence ratings. Topic maintenance and inclusion of unnecessary information were also

both significantly correlated with global coherence. Untrained listeners' ratings of their level of interest and attention for a sample were significantly different between trained medium-high and low-high global coherence ratings. Interest and attention ratings were also significantly correlated with ratings of global coherence.

Conclusions: Untrained raters did differentiate between levels of global coherence using ratings of topic maintenance, inclusion of unnecessary information, and their level of attention and interest. Global coherence was also significantly correlated with the untrained ratings. These findings provide preliminary external validity for the global coherence scale and support its clinical utility.

Global coherence analysis, often considered a measure of topic maintenance, is a type of discourse analysis of spoken language production that can be an informative tool for researchers and clinicians in the field of speech-language pathology. It can help us to understand the underlying cognitive-linguistic impairments of patient populations, as well as to plan and tailor treatment goals. Communication through discourse is often the end goal of cognitive-linguistic rehabilitation. Consequently, there has been a paradigm shift in aphasiology, moving from a word or sentence level focus toward discourse as an outcome measure for the assessment of spoken language (Dietz & Boyle, 2018). Many of the existing discourse measures, however, have not been tested for validity or reliability (Boyle, 2014, 2015). The result is a myriad of methodological types of discourse analysis and techniques, often with the operational definition of each discourse measure varying from investigator to investigator, making cross-study comparisons more difficult (Dietz & Boyle, 2018; Linnik, Bastiaanse, & Höhle, 2016). Moreover, despite speech-language pathologists reporting that inclusion of discourse analysis in assessment is important (Bryant, Spencer, & Ferguson, 2017), a number of studies have presented evidence revealing that use of discourse analysis in clinical practice is limited (Bryant et al., 2017; Frith, Togher, Ferguson, Levick, & Docking, 2014; Maddy, Howell, & Capilouto, 2015; Verna, Davidson, & Rose, 2009; Westerveld & Claessen, 2014). In Bryant and colleagues survey (2017), 41.5% of speech-language pathologists reported difficulty in selecting appropriate analysis techniques as a barrier to implementing discourse analysis into practice, along with lack of time, training, expertise, and resources. It is imperative, therefore, to hone in on tested measures that are time-saving, accessible, easy to employ and require limited training—measures that have been used to analyze the discourse of a variety of populations, thus contributing to a resource base on which clinical comparisons can be made. Importantly, to ensure that discourse measures are clinically meaningful, they should demonstrate a degree of external validity. More clinicians might incorporate discourse analysis into practice if they knew that their clients' gains on discourse measures meant that average listeners could detect these changes.

The current study examines the external validity of the 4-point global coherence (GC) rating scale, a relatively easy to use scale that has demonstrated reliability and convergent validity with the Glosser and Deser (1990) 5-point GC scale (Wright, Capilouto, & Koutsoftas, 2013). As will be discussed below, both scales have been used by a variety of researchers to study the spoken discourse patterns of a number of different populations with and without acquired neurogenic communication disorders, lending credence to the clinical utility of

these scales, and the rationale for examining the external validity of the already tested 4-point GC scale.

Background of the 4-point and 5-point GC Rating Scales

Discourse can refer to the comprehension or production of spoken or signed utterances or written text and is defined differently depending on whether the theoretical approach to analysis is structural, functional, or strategic. The current paper examines discourse from the structuralist perspective, which defines discourse as language that is connected by successive spoken utterances or successive written sentences (e.g., Grimes, 1975; Harris, 2013), with the analysis based on *form*. This differs from the functionalist approach, which is language as it is used within context regardless of discourse length (e.g., Christie & Martin, 2009; Halliday, Matthiessen, & Halliday, 2004), and the “strategic” approach, which is based on the integration of the linguistic and semantic content of a text with the comprehender’s world knowledge (e.g., van Dijk & Kintsch, 1983).

The term “coherence,” like discourse, can have a number of different definitions depending on various theoretical origins. Global coherence is one type of discourse analysis that has been described and operationalized in a number of different ways (for a review, see Ellis, Henderson, Wright, & Rogalski, 2016). The focus of the current study is the construct of global coherence as theoretically derived from Agar and Hobbs (1982), whose ethnographic interviews outlined three types of coherence (global, local, and themal) to characterize organizational aspects of coherence in the transcripts of extended monologues. Global coherence is defined as a top-down approach where each utterance can be analyzed in terms of its relationship to the overall plan or topic of a discourse segment. This is in contrast to local coherence, a bottom-up approach where each utterance is related to the immediately preceding utterance. It is also different from themal coherence, which refers to the recurring “threads” of narrative strategies, underlying assumptions, functions, and frequency of coherent structures that relate one segment of discourse to other segments (Agar & Hobbs, 1982). The current paper defines global coherence as a macrostructure analysis thought to measure higher order macrolinguistic language processes that are cognitive-linguistic in nature, as opposed to a microstructure analysis which is considered a within-sentence measurement of microlinguistic processing such as the phonological, lexical and syntactic components of language.

Glosser and Deser (1990, 1992) were the first to operationalize global coherence as a 5-point rating scale that was theoretically derived from Agar and Hobbs (1982) and used to analyze discourse macrostructure compared to microstructure. The following is a broad outline of Glosser and Deser’s (1990, 1992) methods for their global coherence analysis, which laid the foundation for the 4-point GC scale used in the current study (Wright et al., 2013). First, discourse was elicited from patients during recorded interviews where they were asked to speak on the topic of their work experience or their families. Next, the discourse was transcribed and divided into segments called communication units (C-units) according to published procedures (e.g., Hunt, 1965; Loban, 1963) where each C-unit represented an independent clause plus any related clause. The first 20 C-units were scored for global coherence irrespective of any microstructural errors occurring at the lexical or syntactic

level. Then, each of these C-units was rated on a 5-point GC scale in terms of its relatedness to the topic (either work or family), where C-units receiving higher scores provided “substantive information directly related to the designated topic” (Glosser & Deser, 1992, p. 268). Finally, a mean global coherence score was obtained for each sample by adding together the scores of each of the C-units and dividing by the total number of C-units. Glosser and Deser (1990, 1992) then used the mean values of the discourse samples to make comparisons across groups of patients and non-brain-injured adults, introducing the usefulness of the 5-point GC scale as a measure that potentially reflects the underlying cognitive-linguistic aspects of discourse.

With permission from the authors, Van Leer and Turkstra (1999) adapted Glosser and Deser’s (1987) unpublished guidelines for the 5-point GC rating scale used in the 1990 and 1992 studies, and were the first to publish the 5-point GC rating scale guidelines. Subsequently, Wright, Capilouto, and Koutsoftas (2013) developed a 4-point GC scale from the Van Leer and Turkstra (1999) adaptation of the Glosser and Deser scale and were the first to provide empirical validation for the scale. The modifications introduced by Wright et al. (2013) included reducing the rating range from five points to four points, since prior studies had not found much use for the “2” and “4” ratings of the five-point scale (e.g., Rogalski, Altmann, Plummer-D’Amato, Behrman, & Marsiske, 2010; Rogalski & Edmonds, 2008; Van Leer & Turkstra, 1999), and providing a more extensive set of criteria for each rating point in the scale (see Wright et al., 2013 for rating scale). Wright and colleagues (2013) tested both scales in non-brain-injured adults and reported that the 4-point GC scale yielded high measurement reliability. Moreover, both the 4-point and 5-point scales were significantly correlated, thus demonstrating convergent validity.

The Utility of the GC Rating Scales

Given the significant correlation between the scales, the next section focusses on the utility of the scales in general, demonstrating an already extant base of studies that have used the 5-point or 4-point scales to characterize the global coherence of a variety of populations. Researchers have used the GC scales to further our understanding of: the relationship between macrolinguistic and microlinguistic processing, the cognitive-linguistic underpinnings of discourse coherence, and coherence as a diagnostic or treatment outcome measure.

First, global coherence rating scales have helped further our understanding of macro- versus microlinguistic processing and potential patterns that exist among patient populations and healthy older adults. Global coherence, among other macrostructure measures, has been reported as more impaired than microstructure measures in Alzheimer’s dementia (Glosser & Deser, 1990), traumatic brain injury (Hough & Barrow, 2003) and healthy older adults (Glosser & Deser, 1992), but in closed head injury patients, microstructure and macrostructure measures have been found to be similarly impaired (Glosser & Deser, 1990). In patients with stroke-related fluent aphasia, global coherence has been reported as preserved compared to microstructure impairments (Glosser & Deser, 1990). Moreover, Wright and Capilouto (2012) have reported that microlinguistic processes contribute to the maintenance of global coherence in people with aphasia.

Second, global coherence scales have been included in several studies as one of the means to explore the underlying cognitive-linguistic processes that might contribute to discourse coherence. In their study comparing discourse in vascular and Alzheimer's dementia, Laine, Laakso, Vuorinen, and Rinne (1998) found associations between global coherence and semantic processing tasks, suggesting that global coherence was deeply rooted in conceptual/semantic processing. Rogalski and colleagues (2010) examined the role of coherence and cognition in post-stroke patients without aphasia and found correlations between global coherence and performance on tests measuring sustained attention and processing speed. Global coherence has also been found to correlate with a measure of executive function (Lê, Coelho, Mozeiko, Krueger, & Grafman, 2014) and tasks measuring selective attention and episodic memory (Wright, Koutsoftas, Capilouto, & Fergadiotis, 2014). Associations between global coherence and working memory have been found in aphasia (Cahana-Amitay & Jenkins, 2018) and in individuals with damage to the dorsolateral prefrontal cortex (Coelho, Lê, Mozeiko, Krueger, & Grafman, 2012; Lê et al., 2014). In contrast, no associations between global coherence and working memory have been found in focal neurodegenerative disease (Gola et al., 2015) or stroke survivors without aphasia (Rogalski et al., 2010). Surprisingly, negative correlations have been found between global coherence and working memory in cognitively healthy older adults on discourse tasks requiring explanations of procedures (e.g., how to make a peanut butter and jelly sandwich) (Wright et al., 2014).

Third, global coherence scales have been used in studies investigating the contribution of localized lesions on cognitive and discourse functioning. Kurczek and Duff (2011) compared participants with chronic hippocampal amnesia characterized by severe deficits in declarative memory but preserved cognitive functioning to control participants and documented differences in local coherence but no differences in global coherence scores. These findings suggest that declarative memory plays a potential role in some macrolinguistic processes of discourse, but not in global coherence processing. Replicating their 2011 methods, Kurczek and Duff (2012) compared patients with bilateral ventromedial prefrontal cortex lesions to healthy controls and found no differences in global coherence between the patient and control groups. Based on these findings they propose that the bilateral ventromedial prefrontal cortices subserve functions that are more affective in nature rather than cognitive. Additionally, Coelho and colleagues (2012) reported that participants with injury to the left dorsolateral prefrontal cortex scored more poorly on global coherence than a comparison group, indicating that global coherence impairments can result from focal damage (Coelho et al., 2012). A follow-up study indicated that regardless of the lesion locale of a penetrating head injury, global coherence impairments were found to be similar to a comparison group of individuals with closed head injury (Coelho et al., 2013).

As has been noted, global coherence is a macrolinguistic processing measure that can be distinguished from microlinguistic processing in certain populations. Additionally, many of the above-mentioned studies have shown a link between global coherence and cognitive processing using the global coherence scales. Together, these findings suggest that incorporation of a global coherence scale into a clinical evaluation battery could be useful diagnostically and could inform patterns of recovery. Indeed, Coelho and Flewellyn (2003) used the global coherence scales to longitudinally track macrolinguistic progress in a single

patient with anomic aphasia and found no change in global coherence compared to improvements in microlinguistic abilities. The authors stressed the importance of including a macrostructure measure such as global coherence as part of an aphasia evaluation, since focal lesions such as stroke can result in macrostructure impairments similar to patients with diffuse lesions, such as Alzheimer's dementia and closed head injury (Coelho & Flewellyn, 2003).

Just as using a global coherence scale could help inform diagnosis, it could also be beneficial as an outcome measure for discourse-based therapies. Rogalski and Edmonds (2008) developed a cognitive-linguistic discourse-based treatment titled Attentive Reading and Constrained Summarization (ARCS) and explored the effects of ARCS on discourse production in an individual with primary progressive aphasia. In addition to microlinguistic constraints (e.g., refrain from using pronouns and nonspecific words during summaries), ARCS incorporates a macrolinguistic constraint (e.g., refrain from using opinion during summaries) with the intent to improve coherence. Results of the case study indicated that global coherence was improved post-treatment with ARCS in one person with primary progressive aphasia who had concomitant attentional difficulties (Rogalski & Edmonds, 2008). Lastly, Attentive Reading with Constrained Summarization–Written (ARCS-W; Obermeyer & Edmonds, 2018) was developed to address writing treatment at the paragraph level in people with mild aphasia. Obermeyer and colleagues used a global coherence rating scale as one of their macrolinguistic outcome measures and found increases in global coherence for spoken discourse (Participant 1 and 2) and for written discourse (Participant 2) following treatment with ARCS-W (Obermeyer, Rogalski, & Edmonds, Under Review).

The Current Study

Despite having been used in a variety of studies, neither of the global coherence rating scales has yet been tested to determine its clinical meaningfulness. In other words, if a client receives an overall global coherence score of 2.75 on the 4-point scale, how does this correspond to what the average listener hears? If the same client then increases that global coherence score from 2.75 to 3.66 post-treatment, would that demonstrate a clinically meaningful change that the average listener would be able to detect? Fundamentally, would a change in the global coherence rating scale affect an average listeners' qualitative level of attention/interest?

Several studies have used untrained raters' judgments of discourse to assess the external validity of their treatments (e.g., Ballard & Thompson, 1999; Cupit, Rochon, Leonard, & Laird, 2010; Hickey & Rondeau, 2005; Jacobs, 2001; Lustig & Tompkins, 2002; Ross & Wertz, 1999) or of their trained discourse analysis techniques (Body & Perkins, 1998; Doyle, Tsironas, Goda, & Kalinyak, 1996; Kong, Linnik, Law, & Shum, 2018; Kong & Wong, 2018). External validity can provide a measure of clinical meaningfulness to the analysis of discourse and can strengthen the interpretability of trained measures. For example, Doyle et al. (1996) found that trained and untrained measures of informativeness were highly correlated, and that trained measures accurately predicted how unfamiliar individuals would perceive the informativeness of people with aphasia. Body and Perkins also found that the discourse ratings of trained and untrained individuals were highly

correlated (Body & Perkins, 1998) and that linguistic measures predicted trained and untrained raters' judgment of content and clarity in the discourse of individuals with traumatic brain injury (Body & Perkins, 2004). Finally, Kong and colleagues (2018) reported that untrained listeners judged people with aphasia's discourse as having a lower level of coherence and these untrained ratings of coherence were correlated with a macrostructure framework capturing several elements thought to contribute to coherence.

Past studies have similarly found a positive relationship between trained measures of discourse and untrained listeners' ratings of discourse quality (e.g., Christensen, Wright, Ross, Katz, & Capilouto, 2009; Doyle et al., 1996; Olness, Ulatowska, Carpenter, Williams-Hubbard, & Dykes, 2005). In Doyle et al.'s (1996) study, correlations were found between untrained listeners' judgments of informativeness and objective measures of informativeness in discourse elicited from people with aphasia. Christensen et al., (2009) found that their naïve raters' perceptions of discourse quality differed depending on the discourse elicitation task used. Both Christensen et al. (2009) and Olness et al. (2005) found that longer stories were rated more favorably than those with fewer words. Critically, listener perceptions of discourse in individuals with speech and language disorders matter, as unfavorably rated discourse has been shown to affect the perception of traits such as employability, intelligence, and self-esteem (Allard & Williams, 2008).

To our knowledge, no studies have explored the relationship between untrained listeners' perceived judgment of story quality (e.g., how well a story holds their attention and interest) and a validated global coherence rating scale. Although the aforementioned studies indicate a positive relationship between objective measures of discourse and naïve raters' perceptions, a review of literature that is more closely related to global coherence indicates a relationship that is less clear. The term "off-topic verbosity" or "off topic speech" (OTS) (Arbuckle & Gold, 1993) can be considered a correlate of global coherence (i.e., the greater the inclusion of irrelevant speech, the greater the impairment in topic maintenance or global coherence). Several studies on OTS have incorporated ratings of story quality but increases in OTS or tangential discourse are not necessarily rated as having poorer quality, and in fact may be considered more interesting than stories that are on target. For example, James, Burke, Austin, and Hulme (1998) revealed that although the older adults in their study produced more verbose and less focused narratives, their autobiographical stories were rated as having higher quality than the young adults' narratives. In Beaudreau, Storandt, and Strube's study (2005), few differences were found between the number of irrelevant utterances and story quality ratings.

In the current study, we explored how untrained listeners would rate extended monologues that had been scored for global coherence by a trained rater using the 4-point GC scale. The listeners were asked to rate the samples on their topic maintenance and their inclusion of unnecessary information. Our rationale is that staying focused on a topic is considered the operational definition of global coherence according to Agar and Hobbs (1982), and inclusion of unnecessary information detracts from topic focus. If untrained listeners are able to judge differences among levels of discourse delineated by the 4-point GC ratings, this would provide a degree of external validity for the 4-point GC scale. We also wanted to explore whether an association exists between global coherence and listener judgments of

discourse quality with regard to attention/interest. If the untrained raters' judgements of discourse quality are strongly associated with trained levels of global coherence, it could have important clinical implications for how individuals with global discourse impairments are perceived by the average listener. If a listener's attention and interest decline with decreases in global coherence, then deficits in global coherence could substantially impede communicative effectiveness.

Specific Aims

The study was designed to answer the following questions:

1. Do untrained listeners differentiate between recorded discourse samples representing low, medium, and high levels of global coherence when using a Likert Scale to rate for a) topic maintenance, and b) inclusion of unnecessary information?
2. Is there a relationship between trained ratings of global coherence and untrained listeners' ratings of a) topic maintenance and b) inclusion of unnecessary information?
3. Do untrained listeners differentiate between recorded discourse samples representing low, medium, and high levels of global coherence when using a Likert Scale to rate their level of interest and attention?
4. Is there a relationship between trained ratings of global coherence and untrained listeners' ratings of their level of attention and interest?

Method

Discourse Stimuli Preparation and Coding

Initial Discourse Samples.—The discourse samples used in the current study are secondary data that were initially collected during the pre-treatment phase of a larger set of studies on the effects of dual tasks on cognitive and language performance in healthy older adults and adults with Parkinson's disease (Altmann et al., 2016; Altmann et al., 2015; Wilson, 2013). A complete description of the participants and procedures for those studies is available in Wilson's (2013) dissertation. All personal identifying information was removed from the original data set and transcripts before use in the current study, which was approved as exempt from review by the Institutional Review Board of Ithaca College.

Discourse was elicited in the initial group of studies (Altmann et al., 2016; Altmann et al., 2015; Wilson, 2013) by asking one of four open-ended questions that have been published in previous studies examining expressive discourse in healthy young and older adults (Kemper, Herman, & Lian, 2003; Kemper, McDowd, Pohl, Herman, & Jackson, 2006; Kemper, Schmalzried, Herman, Leedahl, & Mohankumar, 2009) and in adults who have survived a stroke but did not have aphasia (Rogalski et al., 2010). These four questions were: 1) "What do you think is the most important invention of the last 100 years and why?"; 2) "What do you think is the most important event of the last 100 years and why?"; 3) "Tell me about a

person who had an important impact in your life”; and 4) “Tell me about a vacation or event from your past that you remember very well.”

Participants were told they would be given a topic, and they would have three minutes to talk about it, during which time the prompt was visible on a screen. Participants’ vocal responses were recorded and the digital files were transcribed verbatim by trained research assistants from the last authors’ Language Over the Lifespan Laboratory at the University of Florida. Personal identifiers were removed from the transcripts and data set before sending the information to the first author’s Adult Cognition and Language Lab at Ithaca College for discourse coding.

Trained Global Coherence Rater.—The third author, a master’s level graduate student, was trained on the methods of C-unit segmentation and global coherence scoring according to the comprehensive training procedures developed by Wright and colleagues and used with their permission for the current study. The procedures for training include definitions of the terms “C-unit” and “global coherence,” rules for segmenting and scoring, examples of correct and incorrect segmentation and global coherence scores, and transcripts from different types of discourse (e.g., a wordless picture book description, a “recount” or retelling of an event, and procedural discourse) to be used for practice. Briefly, the definition of a C-unit is any independent clause plus any related clauses (Loban, 1976). For example: “I laughed so hard my belly hurt” would be 1 C-unit; whereas, “My belly hurt like crazy BUT I couldn’t stop laughing” would be 2 C-units. Each C-unit is then scored on a scale of 1 – 4 in relation to the topic of the discourse, where 1 = “The utterance is entirely unrelated to the stimulus or topic” and 4 = “The utterance is overtly related to the stimulus as defined by the mention of actors, actions, and/or objects present in the stimulus which are of significant importance to the main details of the stimulus” (Wright et al., 2013, p. 252). A global coherence score is determined when the sum of the global coherence scores of the individual C-units is divided by the total number of C-units. In order to pass the training procedures, the third author needed to achieve a minimum of 80% agreement between her practice transcripts for C-unit segmentation and global coherence scores, and the answer key provided by Wright and colleagues. Any disagreements were discussed by and resolved between the third author and the first author, an experienced global coherence rater. From this point forward, the third author will be referred to as the trained rater.

One hundred transcripts from the original data set (Altmann et al., 2016; Altmann et al., 2015; Wilson, 2013) were available to be coded by the trained rater. The trained rater first divided each transcript into C-units, then scored each transcript for global coherence according to Wright and colleagues’ 4-point rating scale (Wright et al., 2013). A random selection of 20% of the 100 transcripts were then re-scored for reliability by the trained rater and by an additional trained rater, a master’s level student from the first author’s laboratory. Intra-rater reliability for global coherence was 91.4% and inter-rater reliability was 81%. Previously reported ranges for global coherence reliability are commensurate with the current study’s reliability findings (Rogalski et al., 2010; Wright et al., 2014).

Discourse Samples and Survey Used in the Current Study.—Of the 100 discourse transcripts scored for global coherence, the trained rater searched for those with ratings

scores that met the criterion of low (less than 2 out of 4), medium (ratings between 2.5 and 3) and high (between 3.8 and 4). She selected 12 to be recorded in her own voice: 4 low (range = 1.52 – 1.79), 4 medium (range = 2.69 – 2.74) and 4 high (range = 3.86 – 3.93). The 12 transcripts were from 11 native English speaking participants (mean age = 68.7 years; mean level of education = 18.6 years), of whom 8 had a diagnosis of Parkinson's disease and 3 were non-brain injured age-matched older adults. The trained rater removed filled and unfilled pauses from the transcripts and corrected articulatory errors so that the focus of each transcript was on the coherence level and not on any distracting features. Each transcript was then recorded by the trained rater, a woman in her twenties, using a Marantz digital recorder. This was done to avoid any listener bias toward speaking patterns or behaviors from the original recordings (e.g., listeners giving poorer discourse ratings to speakers who included multiple fillers) and to protect the original speakers from being identified by their recordings. When recording, the trained rater attempted to maintain the same level of loudness and degree of prosodic range across samples to avoid having the listeners bias their ratings toward vocal performance.

Statistical analyses were completed on the discourse samples with an Alpha level set at .05. A Shapiro Wilks test for small sample sizes revealed that the number of words (word count) and the number of seconds taken to read aloud the discourse passages (spoken time) met assumptions of normality, thus an analysis of variance (ANOVA) test was used. There were no significant differences in word counts among the low, medium, or high GC samples, $F(2, 9) = 1.84, p = .21$, and no significant differences in spoken time among the low, medium, and high GC passages, $F(2, 9) = .93, p = .43$. Please see Table 1 for the descriptive and quantitative data related to the discourse samples. The 12 transcripts, the topic questions used to elicit the discourse samples, the computed global coherence scores by the trained rater (single score per sample) and the global coherence levels (low, medium, high) of each discourse sample are available as Supplemental Material.

In addition to recording the discourse samples, the trained rater developed a rating form for the untrained listeners (see Appendix). The rating form included six questions, each of which was to be rated in conjunction with the discourse sample on a scale of 1–8 (1= “not at all” and 8= “completely” or “all the time”). Three of the questions were the focus of the current study. Two questions targeted global coherence (questions 1 and 4) and one was related to listener levels of attention and interest (question 5). The remaining three questions, although included on the rating form, were developed to target local coherence and were not analyzed as part of the current study.

Participants

Untrained raters were a convenience sample of 24 individuals (18 female, 6 male), aged 18–54 (mean age = 27.5 years), who were recruited by flyers and announcements posted at Ithaca College. Education levels ranged from one year of college/university to completion of a Master's degree (mean education level = 14.75 years). All were native speakers of English and none had a history of a learning disability, or any cognitive, speech, language, or hearing impairments. Twenty participants self-reported as Caucasian, two were Asian, one was African American, and one was Hispanic. None of the participants reported any previous

experience with discourse analysis. The study was approved by the Ithaca College Institutional Review Board. Please see Table 2 for the untrained raters' demographic information.

Procedure

Untrained raters attended individual or group sessions that lasted approximately one hour. They were not compensated for their time. They were told that they would be listening to twelve language samples from real people that had been re-recorded in a woman's voice. Participants were given the topic of each discourse sample prior to listening to it. Topics included 1) favorite vacation, 2) most important event in the last 100 years, 3) most important invention of the past 100 years, and 4) person who had an important impact on your life. They were instructed to listen to each sample until it was complete, then rate that sample using a rating form (See Appendix). Each participant received 12 rating forms printed on 8.5" × 12" paper, one for each of the 12 discourse samples.

The order of presentation of the 12 recorded discourse samples was pseudorandomized to four different presentation clusters of three blocks, with each block containing one low global coherence, one medium global coherence, and one high global coherence sample. The four presentation clusters were counter-balanced such that each participant or participant group listened to a different presentation cluster than the previous participant or participant group.

Data Preparation for Analysis

The untrained listeners' rating form consisted of six questions, three were chosen a priori to address the specific aims of the current study. Two of the six questions on the rating form address the first and second aims of the study, which focus on global coherence: "Did the speaker stay focused on the topic?" and "Did the speaker include unnecessary information?" The "unnecessary information" question was reverse coded (8=1, 7=2 etc.), such that lower numbers indicated greater inclusion of unnecessary information, and higher numbers indicated less inclusion of unnecessary information. The reverse coding was done to ease interpretability of the correlations, where higher ratings on the scale would uniformly mean qualitatively better discourse for each of the outcome measures. The question: "Did the speaker hold your attention and interest?" was analyzed to address the third and fourth research aims.

Statistical analyses were completed using an alpha level of .05. Friedman non-parametric tests of differences among repeated measures of three related samples (low, medium, and high GC) were used. Post hoc orthogonal contrasts were performed using the Wilcoxon Signed-Ranks tests with Bonferroni corrections set at .017. To examine the relationships between variables, we used the Spearman Rank Order test of correlations.

Results

The first aim was to determine if untrained listeners differentiate between recorded discourse samples representing low, medium, and high levels of global coherence when using a Likert Scale to rate for topic maintenance and inclusion of unnecessary information. For topic

maintenance, a non-parametric Friedman test rendered a Chi-square value of 42.35 which was significant, $p < .001$, indicating there were differences among the three ranks. Comparisons were significant between low and medium, $Z = -4.11$, $p < .001$, $r = -.84$; medium and high, $Z = -3.46$, $p = .001$, $r = -.71$; and low and high global coherence levels, $Z = -4.29$, $p < .001$, $r = -.88$. See Figure 1. For inclusion of unnecessary information, a Friedman test resulted in a Chi-square value of 28.44 which was significant, $p < .001$, with significant comparisons between low and medium, $Z = -3.96$, $p < .001$, $r = -.81$; medium and high, $Z = -2.84$, $p = .005$, $r = -.58$; and low and high global coherence, $Z = -4.25$, $p < .001$, $r = -.87$. See Figure 2.

To further support the first aim and to address the second aim, Spearman rank-order correlations indicated that trained ratings of global coherence had strong positive correlations with topic maintenance, $r_s(12) = .94$, $p < .001$, and moderate positive correlations with inclusion of unnecessary information (reversed scored such that higher ratings indicate less inclusion of unnecessary information), $r_s(12) = .57$, $p = .05$. See Table 3.

The third aim was to determine if untrained listeners differentiate between recorded discourse samples representing low, medium, and high levels of global coherence when using a Likert Scale to rate their level of interest and attention. A Friedman nonparametric test revealed a Chi-square value of 22.16 which was significant, $p < .001$. Subsequent Wilcoxon Signed Ranks tests indicated comparisons were not significant between low and medium global coherence samples, $Z = -1.68$, $p = .093$, $r = -.34$, but were significant between medium and high, $Z = -3.57$, $p = .001$, $r = -.71$, and low and high global coherence samples, $Z = -4.08$, $p < .001$, $r = -.83$. See Figure 3.

To address aim 4, to examine if there is a relationship between untrained listeners' ratings of attention/interest and trained ratings of global coherence, Spearman nonparametric correlations were performed. Results indicated strong positive correlations between untrained listener ratings of interest/attention and trained ratings of global coherence, $r_s(12) = .63$, $p = .029$. See Table 3.

Discussion

In the current study we explored the external validity of global coherence operationalized as a 4-point rating scale that has previously demonstrated measurement reliability and convergent validity (Wright et al., 2014). A convenience sample of untrained participants of a variety of ages and backgrounds listened to and rated discourse samples previously analyzed for global coherence by a trained rater using the 4-point GC scale.

With regard to specific aims 1 and 2, we found that untrained listeners did differentiate between levels of global coherence in terms of their ratings of topic maintenance and inclusion of unnecessary information. For example, discourse samples rated as having high global coherence by a trained rater were also rated as having greater topic maintenance and less inclusion of unnecessary information by untrained listeners. Moreover, a relationship was found among global coherence, topic maintenance, and inclusion of unnecessary

information, indicating greater trained ratings of global coherence were associated with greater untrained ratings of topic maintenance and less inclusion of unnecessary information. Our findings are consistent with previous external validity studies that have demonstrated a relationship between analyses performed by trained raters and judgments of untrained listeners (e.g., Body & Perkins, 1998, 2004; Doyle et al., 1996; Kong et al., 2018; Kong & Wong, 2018).

In terms of specific aims 3 and 4, results indicated that discourse samples with high levels of global coherence were rated highly by untrained listeners on a scale of attention and interest; whereas, the low and medium level global coherence samples were rated as having lower levels of attention and interest. Thus, the untrained listeners judged that the stories that held their attention and interest more were only those that had high global coherence ratings. These findings were supported by a strong positive relationship between trained ratings of global coherence and untrained listeners ratings of attention and interest and are consistent with prior research reporting a relationship between trained measures of discourse analysis and naïve listeners' ratings of quality (e.g., Christensen et al., 2009; Olness et al., 2005).

Our findings are in contrast to those of James et al. (1998), whose participants rated the off topic personal narratives of older adults more positively than the more concise ones produced by younger adults. However, the older adults' narratives were more verbose, containing twice as many words. Relatedly, both Christensen et al., (2009) and Olness et al., (2005) found that stories rated as having higher quality had more words than those rated as having lower quality. In our study, the 12 discourse samples were not statistically different in terms of word count or time taken to read the samples aloud, providing further support that the listeners were perceiving differences in quality based on levels of global coherence, and not on other discourse features.

It is not clear why listeners' attention/interest ratings were similarly low for passages with low and medium global coherence, despite having ratings that differentiate between low, medium, and high levels of topic maintenance and inclusion of unnecessary information. It could be that the line between low and medium global coherence quality is simply less distinguishable. It could also be that there are other components contributing to the ratings of attention/interest in the low and medium monologues that are not accounted for by global coherence. This requires further exploration.

The Clinical Utility of the GC Scales

Global coherence has been used as a measure to understand the differences between macro-versus microlinguistic processing, as a means to explore potential cognitive and linguistic contributions to discourse, and as a way of tracking progress in treatment and recovery. The 4-point and 5-point GC scales have been used to analyze the spoken discourse patterns of a number of different clinical and non-brain-injured populations (e.g., Cahana-Amitay & Jenkins, 2018; Coelho & Flewellyn, 2003; Coelho et al., 2012; Gola et al., 2015; Hough & Barrow, 2003; Kurczek & Duff, 2011, 2012; Laine et al., 1998; Lê et al., 2014; Rogalski et al., 2010; Rogalski & Edmonds, 2008; Sanchez & Spencer, 2013; Van Leer & Turkstra, 1999; Wright & Capilouto, 2012; Wright et al., 2013; Wright et al., 2014). The current study provides preliminary evidence of external validity for the 4-point GC scale. By using an

already-established global coherence analysis method, we are contributing to methodological consistency in a field where a lack of consistency in discourse methodology can greatly influence the interpretability of results, as noted by Ellis et al. (2016) and Linnik et al. (2016), and can make comparisons across studies more difficult (Dietz & Boyle, 2018).

Our findings support that the 4-point GC analysis technique can translate to meaningful differences in coherence judgments by untrained listeners. It is unclear, however, how often a global coherence rating scale is used in clinical settings. What we do know is that discourse analysis techniques in general are not widely implemented in clinical practice (Bryant et al., 2017; Frith et al., 2014; Maddy et al., 2015; Verna et al., 2009; Westerveld & Claessen, 2014). We also know that choosing an appropriate discourse analysis technique can be difficult and that there is a need to simplify this process for clinicians (Bryant et al., 2017). If clinicians knew that they could employ a relatively easy to use rating scale as a measure of pre- and post-treatment global coherence (see Rogalski & Edmonds, 2008), and that the measure would indicate meaningful changes in global coherence, they might be more likely to incorporate the use of a global coherence scale in their clinical practice. To further close the research-to-practice gap there is a need for more education and professional development training on discourse analysis methods (Bryant et al., 2017). Bryant and colleagues have begun to investigate the outcomes of a discourse analysis training program with participating speech-language pathology students (Bryant, Ferguson, Valentine, & Spencer, 2019). Similar training programs using the 4-point GC scale would be beneficial.

Of greatest importance, if untrained raters' judgements of discourse quality are strongly associated with global coherence, it could have substantial clinical implications for how individuals with global coherence impairments are perceived by the average listener. Allard and Williams (2008) have demonstrated that listener judgments of individuals with speech and language disorders confirm stereotypes and negatively influence perceptions of traits such as employability, intelligence, and self-esteem. If a listener's attention and interest decline with decreases in global coherence, then individuals' deficits in global coherence could potentially impede their communicative effectiveness. This finding would also indicate that treatment programs focused on strengthening global coherence (e.g., Obermeyer & Edmonds, 2018; Rogalski & Edmonds, 2008) are warranted and could result in important communicative benefits. Two techniques for improving global coherence are ARCS (Rogalski & Edmonds, 2008) and ARCS-Written (Obermeyer & Edmonds, 2018). Both are cognitive-linguistic therapy techniques that require reading small units of discourse (several sentences or one paragraph at a time) then summarizing orally or in writing using constraints that emphasize meaningful language. The ARCS and ARCS-W techniques in combination with pre- and post-treatment analyses of global coherence offer clinicians a way of treating and tracking global coherence.

Limitations and Future Directions

One limitation of the current study is that all discourse samples were "cleaned" of filled and unfilled pauses and articulatory errors. Additionally, they were all re-recorded by a single speaker in a similar manner. We purposefully controlled the stimuli so that listeners would focus on the content of the samples without the distraction of extraneous features (e.g.,

speech errors or differences in speaking styles). Understandably, this level of control over discourse is unnatural, not feasible, and not representative of what someone sounds like in the real world. Thus, it is unknown how the inclusion of discourse interruptions or differences in speaking styles or the presence of different types of communication impairments might influence a naïve listener's judgments of adherence to topic, inclusion of unnecessary information, and level of attention and interest. Future studies should address these questions.

Our study used a convenience sample of a wide variety of untrained listeners with a range of ages and educational backgrounds. Future directions should consider how the age of untrained listeners affects ratings of global coherence, as prior studies have found that younger and older adults may diverge in their ratings of a speaker's adherence to topic during personal narratives (James et al., 1998; Odatto & Keller-Cohen, 2009). Moreover, all of the narratives in the study were re-recorded in a young woman's voice. It would be interesting to compare listener judgments of different genders of speakers producing the same discourse samples, given that female speakers have been judged as more on-target during personal narratives than male speakers (Odatto & Keller-Cohen, 2009). Finally, the number of passage recording samples is relatively small, and though there were still significant findings, it would be useful to look at a larger sample.

Conclusion

This exploratory study contributes to the external validity to the 4-point GC rating scale, which has demonstrated convergent validity and reliability (Wright et al., 2013). Findings suggest that untrained listeners can accurately distinguish between discourse samples rated as high, medium and low global coherence by asking them to judge how well the speaker stayed on topic and whether the speaker included unnecessary information. Most importantly, the study suggests that passages with high global coherence are more likely to hold the listeners' attention and interest. These findings could have clinically meaningful implications for how clients with reduced topic maintenance are perceived by others. Thus, in addition to providing support for the incorporation of the 4-point GC scale into clinical assessment, improvements in global coherence should be considered an important treatment goal for those with acquired communication impairments.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was partially funded by the National Institute on Aging (R21AG0033284-01A2) at the National Institutes of Health, as well as the National Parkinson Foundation Center of Excellence, and the UF-INFORM database.

Appendix

Rating Scale for Untrained Listeners

1. Did the speaker stay focused on the topic?

- 1 2 3 4 5 6 7 8
(not at all) (completely)
2. Was the story easy to follow?
- 1 2 3 4 5 6 7 8
(not at all) (completely)
3. Did one sentence flow logically and continuously into another?
- 1 2 3 4 5 6 7 8
(not at all) (completely)
4. Did the speaker include unnecessary information?
- 1 2 3 4 5 6 7 8
(not at all) (all the time)
5. Did the speaker hold your attention and interest?
- 1 2 3 4 5 6 7 8
(not at all) (all the time)
6. Was the story full of details and expanding of ideas?
- 1 2 3 4 5 6 7 8
(not at all) (all the time)

References

- Agar M, & Hobbs JR (1982). Interpreting discourse: Coherence and the analysis of ethnographic interviews. *Discourse Processes*, 5(1), 1–32.
- Allard ER, & Williams DF (2008). Listeners' perceptions of speech and language disorders. *Journal of Communication Disorders*, 41(2), 108–123. [PubMed: 17586521]
- Altmann LJ, Stegemöller E, Hazamy AA, Wilson JP, Bowers D, Okun MS, & Hass CJ (2016). Aerobic exercise improves mood, cognition, and language function in Parkinson's disease: Results of a controlled study. *Journal of the International Neuropsychological Society*, 22(9), 1–12. [PubMed: 26714882]
- Altmann LJ, Stegemöller E, Hazamy AA, Wilson JP, Okun MS, McFarland NR, ... Hass CJ (2015). Unexpected dual task benefits on cycling in Parkinson disease and healthy adults: A neuro-behavioral model. *PloS one*, 10(5), 1–13.
- Arbuckle TY, & Gold DP (1993). Aging, inhibition, and verbosity. *Journals of Gerontology*, 48(5), 225–232.
- Ballard KJ, & Thompson CK (1999). Treatment and generalization of complex sentence production in agrammatism. *Journal of Speech, Language, and Hearing Research*, 42(3), 690–707.
- Beaudreau SA, Storandt M, & Strube MJ (2005). A comparison of narratives told by younger and older adults. *Experimental Aging Research*, 32(1), 105–117.
- Body R, & Perkins MR (1998). Ecological validity in assessment of discourse in traumatic brain injury: Ratings by clinicians and non clinicians. *Brain Injury*, 12(11), 963–976. [PubMed: 9839029]
- Body R, & Perkins MR (2004). Validation of linguistic analyses in narrative discourse after traumatic brain injury. *Brain Injury*, 18(7), 707–724. [PubMed: 15204331]
- Boyle M (2014). Test–retest stability of word retrieval in aphasic discourse. *Journal of Speech, Language, and Hearing Research*, 57(3), 966–978.
- Boyle M (2015). Stability of word-retrieval errors with the AphasiaBank stimuli. *American Journal of Speech-Language Pathology*, 24(4), S953–S960. [PubMed: 26151900]

- Bryant L, Ferguson A, Valentine M, & Spencer E (2019). Implementation of discourse analysis in aphasia: investigating the feasibility of a Knowledge-to-Action intervention. *Aphasiology*, 33(1), 31–57.
- Bryant L, Spencer E, & Ferguson A (2017). Clinical use of linguistic discourse analysis for the assessment of language in aphasia. *Aphasiology*, 31(10), 1105–1126.
- Cahana-Amitay D, & Jenkins T (2018). Working memory and discourse production in people with aphasia. *Journal of Neurolinguistics*, 48, 90–103.
- Christensen SC, Wright HH, Ross K, Katz R, & Capilouto G (2009). What makes a good story? The naive rater's perception. *Aphasiology*, 23(7–8), 898–913.
- Christie F, & Martin JR (Eds.). (2009). *Language, knowledge and pedagogy: Functional linguistic and sociological perspectives*. New York, NY: Continuum.
- Coelho CA, & Flewellyn L (2003). Longitudinal assessment of coherence in an adult with fluent aphasia: A follow-up study. *Aphasiology*, 17(2), 173–182.
- Coelho CA, Lê K, Mozeiko J, Hamilton M, Tyler E, Krueger F, & Grafman J (2013). Characterizing discourse deficits following penetrating head injury: A preliminary model. *American Journal of Speech-Language Pathology*, 22(2), S438–S448. [PubMed: 23695915]
- Coelho CA, Lê K, Mozeiko J, Krueger F, & Grafman J (2012). Discourse production following injury to the dorsolateral prefrontal cortex. *Neuropsychologia*, 50(14), 3564–3572. [PubMed: 22982512]
- Cupit J, Rochon E, Leonard C, & Laird L (2010). Social validation as a measure of improvement after aphasia treatment: Its usefulness and influencing factors. *Aphasiology*, 24(11), 1486–1500.
- Dietz A, & Boyle M (2018). Discourse measurement in aphasia: Consensus and caveats. *Aphasiology*, 32(4), 487–492.
- Doyle PJ, Tsironas D, Goda AJ, & Kalinyak M (1996). The relationship between objective measures and listeners' judgments of the communicative informativeness of the connected discourse of adults with aphasia. *American Journal of Speech-Language Pathology*, 5(3), 53–60.
- Ellis C, Henderson A, Wright HH, & Rogalski Y (2016). Global coherence during discourse production in adults: A review of the literature. *International Journal of Language and Communication Disorders*, 51(4), 359–367. [PubMed: 26899279]
- Frith M, Togher L, Ferguson A, Levick W, & Docking K (2014). Assessment practices of speech-language pathologists for cognitive communication disorders following traumatic brain injury in adults: An international survey. *Brain Injury*, 28(13–14), 1657–1666. [PubMed: 25158134]
- Glosser G, & Deser T (1987). Guidelines for rating discourse coherence. Unpublished rating scale.
- Glosser G, & Deser T (1990). Patterns of discourse production among neurological patients with fluent language disorders. *Brain and Language*, 40(1), 67–88.
- Glosser G, & Deser T (1992). A comparison of changes in macrolinguistic and microlinguistic aspects of discourse production in normal aging. *Journals of Gerontology*, 47(4), 266–272.
- Gola KA, Thorne A, Veldhuisen LD, Felix CM, Hankinson S, Pham J, ... Glenn S (2015). Neural substrates of spontaneous narrative production in focal neurodegenerative disease. *Neuropsychologia*, 79, 158–171. [PubMed: 26485159]
- Grimes JE (1975). *The thread of discourse*. The Hague: Mouton.
- Halliday MAK, Matthiessen C, & Halliday M (2004). *An introduction to functional grammar*. New York: Oxford University Press Inc.
- Harris ZS (2013). *Papers in structural and transformational linguistics*. Philadelphia: Springer.
- Hickey E, & Rondeau G (2005). Social validation in aphasiology: Does judges' knowledge of aphasiology matter? *Aphasiology*, 19(3–5), 389–398.
- Hough MS, & Barrow I (2003). Descriptive discourse abilities of traumatic brain-injured adults. *Aphasiology*, 17(2), 183–191.
- Hunt KW (1965). *Grammatical structures written at three grade levels*. Champaign, IL: National Council of Teachers of English.
- Jacobs BJ (2001). Social validity of changes in informativeness and efficiency of aphasic discourse following Linguistic Specific Treatment (LST). *Brain & Language*, 78, 115–127. [PubMed: 11412020]

- James LE, Burke DM, Austin A, & Hulme E (1998). Production and perception of 'verbosity' in younger and older adults. *Psychology and Aging*, 13(3), 355–367. [PubMed: 9793112]
- Kemper S, Herman RE, & Lian CH (2003). The costs of doing two things at once for young and older adults: talking while walking, finger tapping, and ignoring speech or noise. *Psychology and Aging*, 18(2), 181–192. [PubMed: 12825768]
- Kemper S, McDowd J, Pohl P, Herman R, & Jackson S (2006). Revealing language deficits following stroke: The cost of doing two things at once. *Aging, Neuropsychology and Cognition*, 13(1), 115–139.
- Kemper S, Schmalzried R, Herman R, Leedahl S, & Mohankumar D (2009). The effects of aging and dual task demands on language production. *Aging, Neuropsychology, and Cognition*, 16(3), 241–259.
- Kong AP-H, Linnik A, Law S-P, & Shum WW-M (2018). Measuring discourse coherence in anomic aphasia using Rhetorical Structure Theory. *International Journal of Speech-Language Pathology*, 20(4), 406–421. [PubMed: 28306394]
- Kong AP-H, & Wong CW-Y (2018). An Integrative Analysis of Spontaneous Storytelling Discourse in Aphasia: Relationship With Listeners' Rating and Prediction of Severity and Fluency Status of Aphasia. *American Journal of Speech-Language Pathology*, 27(4), 1491–1505. [PubMed: 30458505]
- Kurczek J, & Duff MC (2011). Cohesion, coherence, and declarative memory: Discourse patterns in individuals with hippocampal amnesia. *Aphasiology*, 25(6–7), 700–712. [PubMed: 23136461]
- Kurczek J, & Duff MC (2012). Intact discourse cohesion and coherence following bilateral ventromedial prefrontal cortex. *Brain and Language*, 123(3), 222–227. [PubMed: 23102898]
- Laine M, Laakso M, Vuorinen E, & Rinne J (1998). Coherence and informativeness of discourse in two dementia types. *Journal of Neurolinguistics*, 11(1–2), 79–87.
- Lê K, Coelho C, Mozeiko J, Krueger F, & Grafman J (2014). Does brain volume loss predict cognitive and narrative discourse performance following traumatic brain injury? *American Journal of Speech-Language Pathology*, 23(2), S271–S284. [PubMed: 24686463]
- Linnik A, Bastiaanse R, & Höhle B (2016). Discourse production in aphasia: A current review of theoretical and methodological challenges. *Aphasiology*, 30(7), 1–36.
- Loban W (1963). *The language of elementary school children*. Champaign, IL: National Council of Teachers of English.
- Loban W (1976). *Language development: Kindergarten through grade twelve* (Vol. 18). Urbana, IL: National Council of Teachers of English.
- Lustig AP, & Tompkins CA (2002). A written communication strategy for a speaker with aphasia and apraxia of speech: Treatment outcomes and social validity. *Aphasiology*, 16(4–6), 507–521.
- Maddy KM, Howell DM, & Capilouto GJ (2015). Current practices regarding discourse analysis and treatment following non-aphasic brain injury: A qualitative study. *Journal of Interactional Research in Communication Disorders*, 6(2), 211–236.
- Obermeyer JA, & Edmonds LA (2018). Attentive Reading with Constrained Summarization adapted to address written discourse in people with mild aphasia. *American Journal of Speech-Language Pathology*, 27, 392–405. [PubMed: 29497751]
- Obermeyer JA, Rogalski Y, & Edmonds LA (Under Review). Attentive Reading with Constrained Summarization-Written, a multi-modality discourse level treatment for mild aphasia. *Aphasiology*.
- Odato CV, & Keller-Cohen D (2009). Evaluating the speech of younger and older adults: Age, gender, and speech situation. *Journal of Language and Social Psychology*, 28(4), 457–475.
- Olness GS, Ulatowska HK, Carpenter CM, Williams-Hubbard LJ, & Dykes JC (2005). Holistic assessment of narrative quality: A social validation study. *Aphasiology*, 19(3–5), 251–262.
- Rogalski Y, Altmann LJ, Plummer-D'Amato P, Behrman AL, & Marsiske M (2010). Discourse coherence and cognition after stroke: A dual task study. *Journal of Communication Disorders*, 43(3), 212–224. [PubMed: 20219209]
- Rogalski Y, & Edmonds LA (2008). Attentive Reading and Constrained Summarisation (ARCS) treatment in primary progressive aphasia: A case study. *Aphasiology*, 22(7–8), 763–775.
- Ross KB, & Wertz RT (1999). Comparison of impairment and disability measures for assessing severity of, and improvement in, aphasia. *Aphasiology*, 13, 113–124.

- Sanchez J, & Spencer KA (2013). Preliminary evidence of discourse improvement with dopaminergic medication. *Advances in Parkinson's Disease*, 2(02), 37–42.
- van Dijk TA, & Kintsch E (1983). *Strategies of discourse comprehension*. New York, NY: Academic Press.
- Van Leer E, & Turkstra L (1999). The effect of elicitation task on discourse coherence and cohesion in adolescents with brain injury. *Journal of Communication Disorders*, 32(5), 327–349. [PubMed: 10498013]
- Verna A, Davidson B, & Rose T (2009). Speech-language pathology services for people with aphasia: A survey of current practice in Australia. *International Journal of Speech-Language Pathology*, 11(3), 191–205.
- Westerveld MF, & Claessen M (2014). Clinician survey of language sampling practices in Australia. *International Journal of Speech-Language Pathology*, 16(3), 242–249. [PubMed: 24447161]
- Wilson J (2013). Dual task effects on language production in sentence and discourse contexts in Parkinson's disease (Unpublished doctoral dissertation). University of Florida: Gainesville, FL.
- Wright HH, & Capilouto GJ (2012). Considering a multi-level approach to understanding maintenance of global coherence in adults with aphasia. *Aphasiology*, 26(5), 656–672. [PubMed: 22962509]
- Wright HH, Capilouto GJ, & Koutsoftas A (2013). Evaluating measures of global coherence ability in stories in adults. *International Journal of Language and Communication Disorders*, 48(3), 249–256. [PubMed: 23650882]
- Wright HH, Koutsoftas AD, Capilouto GJ, & Fergadiotis G (2014). Global coherence in younger and older adults: Influence of cognitive processes and discourse type. *Aging, Neuropsychology, and Cognition*, 21(2), 174–196.

Figure 1.

Untrained listener ratings of topic maintenance on discourse representing low, medium, and high trained ratings of global coherence (GC). Boxes indicate the interquartile range, lines indicate the median, and whiskers indicate the minimum and maximum. Comparisons were significant among low, medium, and high GC levels.

Figure 2.

Untrained listener ratings of unnecessary information on discourse representing low, medium, and high trained ratings of global coherence (GC). Boxes illustrate the interquartile range, lines indicate the median, and whiskers indicate the minimum and maximum. Comparisons were significant among low, medium, and high GC levels.

Note. *Higher scores indicate less inclusion of unnecessary information.

Figure 3.

Untrained listener ratings of attention/interest on discourse representing low, medium, and high trained ratings of global coherence (GC). Boxes indicate the interquartile range, lines indicate the median, and whiskers indicate the minimum and maximum. Comparisons were significant between low and high and between medium and high GC levels.

Table 1

Descriptive and quantitative data for the 12 discourse samples

GC Level	GC Trained Rater Score (/4)	Topic of Discourse Sample	Word Count	Spoken Time (seconds)
High	3.91	Vacation	439	160
High	3.86	Person	418	158
High	3.93	Vacation	448	177
High	3.88	Event	398	163
Medium	2.71	Vacation	272	105
Medium	2.69	Event	265	121
Medium	2.74	Event	426	166
Medium	2.70	Person	447	182
Low	1.79	Invention	351	150
Low	1.68	Event	373	143
Low	1.60	Invention	368	128
Low	1.52	Event	366	166

Note. Invention = “What do you think is the most important invention of the last 100 years and why?”; Event = “What do you think is the most important event of the last 100 years and why?”; Person = “Tell me about a person who had an important impact in your life”; Vacation = “Tell me about a vacation or event from your past that you remember very well.”

Table 2

Individual Characteristics of Untrained Listeners

Untrained Listener	Gender	Age	Race/Ethnicity	Years of Education	Occupation
1	Female	51	Caucasian	16	Student
2	Female	23	Caucasian	16	Student
3	Female	23	Caucasian	16	Student
4	Female	30	Caucasian	16	Student
5	Female	33	Caucasian	14	Student
6	Female	22	Caucasian	16	Student
7	Female	19	Caucasian	13	Student
8	Male	34	Caucasian	18	College Professor
9	Female	19	Caucasian	13	Student
10	Male	18	Hispanic	12	Student
11	Female	18	Asian	12	Student
12	Female	18	African American	13	Student
13	Female	51	Caucasian	18	Dir of Resident Life
14	Female	18	Asian	12	Student
15	Male	18	Caucasian	12	Student
16	Female	54	Caucasian	14	Support Specialist for ITS
17	Female	27	Caucasian	16	Student
18	Female	26	Caucasian	16	Student
19	Female	24	Caucasian	16	Student
20	Female	23	Caucasian	16	Student
21	Male	20	Caucasian	13	Student
22	Male	39	Caucasian	16	Dir of Advancement Services
23	Female	18	Caucasian	12	Student
24	Male	34	Caucasian	18	Office of Admissions/Lecturer

Note. Dir = Director; ITS = Information Technology Services; Director of Advancement Services works with fundraising.

Table 3

Spearman rank order correlations among trained global coherence ratings and untrained listener ratings

Variable	2	3	4
1. Trained Ratings of Global Coherence	.75 **	.57 *	.63 *
2. Untrained Ratings of Topic Maintenance		.94 **	.69 *
3. Untrained Ratings of Inclusion of Unnecessary Info ^a			.50
4. Untrained Ratings of Interest/Attention			

Note. Correlation is significant at:

* p .05;

** p .01;

^aUntrained Ratings of Inclusion of Unnecessary Information is reversed scored. Positive correlations indicate less inclusion of unnecessary information.