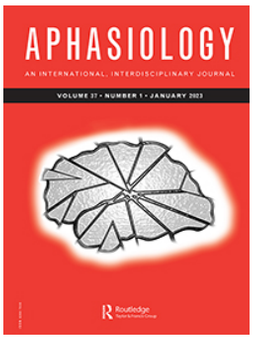

Aphasiology

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

Phase I test development for a brief assessment of transactional success in aphasia: methods and preliminary findings of main concepts in non-aphasic participants

Jacquie Kurland, Anna Liu & Polly Stokes

To cite this article: Jacquie Kurland, Anna Liu & Polly Stokes (2023) Phase I test development for a brief assessment of transactional success in aphasia: methods and preliminary findings of main concepts in non-aphasic participants, Aphasiology, 37:1, 39-68, DOI: 10.1080/02687038.2021.1988046

To link to this article: https://doi.org/10.1080/02687038.2021.1988046

APHASIOLOGY
2023, VOL. 37, NO. 1, 39–68
https://doi.org/10.1080/02687038.2021.1988046

Routledge
Taylor & Francis Group

Phase I test development for a brief assessment of transactional success in aphasia: methods and preliminary findings of main concepts in non-aphasic participants

Jacquie Kurland[a], Anna Liu[b] and Polly Stokes[a]

[a]Department of Communication Disorders, University of Massachusetts Amherst, Amherst, MA, USA;
[b]Department of Mathematics and Statistics, University of Massachusetts Amherst, Amherst, MA, USA

ABSTRACT

Background: One obstacle for clinicians and third-party payers embracing a participation-based framework for assessing and treating aphasia is the dearth of clinically convenient instruments for measuring change in functional communication. Traditional assessments often do not capture subtle improvements in communicative success. However, analyzing conversation and other discourse is too labor intensive to be a useful, practical tool in clinical settings.

Aims: The purpose of this study was to acquire a set of story-retelling normative references from a sample of non-aphasic volunteers, and to develop checklists for the Brief Assessment of Transactional Success in conversation in aphasia (BATS).

Methods & Procedures: We examined 768 narratives from a sample of 96 healthy, non-aphasic volunteers from three age cohorts. We focus here on one macrolinguistic measure of discourse analysis, main concepts (MCs), that assesses a person's ability to convey a story's gist. Forty-eight narratives were elicited from each of 16 short video and/or audio stimuli from four categories that varied in the degree of reliance on auditory comprehension for story gist. Transcripts were analyzed for MCs using the methods of Richardson and Dalton (2016, 2020).

Results: Our analysis generated checklists, including essential elements of MCs that were produced by at least 33% of the normative sample, along with examples of alternative productions. Reference thresholds were established for "non-normal" scores falling below the 5% quantile in the ratio of an MC composite score to the number of MCs (MCComp/MCs). Similar to earlier studies, a younger third of participants produced narratives that were scored significantly higher in the ratio of MCComp/MCs. Whereas we have hypothesized that the non-verbal video stimuli would ultimately prompt the most accurate and complete narratives in *aphasic* narrative retells, in the current non-clinical sample, we expected and found that the narrated stimuli were more likely to elicit accurate and complete main concepts.

***Conclusions*:** The next phase of development will involve testing the stimuli on a large clinical sample to acquire: 1) aphasic narratives; 2) topic-constrained conversations with non-aphasic

ARTICLE HISTORY
Received 8 February 2021
Accepted 27 September 2021

KEYWORDS
Aphasia; discourse analysis; main concepts; story retelling; transactional success

CONTACT Jacquie Kurland jkurland@comdis.umass.edu Department of Communication Disorders, University of Massachusetts Amherst, 358 North Pleasant Street, Amherst, MA 01003-9296, USA

conversation partners to establish intersubjectivity regarding story gist; and 3) conversation partner narratives. It is hoped that the BATS will become a popular, free, and accessible tool for clinicians and clinical researchers. Utilizing these short, engaging video/audio clips and checklists of MCs will help to narrow the chasm between standardized aphasia batteries and an elusive measure of communicative success.

Introduction

Although we take it for granted, few things are as fundamental to human existence as ordinary everyday conversation. The "basic and primary use of language" (Fillmore, 1981, p. 152) and the most frequent communicative activity of daily life for older adults (Davidson et al., 2003), we use it to express our desires and needs, exchange information with others, share our experience, and come to understand the world through other people's points of view. Moreover, in daily interactions with others, we create and maintain interpersonal connections, a cornerstone of achieving a satisfying quality of life (Ross & Wertz, 2003). For the estimated more than 2.5 million Americans living with aphasia (Simmons-Mackie, 2018), speech and language impairments drastically diminish quality of life, in part, because they reduce opportunities for interpersonal connection and create barriers to efficient, effective conversational interactions. This frequently leads to social isolation and a host of other negative psychosocial consequences (Brumfitt, 1993; Ferro et al., 2009; Wray & Clarke, 2017) that can exacerbate long-term disability (Teoh et al., 2009).

While conversation is critical to a meaningful life, there are virtually no valid, reliable instruments for directly measuring conversational treatment outcomes in aphasia that are also compatible with the conventional time constraints for assessing outcomes in clinical settings. Indirect or observational measures have been available for as long as clinical researchers have attempted to develop "functional" measures of communication. These include the Functional Communication Profile (FCP; Sarno, 1969), the Communicative Effectiveness Index (CETI; Lomas et al., 1989), and more recently, the Functional Outcome Questionnaire for Aphasia (FOQ-A; Glueckauf et al., 2003), the Communicative Activity Log (CAL; Pulvermuller & Berthier, 2008), the Aphasia Communication Outcome Measure (ACOM; Hula et al., 2015), and the Measures of Skill in Supported Conversation and of Participation in Conversation (MSC and MPC; Kagan et al., 2018, 2004). These observational profiles have been utilized to demonstrate the effects of therapy on everyday communication. Rated by the person with aphasia, a significant other, or a clinician, they are relatively quick and easy to administer. Some include skills and behaviors that influence interactional and transactional success in conversation. While these measures are useful in clinical research and practice, they are all subject to the same criticism of any observational profile, i.e., that they are subjective and indirect measures, and subject to the biases of the raters. Moreover, as Webster et al. (2015) note, these measures assess the impact of therapy rather than treatment effects per se.

There have been increasing efforts in the last two decades to address the limited tools available to clinicians to assess improvements in real-world communication. Although there is a lack of consensus on a clear definition of what real-world communication entails (Doedens & Meteyard, 2020), or how best to measure it in aphasia (Wallace et al., 2019), one helpful theoretical framework for guiding discourse-based assessment and intervention is the Linguistic Underpinnings of Narrative in Aphasia (LUNA; Dipper et al., 2021). The LUNA framework provides a holistic theory for conceptualizing the trouble that frequently occurs in aphasic discourse, whether due to, or exacerbated by, impairment at linguistic, propositional, planning, and/or pragmatic levels.

Heightened interest in discourse analysis may reflect the fact that models for assessing and treating aphasia have gradually moved from an impairment-based to a participation-based framework (Brady et al., 2016), with much recent focus on "life participation" approaches (Simmons-Mackie et al., 2014). A greater emphasis on what many still call "functional communication" has challenged the field to investigate therapeutic methods of promoting language recovery that can generalize to, if not focus directly on, language and communication skills in the real world. Conversation-based therapies that treat persons with aphasia and/or conversation partners can meet this challenge (Beeke et al., 2015; Carragher et al., 2015; DeDe et al., 2019; Elman, 2007; Elman & Bernstein-Ellis, 1999; Finch et al., 2020; Kagan, 1995, 1998; McVicker et al., 2009; Wilkinson, 2010; Wilkinson & Wielaert, 2012). Practical and reliable methods of assessing conversation-based treatment outcomes, however, are lagging.

Following their randomized controlled trial examining group size in conversation treatment in 46 persons with aphasia, DeDe et al. (2019) initially only reported on one of multiple discourse samples collected, not including conversation at all. Two years later, DeDe and Hoover (2021) compared four discourse measures across two individuals with aphasia from the RCT study, one mild and fluent, the other severe and nonfluent. Different measures were sensitive to the presence of aphasia and effects of treatment in these two individuals. It is hoped that analysis of the rich database of conversations acquired during their study is ongoing. However, a plausible reason for analyzing the content of picture descriptions rather than conversations may come down to very practical issues such as funding and time to complete the complex training and labor-intensive analysis of aphasic conversation. Analyzing monologic aphasic discourse is already time- and cost-prohibitive (Prins & Bastiaanse, 2004). Clinicians also blame lack of training, expertise, and resources (Bryant et al., 2017) for avoiding the use of detailed transcription-based discourse analysis. As DeDe and Hoover note in citing a recent survey, multiple constraints discourage clinicians from using discourse measures at all, including clinician uncertainty about selecting appropriate measures, and most notably, time (M Cruice et al., 2020).

Perhaps owing to these major obstacles, when discourse is included in outcomes measures, it is often limited to monologic descriptions of single or series of pictures, procedural discourse, or the Cinderella story retell (e.g., MacWhinney et al., 2011). One exception is the Story Retell Procedure (SRP; Doyle et al., 2000, 1998; McNeil et al., 2001, 2002). The SRP, which uses auditorily presented stories, with and without the support of pictures (Doyle et al., 1998), and includes four parallel forms to avoid undesirable learning effects in longitudinal testing, has been shown to be a useful tool for acquiring narrative

discourse samples in aphasia. All of these discourse elicitation tasks capture some essential aspects of the ability to convey information during monologic connected speech tasks (Fergadiotis et al., 2019).

It is more and more common for measures of discourse to be included as an outcome measure in aphasia treatment research. A recent review identified 165 studies over the last four decades that included 536 linguistic measures utilized to examine performance in aphasic discourse (Bryant et al., 2016). It is no wonder that Dietz and Boyle suggest that the field has reached a "tipping point" in the number and diversity of methods for eliciting, scoring, and analyzing discourse, such that it is nearly impossible to compare and synthesize treatment intervention studies (Dietz & Boyle, 2018). Our contention is that, not only is it time to establish a core outcome set for discourse (D-COS) in aphasia research (Dietz & Boyle, 2018; Stark et al., 2020), but such a D-COS must include conversational discourse if we are to understand how people with aphasia manage the co-construction of shared meaning in everyday conversation (Kurland & Stokes, 2018).

In order to gain a holistic impression of a person's linguistic skills both contexts of monologue and dialogue should be examined (Armstrong et al., 2011). Even when discourse analysis is limited to grammatical profiles, it is clear that grammar in the context of clinical assessments including picture descriptions and the Cinderella story retelling differs from grammar in conversation (Beeke et al., 2003). From a pragmatic perspective, monologic tasks cannot capture the dyadic, interactive nature of real-world communication. They lack the communicative intention of reaching a shared interpretation (Klippi, 1996) – something that occurs naturally in ordinary conversation (Sacks et al., 1974). As such, they also fail to capture the collaborative nature of everyday communication (Clark & Wilkes-Gibbs, 1986).

Some measures have sought to make assessment of conversational gains more accessible to clinicians and clinical researchers. While making important contributions to this effort, tools such as the Conversation Analysis Profile for People with Aphasia (CAPPA; Whitworth et al., 1997), the Profile of Word Errors and Retrieval in Speech (POWERS; Herbert et al., 2008), and the Correct Information Unit in conversation (CIU_{conv}; Leaman & Edmonds, 2019) have mostly not found their way into everyday clinical practice, again likely due to the time constraints of transcribing and analyzing aphasic discourse.

Ramsberger and Rende (2002) introduced an innovative workaround to these time and labor constraints. Their measure of transactional success in conversation was shown to have construct validity, test-retest stability, and interrater reliability. In their study, people with aphasia watched *I Love Lucy* episodes and then worked with unfamiliar conversation partners who were naïve to the episodes, in order to "co-construct" the stories (Goodwin, 1995) and reach intersubjectivity, i.e., a shared interpretation of the main ideas. Scoring the partner's narrative retell as a measure of transactional success in conversation in aphasia provided evidence of content-related validity, i.e., that the test content is relevant to the proposed use of the test (Messick, 1995). Drawbacks to their measure of transactional success that may have limited its clinical implementation mostly centered around time: stimuli were too long (~25 minutes) and conversations between dyads were unlimited in time.

Ramsberger and Rende's idea of measuring transactional success in storytelling via the conversation partner's retelling was recently applied to assess the outcome of a conversation-based therapy in four persons with aphasia (Carragher et al., 2015). Some

limitations of their study included a small sample size, reliance on evidence of validity and reliability from prior studies, and a choice of pre/post-treatment untrained stimuli that may have given an advantage to the outcome measure. Both pre- and post-treatment stimuli were short "Mr. Bean" video clips. It is reasonable to suspect that the dyadic partners, having spent time and effort to establish joint reference regarding the genre and main actor would require fewer words and turns on a subsequent attempt (Clark & Wilkes-Gibbs, 1986).

The Brief Assessment of Transactional Success in conversation in aphasia (BATS) is currently in development to provide a valid, reliable, practical tool for measuring change in conversation outcomes that could be included in a D-COS. Inspired by Ramsberger and Rende (2002), one goal of developing the BATS is to allow clinicians and clinical researchers to measure the number of main concepts and correct information exchanged between persons with aphasia and non-aphasic conversation partners, without the labor-intensive requirement to transcribe and analyze aphasic discourse. In their study, checklists of main ideas were proposed by the authors and a research assistant, and inarguably reflect what a complete and accurate story retelling *should* include. This method of developing lists of propositions is common in discourse analysis in aphasia research, but has the disadvantage of creating target lists that may be biased due to sampling from a small, selective, homogeneous group (e.g., regarding age, years of education, race/ethnicity, gender). An additional concern is how completely and accurately non-aphasic subjects retell the stories that we ask persons with aphasia to retell, an area that is still under rigorous examination.

To address the issue of potentially biased, predetermined Richardson and Dalton (2016); Richardson & Dalton (2020) utilized a sizeable sample of non-aphasic participants available through AphasiaBank (MacWhinney et al., 2011), a database that includes a handful of monologic discourse tasks (i.e., picture descriptions, procedural discourse, and retelling of a wordless picture book) elicited from both aphasic and non-aphasic participants. They developed lists of main concepts and preliminary normative references for these standardized discourse tasks, and investigated differences between proposition generation in four age groups (n = 23 each). Comparing the older half to the younger half of participants, Richardson and Dalton (2016) found that younger participants (59 and younger) produced more accurate and complete concepts, on average, on two of the three discourse tasks (Cinderella and PB&J).

In a similar vein, in this first phase of test development of the BATS, we investigated non-aphasic performance in story retelling in three age groups, and developed checklists of main concepts (MCs). Like Richardson and Dalton (2016) and Richardson & Dalton (2020), we define an MC as an utterance that contains a subject, one main verb, and an optional object, with or without subordinate clauses. We used a comprehensiveness rating that scores accuracy and completeness of story elements, similar to the methods of Richardson and Dalton and others (Nicholas & Brookshire, 1995). MCs reflect a proposition-level of discourse analysis. They characterize a person's ability to recall and express story gist at a macrolinguistic level. The impetus to develop checklists of MCs in this first phase of testing was to have them for analysis of later phases of data collection which involve persons with aphasia. Checklists are increasingly viewed as a "clinician-friendly" method of bridging the gap between ecologically valid measurement of language at the level of discourse which is often time- and resource-prohibitive and clinical usability (Kim et al., 2021, 2019).

In line with prior studies of monologic spoken discourse (e.g., Richardson & Dalton, 2016, 2020), we expected our non-aphasic group to demonstrate variability in what they included in their story retellings. Given prior research suggesting age-related declines in narrative production that might be related to memory and attention (Wright et al., 2011), we hypothesized that accuracy and completeness of MCs in younger participants would be significantly greater than in older participants, but not necessarily for all stimulus types, similar to the findings by Richardson and Dalton (2016). Fergadiotis and colleagues also found age-related differences – in lexical diversity – with older participants producing a greater range of vocabulary, but only when stimuli were verbally presented. Measures of lexical diversity in both groups followed a similar pattern of production according to the stimulus type (Fergadiotis et al., 2011). The BATS stimuli vary in the degree to which story gist is dependent on auditory comprehension, and the degree of visual support accompanying any story narration. Given age-related differences in micro- and macrolinguistic measures previously found to vary across stimulus types, we expected that the BATS stimulus type might reflect differences associated with participants' age.

Materials and methods

Participants

Ninety-six volunteers (73 females; mean age = 46.04; *SD* = 19.06; range = 18–76) were recruited from the University of Massachusetts Amherst and surrounding communities via flyer and word of mouth. The Institutional Review Board of the university approved the study, and signed informed consent was obtained. Volunteers were recruited in three age bins (n = 32 each), and the sample included young adult (YA; 26 females; age range = 18–35), middle-aged (MA; 22 females; age range = 36–58), and older adults (OA; 25 females; age range = 60–76). The sample is predominantly self-identified as Caucasian (n = 85), and included African American (n = 2), Asian American (n = 8), and Hispanic/Latino (n = 1). Inclusionary criteria included healthy adults with normal or corrected vision and hearing, English as their primary language, with no history of neurological conditions, and who were willing to be videotaped retelling stories. Prior to the data collection session, participants were pre-screened for issues that might affect their performance on the task, notably: 1) whether they had ever sustained a brain injury or been diagnosed with a brain disease; 2) whether they had ever been diagnosed with a psychiatric disease; 3) whether they wore glasses when using a computer; and 4) whether they wore hearing aids. The first two questions were exclusionary criteria and no participants in this study responded positively to them. Participants who responded positively to wearing glasses (n = 60; YA = 7; MA = 23; OA = 30) and those who responded positively to wearing hearing aids (n = 3; MA = 1; OA = 2) were told to bring them on the day of their participation in the study. All participants complied with this request. The Mini-Mental State Exam (MMSE; Folstein et al., 1975) was used as a cognitive screen and administered on the same day as the data collection session. Scores were all within normal limits (range: 24–30) and no-one was excluded. Demographic data are reported in Table 1.

Table 1. Demographic information for the normative sample of 96 non-aphasic adults.

Age group	*N*	Age (years)	Gender	Education (years)	Race/ethnicity	MMSE scores
ALL (18–76)	96	46.04 *(19.06)*	73 F 23 M	17.08 *(3.15)*	85 Caucasian 2 African American 8 Asian American 1 Hispanic/Latino	28.84 *(1.98)*
YA (18–35)	32	23.63 *(4.63)*	26 F 6 M	15.81 *(2.01)*	31 Caucasian 1 Hispanic/Latino	28.94 *(2.41)*
MA (36–58)	32	46.59 *(7.08)*	22 F 10 M	18.35 *(3.76)*	27 Caucasian 1 African American 4 Asian American	29.03 *(1.86)*
OA (60–76)	32	67.91 *(5.41)*	25 F 7 M	16.5 *(4.09)*	31 Caucasian 1 Hispanic/Latino	28.88 *(1.64)*

Notes: YA = young adults; MA = middle-aged adults; OA = older adults; MMSE = Mini-Mental State Exam (Folstein et al., 1975); F = female; M = male; Age, Education, and MMSE scores are mean (*sd*)

Stimuli

A library of 16 short video/audio clips (mean = 2.55 minutes; SD = 0.50 minutes) was created. Stimulus brevity was intended to alleviate one of the major obstacles to implementing a clinically feasible test of transactional success in conversation, i.e., time. We aimed for a variety of stimuli, some of which mirror procedural discourse tasks, but many of which would evoke some emotional responses, including humor, sadness, awe, and inspiration. Four sets of four video/audio clips per set include "how to" videos, news clips, and other short stories that vary in several ways, including modality of delivery, novelty, abstraction, and emotional content. With regard to modality, the experimental variable of interest, the stimuli vary along a continuum of dependency on verbal comprehension for a complete appreciation of the story. Four non-verbal (NV) video clips can be understood without any spoken or written word comprehension, including two from a Chaplin silent film and two others that tell stories with no verbal information. Four "how to" video clips include approximately equivalent and synchronized visual and verbal (VV) information, and thus could be mostly understood even with compromised auditory comprehension. Four video biographical stories from the "Brief But Spectacular" PBS series rely heavily on verbal information with some visual support (VS). Finally, four audio clips from the NPR "StoryCorps" series are almost entirely reliant on speech comprehension, i.e., mainly speech-dependent (SD), with only a still photograph displayed during the entire audio story. Permission was obtained (from Lowes, PBS, NPR, a private foundation, and an individual videographer) to use 14 copyrighted video/audio clips, while two (Chaplin) were in the public domain. Descriptive data on the stimuli are reported in Table 2.

Discourse elicitation

Discourse samples were digitally recorded in a single session wherein participants each viewed and/or listened to eight video/audio clips, including two from each of the four stimulus types. Order of presentation utilized a custom randomization constraint, such that no two stimuli presented back-to-back were from the same condition. Both the flyer and the word-of-mouth script advertised that we were recruiting participants for a study of story retelling and that their responses to video and audio stimuli would inform our

Table 2. Descriptive information on video and audio stimuli.

#	Title	Condition	time (s)	Description	Source
1	Bicycle Boy	NV	119	Silent video about doing good	Maneesh Satheesan YouTube video
2	Chaplin Eat Shoe	NV	158	Silent video (Chaplin)	"The Gold Rush" movie
3	Share Care	NV	98	Silent video about doing good	Naik Foundation
4	Chaplin Shotgun	NV	157	Silent video (Chaplin)	"The Gold Rush" movie
5	Light Switch	VV	163	How to replace a light switch	Lowe's.com/Home101
6	Hang Blinds	VV	118	How to hang blinds	Lowe's.com/Home101
7	Curb Appeal	VV	150	How to improve your curb appeal	Lowe's.com/Home101
8	Fire Pit	VV	115	How to install a backyard fire pit	Lowe's.com/Home101
9	Marcus Yam	VS	198	Marcus Yam: photo journalist	PBS "Brief but Spectacular" series
10	Sylvia Earle	VS	181	Sylvia Earle: marine biologist	PBS "Brief but Spectacular" series
11	Naomi DeLaRosa	VS	194	Naomi DeLaRosa: on family	PBS "Brief but Spectacular" series
12	Robin Steinberg	VS	128	Robin Steinberg: The Bail Project	PBS "Brief but Spectacular" series
13	Ferguson	SD	178	Ferguson protesters find friendship	NPR "StoryCorps" series
14	Sept 11	SD	172	Sept 11: One survivor's story	NPR "StoryCorps" series
15	Aunt Mother	SD	166	Aunt turned mother after tragedy	NPR "StoryCorps" series
16	No Handbook	SD	156	Mother/son discuss school shootings	NPR "StoryCorps" series

Notes: NV = non-verbal ("silent") film clip; VV = visuo-verbal Do-It-Yourself video; VS = visually supported biographical video; SD = speech-dependent audio clip with only a single still photo for visual support

development of a test to measure conversational success in aphasia. Just prior to data collection, participants were instructed to position themselves "at a comfortable distance for viewing" the monitor of a 15-inch MacBook Pro. The volume was pre-set to a comfortable, but relatively loud volume. Four participants lowered the volume at the beginning of the first stimulus. Participants were instructed that they would be "watching, and/or listening to, eight short video or audio clips. After each one, you will turn and face this camera, and we will record you retelling what each clip was about, in as much detail as you can remember". Before beginning, they were asked if they had any questions. None did. At this point, they were reminded, if needed, that they had reported wearing glasses during computer work. Those six participants noted that they were wearing contact lenses. After each retelling, when they were ready, the testing administrator pressed the space bar to view and/or listen to the next clip until all eight were viewed and/or listened to and retold. Each video/audio clip was viewed and the story retold by 48 participants, 16 from each of the three age groups.

Transcripts

De-identified audio files were obtained from the videotaped story retells and the wav files uploaded for transcription to Rev.com©, a paid transcription service. This online platform provides fast transcription services using proprietary automatic speech recognition algorithms, artificial intelligence, and professional freelance transcribers. The fastest and least expensive service produces a machine-generated transcript that is guaranteed to provide at least 80% accuracy, whereas the human professional transcription service takes a little longer and is guaranteed to produce 99% accuracy. After testing one sample from each participant using the machine-generated transcription service, we determined that we could achieve 90–100% accuracy on a first pass transcription if 29% of the 22.6 hours of audio data were transcribed by humans, with the remaining 71% initially transcribed using the machine-generated program. The 223 (29% of a total 768) transcripts which

were completely transcribed by humans consisted of cases in which the first sample of machine-generated transcripts included at least one instance of unintelligible speech. In these instances, the entire sample was transcribed via the human professional service. Instances of unintelligible speech tended to occur mostly for older participants or those with regional dialects that apparently differed from what the machine learning program was trained on. All transcripts and audio and video files were reviewed by the third author and a trained research assistant for accuracy, regardless of whether the first transcript was generated by a machine or a human. Of the 768 discourse samples, 239 of them required 388 minor corrections, a very small fraction of the total words produced.

Main concept analysis

Main concept (MC) analysis followed methods first proposed by Nicholas and Brookshire (1993, 1995) and later developed by Richardson and Dalton (2016); Richardson & Dalton (2020) with minor exceptions. Relevant concepts for the video/audio clips were derived from the narrated scripts, except in the case of the non-verbal video clips wherein two research assistants created narrations and differences were resolved by the first and third authors. As previously defined, a relevant concept is any statement that is relevant to the story and consists of one main verb, and its subject, object, modifiers, and subordinate clauses if appropriate (Richardson & Dalton, 2016). Relevant concepts (RCs) are candidates for main concepts, i.e., they form the pool of all concepts from which a final list of main concepts is established. RCs were coded for presence, accuracy, and completeness in accordance with the five codes proposed by Richardson and Dalton (2016); Richardson & Dalton (2020): 1) absent (AB): the relevant concept is not produced; 2) accurate and complete (AC): all "essential elements", i.e., information essential to the story gist are present and accurate; 3) accurate but incomplete (AI): at least one essential element is missing; 4) inaccurate but complete (IC): all essential elements are produced but at least one is inaccurate; and 5) inaccurate and incomplete (II): at least one essential element is missing and at least one is inaccurate. Relevant concept coding was performed by a research assistant and the third author. Inter-rater reliability was conducted on all RCs for all transcripts. Intra-rater reliability was conducted on 20% of transcripts. Disagreements were resolved by the first and third authors. Inter- and intra-rater reliability is reported in Table 3 for all transcripts and by age group.

Similar to Richardson and Dalton (2016); Richardson & Dalton (2020), a 33% threshold was applied such that any relevant concept that was produced accurately and completely by 16 or more of the non-aphasic story re-tellers was deemed to be a main concept (MC). Thus, by definition, MCs are also statements that are relevant to the story, and consist of one main verb and its subject, object, modifiers, and subordinate clauses if appropriate. Like Richardson and Dalton (2016), we visually inspected frequency plots for natural gaps between MC productions. Although 33% may appear to be a low cut-off for MCs, it corresponded to natural gaps across the stimuli. Only 1% of MCs (2/158 total) just met the 33% minimum threshold. Moreover, only 18% of MCs (29/158) were under the 50% threshold and they contain concepts that seem fundamental to the narratives. These thresholds are reported along with MCs that survived a 66%, and 75% threshold. The additional thresholds provide useful information about the distribution of the MCs and will support future investigations using the BATS stimuli to perform MC analysis. Using the same (33%)

Table 3. Concept-by concept inter- and intra-rater reliability of MC coding by age group.

# Title% INTER-	% INTER-	% INTER-	% INTER-	% INTRA-	% INTRA-	% INTRA-	% INTRA-
	(ALL)	(YA)	(MA)	(OA)	(ALL)	(YA)	(MA)	(OA)
1 Bicycle Boy	98.1	99.0	97.9	97.4	98.3	98.4	97.9	98.4
2 Chaplin Eat Shoe	93.8	93.8	93.8	93.8	96.0	94.3	96.9	96.9
3 Share Care	93.2	92.6	94.3	92.6	94.7	96.0	94.3	93.8
4 Chaplin Shotgun	95.1	92.1	97.1	96.3	96.8	95.0	97.1	98.3
5 Light Switch	95.1	94.4	97.9	93.1	95.6	96.5	95.1	95.1
6 Hang Blinds	96.4	94.6	98.2	96.4	95.8	97.3	95.5	95.0
7 Curb Appeal	94.0	95.8	95.8	90.3	97.0	96.5	97.9	96.5
8 Fire Pit	95.6	97.2	93.2	96.6	94.0	95.5	94.3	92.6
9 Marcus Yam	95.8	94.3	94.9	98.3	95.5	94.9	95.5	96.0
10 Sylvia Earle	87.2	86.7	88.3	86.7	91.4	92.2	92.2	89.8
11 Naomi DeLaRosa	94.3	93.8	95.3	93.8	94.0	97.7	93.8	90.6
12 Robin Steinberg	90.2	92.0	89.3	89.3	97.9	100.0	98.2	95.5
13 Ferguson	91.3	93.8	93.1	86.9	92.7	94.4	88.8	95.0
14 Sept 11	95.7	96.4	96.4	94.3	95.5	96.4	94.8	95.3
15 Aunt Mother	94.0	96.4	96.4	89.3	94.6	96.4	92.0	95.5
16 No Handbook	95.1	96.0	95.5	93.8	95.8	94.9	94.9	97.7

Notes: Inter- and intra-rater reliability was conducted on 100% and 20% of transcripts, respectively.

threshold, elements of MCs were deemed essential or non-essential. For example, in the "Ferguson" story, the MC "The picture won a Pulitzer prize" consists of three essential elements ("the picture", "won", and "a prize") and one non-essential element ("Pulitzer") which was mentioned by fewer than 33% of respondents in the sample. Non-essential elements, along with alternative productions for all elements, are included in checklists to assist researchers and clinicians in scoring MCs (Appendix 1 and Supplemental Materials).

Finally, numerical scores were applied in order to calculate a composite MC score for each speaker per story retelling transcript as proposed in Richardson and Dalton (2016); Richardson & Dalton (2020), where the composite MC score = (3 x AC) + (2 x AI) + (2 x IC) + (1 x II). As noted by these authors, non-aphasic control participants were expected to receive mostly scores of AC or AB, however the multilevel scoring was included as a reference for comparison with future clinical populations.

Data analysis

R (R Core Team, 2017) was used for analyses. Characteristics of the RC and MC distributions are reported for all 16 stimuli, including descriptive statistics, skew, and kurtosis for the whole sample, for stimulus types (NV, VV, VS, SD), and for three age groups (18–35, 36–58, and 60–80). Histograms demonstrated non-normal distributions for all ratios (e.g., AC/MC, AB/MC, MCComposite/MC). Ratios were used to account for differences in number of MCs per stimulus. While these ratios were skewed, the distributions were not noticeably different among the four different stimulus types. Box plots were used to examine effects of age, gender, education, and MMSE scores on the ratio AC/MC for different stimulus types and different stimuli. We tested for significant differences among the scores (AC, AI, IC, II, AB) between the age groups, number of years of education, MMSE score, gender, stimulus, and stimulus types and adjusted for multiple comparisons using the false discovery rate (FDR). Since these counts do not follow a Normal distribution, the log

linear model was used to model rates without the Normal distribution assumption. The exception was the MCComposite response, which has a distribution closer to the normal distribution, for which we ran a linear mixed effects model.

Results

Relevant concepts

Descriptive statistics for relevant concepts produced for each stimulus include the following: mean, standard deviation, median, range, skewness, and kurtosis (Table 4). Mean and median values were close, suggesting a symmetric distribution of the data. Skewness and kurtosis were within acceptable ranges for assuming a normal distribution of the data, i.e., all values of skewness were < |2| and all values of kurtosis were < |4| (Keppel & Wickens, 2004; Kim, 2013).

Main concepts

Main concept checklists include all concepts produced by at least 33% of participants per stimulus. Main concepts produced by at least 50%, 66%, and 75% of the normative sample are indicated in the appendices. Descriptive statistics for main concepts produced for each stimulus include the following: mean, standard deviation, median, range, skewness, and kurtosis (Table 5). Mean and median values were close, and skewness and kurtosis were within acceptable ranges.

Normative references were obtained from the bottom 5% quantile of the ratio of MC composite score divided by number of MCs (MCComp/MCs) per stimulus, and by stimulus type, age group, and MMSE score. MCComp/MCs scores can range between 0 and 3. The distribution of MCComp/MCs from the normative sample describes the normal variation of MCComp/MCs. A score that falls below the bottom 5% quantile of this distribution has at most a 5% chance of coming from a neurotypical participant. Thus, scores falling below the bottom 5% quantile threshold are statistically low and may be considered "non-

Table 4. Descriptive statistics for relevant concepts (RCs) produced during retelling of 16 stories.

#	Title	Mean	*sd*	Median	Range	Skewness	Kurtosis
1	Bicycle Boy	6.94	2.11	7.5	1 to 11	−0.569	0.385
2	Chaplin Eat Shoe	7.63	2.89	8	1 to 12	−0.704	−0.074
3	Share Care	7.15	2.70	7	1 to 12	−0.331	0.000
4	Chaplin Shotgun	10.23	3.90	11	1 to 17	−0.809	0.309
5	Light Switch	7.17	2.86	8	1 to 11	−0.755	−0.148
6	Hang Blinds	4.90	1.70	5	0 to 8	−0.477	0.342
7	Curb Appeal	6.77	2.17	7	1 to 9	−1.448	1.652
8	Fire Pit	6.46	3.17	7	1 to 12	−0.282	−0.750
9	Marcus Yam	8.63	3.21	10	1 to 12	−0.854	−0.457
10	Sylvia Earle	6.73	2.16	7	0 to 10	−0.862	1.004
11	Naomi DeLaRosa	6.42	1.71	6.5	2 to 9	−0.503	−0.129
12	Robin Steinberg	6.15	1.64	6	3 to 9	−0.032	−0.757
13	Ferguson	7.33	3.44	7.5	0 to 14	−0.214	−0.872
14	Sept 11	8.75	2.87	9	0 to 13	−0.832	0.541
15	Aunt Mother	6.13	1.57	6	0 to 8	−1.499	3.870
16	No Handbook	8.21	2.35	9	2 to 12	−1.009	0.671

Table 5. Descriptive statistics for main concepts (MCs) produced during retelling of 16 stories.

#	Title	Mean	*sd*	Median	Range	Skewness	Kurtosis
1	Bicycle Boy	6.63	1.93	7	1 to 10	−0.605	0.580
2	Chaplin Eat Shoe	7.85	2.86	8	1 to 12	−0.794	0.004
3	Share Care	7.19	2.35	7	1 to 11	−0.759	0.648
4	Chaplin Shotgun	10.56	3.77	11.5	1 to 15	−1.119	0.819
5	Light Switch	5.94	2.50	6.5	0 to 9	−0.834	−0.110
6	Hang Blinds	5.15	1.74	5	0 to 7	−0.969	0.631
7	Curb Appeal	6.48	2.10	7	1 to 9	−1.630	2.077
8	Fire Pit	7.10	2.98	7.5	1 to 11	−0.671	−0.430
9	Marcus Yam	7.67	2.86	9	1 to 11	−0.798	−0.596
10	Sylvia Earle	5.50	1.89	6	0 to 8	−0.946	0.721
11	Naomi DeLaRosa	5.42	1.71	5.5	1 to 8	−0.503	−0.129
12	Robin Steinberg	5.10	1.36	5	2 to 7	−0.357	−0.702
13	Ferguson	5.83	2.50	6	0 to 10	−0.411	−0.617
14	Sept 11	8.21	2.67	9	0 to 12	−0.874	0.658
15	Aunt Mother	5.54	1.49	6	0 to 7	−1.584	3.242
16	No Handbook	7.96	2.32	8	2 to 11	−1.167	0.722

Table 6. Bottom 5% quantile for main concept (MC) composite score/MCs ratio by stimulus type and stimulus.

#	Stimulus Type	5% threshold by stimulus type	Stimulus (Title)	5% threshold by stimulus
1	NV	0.81	Bicycle Boy	1.2
2			Chaplin Eat Shoe	0.75
3			Share Care	1.1
4			Chaplin Shotgun	0.43
5	VV	0.55	Light Switch	0.63
6			Hang Blinds	1.01
7			Curb Appeal	0.79
8			Fire Pit	0.49
9	VS	1	Marcus Yam	0.76
10			Sylvia Earle	1.09
11			Naomi DeLaRosa	1.12
12			Robin Steinberg	1.29
13	SD	0.96	Ferguson	0.84
14			Sept 11	1.06
15			Aunt Mother	1.29
16			No Handbook	0.88

Notes: The ratio of MC composite score/MCs is a number between 0 and 3. Numbers falling below the bottom 5% quantile thresholds can be considered "non-normal"; NV = non-verbal ("silent") film clip; VV = visuo-verbal Do-It-Yourself videos; VS = visually supported biographical video; SD = speech-dependent audio clip with only a single still photo for visual support

normal". The bottom 5% quantile for stimulus and stimulus type are shown in Table 6. The bottom 5% quantile for age group and MMSE score are shown in Table 7. The threshold decreases as MMSE scores decrease and as age increases.

While there was no correlation between age and MMSE score, it is noteworthy that the bottom 5% threshold decreases as age increases and as MMSE scores decrease. A score of 23 and lower on the MMSE is considered to be indicative of cognitive impairment, and none of our non-aphasic sample scored at or below 23. There was, however, a significant decline in the bottom 5% threshold distinguishing the majority of participants (88/96) who scored 26–30 from those who scored 24–25 on the MMSE (Table 7).

The data were examined for patterns in non-normative low and high scores by participants, age groups, stimuli, and stimulus types. On the low scoring end, i.e., scores falling below the 5% quantile as a normative reference threshold, a total of 24 participants

Table 7. Bottom 5% quantile for the main concept (MC) composite.

score/MCs (MCComp/MCs) ratio by age group and MMSE

Variable	Level	Number	5% threshold by stimulus
Age groupYA321.24			
	MA	32	0.65
	OA	32	0.61
MMSE score	24	4	0.52
	25	4	0.31
	26	6	0.84
	27	3	0.78
	28	5	0.86
	29	14	1.07
	30	60	1.1

Notes: The MCComp/MCs ratio is a number between 0 and 3. Numbers falling below the bottom 5% quantile thresholds can be considered "non-normal"; YA = young adult; MA = middle-aged adult; OA = older adult; MMSE = Mini-mental state exam (Folstein et al., 1975)

(25%) produced 50 narratives (6.5%) that were scored at or below threshold. Whereas two-thirds of these participants (n = 16) produced only 1–2 narratives with low MCComp/MCs ratio scores, three of these participants (NC-58, NC-66, and NC-67) received low scores on 5 or more of their 8 narrative retells, accounting for over one third of the narratives in the "non-normal" sample. NC-58 is a 52-year-old high school graduate with 15 years of education who is employed as a school janitor. He scored 30 on the MMSE. NC-66 is a 75-year-old retired professional with a masters degree and 18 years of education. She scored 25 on the MMSE. NC-67 is a retired painter with a bachelors degree and 16 years of education. She scored 30 on the MMSE.

Most stimuli elicited low scoring narratives from only 2 or 3 of 48 participants retelling them. Two stimuli ("Robin Steinberg" and "Aunt Mother") elicited narratives that accounted for 20% of the low scores. On the high scoring end, many narratives received "perfect" scores, i.e., those in which all essential elements were scored accurate and complete, thus producing a MCComp/MCs ratio of 3. Forty-four participants (45.8%) produced 84 narratives (10.9%) that received perfect scores. Again, most participants produced only 1–2 narratives with perfect scores; however, 13 participants produced three or more narratives that received perfect scores. Of note, there was very little overlap between these two groups of low and high scoring participants (n = 65 total), with only three participants producing one perfect narrative and 1–3 low scoring narratives. While all stimuli elicited two or more narratives with perfect scores, two stimuli ("Hang Blinds" and "Aunt Mother") elicited narratives that accounted for 31% of the perfect scores. Patterns observed between age groups and stimulus types are discussed below.

Influence of age group

Descriptive statistics for the dominant main concept codes (AC and AB) and main concept composite scores are reported in the context of age group in Table 8. Descriptive statistics for the codes indicating partial accuracy and/or completeness (AI, IC, and II) are not listed given their relative infrequency in this sample. Mean and median values were close, and skewness and kurtosis were within acceptable ranges.

Table 8. Descriptive statistics for dominant MC codes for all participants and by age group.

		ALL	YA (18–35)	MA (36–59)	OA (60–80)
AC	Mean	6.76	7.16	6.59	6.52
	sd	2.76	2.50	2.96	2.78
	Median	7	7	7	7
	Range	0 to 15	1 to 15	0 to 15	0 to 13
	Skewness	0.07	0.46	0.15	−0.21
	Kurtosis	0.18	0.68	0.04	−0.27
AB	Mean	2.57	2.25	2.73	2.71
	sd	2.32	2.01	2.57	2.32
	Median	2	2	2	2
	Range	0 to 14	0 to 9	0 to 13	0 to 14
	Skewness	1.45	1.17	1.37	1.60
	Kurtosis	2.67	1.14	2.12	3.51
MCComp	Mean	21.35	22.38	20.86	20.82
	sd	7.94	7.26	8.46	7.99
	Median	21	21	21	21
	Range	0 to 45	5 to 45	2 to 45	0 to 39
	Skewness	0.07	0.54	0.17	−0.29
	Kurtosis	0.24	0.76	0.05	−0.17

Notes: AC = code for all essential elements of main concepts (MCs) scoreD accurate and complete; AB = code for MC was missing from narrative retell; MCComp = MC composite score; YA = young adults; MA = middle-aged adults; OA = older adults

Results of the log linear model examining odds of different scores (AC/MCs, AI/MCs, IC/MCs, II/MCs, and AB/MCs) according to age group demonstrated that young adults had much higher counts of AC relative to AB than did middle-aged adults (β = 0.34, *SE* = .07, p_{FDR} < .001). This result converges with results of t-tests comparing group means on MCComp/MCs that demonstrate significant differences between the young and middle-aged groups (p= 0.007) and the young and older groups (p= 0.002; see Figure 1). It also converges with results of the linear mixed model using MCComp alone as the response, given that its distribution was closer to a normal distribution (Table 9). In this model, the intercept represents the average MCComp for the middle-aged group, female gender, stimulus type non-verbal, with average number of years of education and average MMSE score. According to this model, there was an effect of age group such that the young adult group scored on average higher on MCComp than the middle-aged group (β = 1.85,

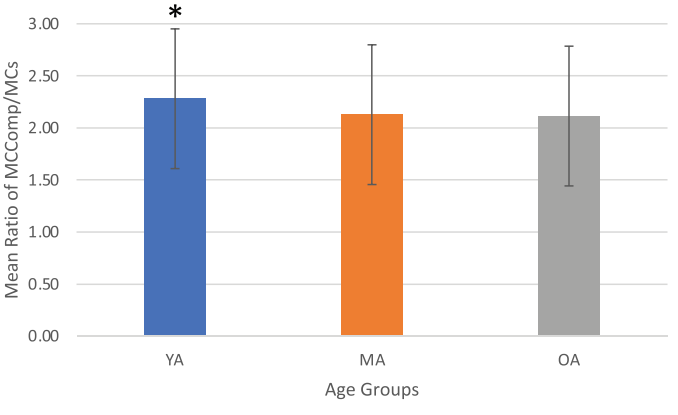

Figure 1. Mean ratio of main concept (MC) composite score to number of MCs by age group.

Table 9. Significant fixed effects in the linear mixed effects model estimating MCComp.

Fixed Effect	β	*SE*	*p*
(Intercept)	24.54	0.69	< .001
Age group, YA	1.85	0.7	0.01
MMSE score	0.66	0.16	< .001
Stimulus type, VV	−5.32	0.76	< .001
Stimulus type, VS	−6.54	0.76	< .001
Stimulus type, SD	−3.53	0.76	< .001

Notes: The intercept represents the overall mean main concept composite score (MCComp), and the coefficient of a variable represents the effect of the variable as the amount of deviation from the overall mean. YA = young adult group; MMSE = Mini-mental State Examination (Folstein et al., 1975); VV = visuo-verbal stimuli; VS = visually supported stimuli; SD = speech-dependent stimuli

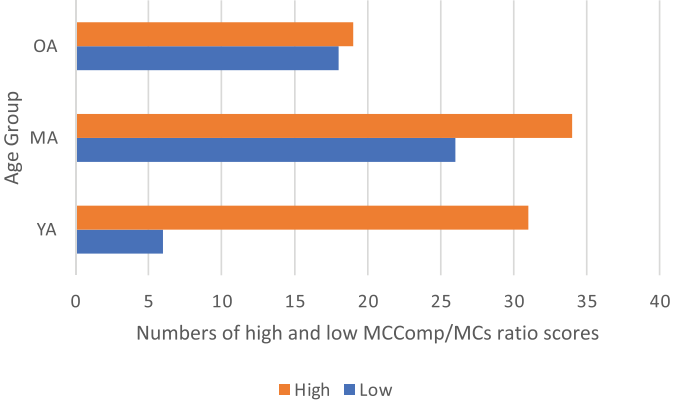

Figure 2. Numbers of lowest (bottom 5% quantile) and highest (perfect) main concept (MC) composite to number of MCs ratio scores by age group.

SE = 0.70, *p* = 0.01). We warn that MCComp by itself is affected by the number of MCs per stimulus, which depends both on the richness of relevant concepts in a stimulus and the capabilities of story retellers.

Age group influenced the number of "non-normal" low range scores (bottom 5% quantile) with respect to the ratio of MCComp/MCs. The young adult group had fewer than older and middle-aged adults (n = 6; 26; 18, respectively). As noted above, the threshold decreases as age increases, especially between the young adult group and the other two age groups (Table 7). On the high end, i.e., "perfect" scores, the middle-aged group had more than young and older adults (n = 34; 31, and 19, respectively). Lowest and highest scores by age group are shown in Figure 2.

Influence of stimulus type

Descriptive statistics for the dominant MC codes (AC and AB) and MC composite scores are reported in the context of stimulus type in Table 10. Mean and median values were close, and skewness and kurtosis were within acceptable ranges.

Table 10. Descriptive statistics for dominant MC codes for all narratives and by stimulus type.

		ALL	NV (15)	VV (11)	VS (11)	SD (12)
AC	Mean	6.76	8.06	6.17	5.92	6.89
	sd	2.76	3.17	2.46	2.26	2.57
	Median	7	8	7	6	7
	Range	0 to 15	1 to 15	0 to 11	0 to 11	0 to 12
	Skewness	0.07	−0.04	−0.56	0.14	−0.35
	Kurtosis	0.18	−0.04	−0.01	−0.16	−0.25
AB	Mean	2.57	3.42	2.14	2.13	2.59
	sd	2.32	2.62	2.16	1.99	2.21
	Median	2	3	2	2	2
	Range	0 to 14	0 to 14	0 to 10	0 to 10	0 to 12
	Skewness	1.45	1.36	1.63	1.31	1.31
	Kurtosis	2.67	2.57	2.72	1.89	1.98
MCComp	Mean	21.35	25.20	19.89	18.66	21.67
	sd	7.94	9.15	6.99	6.41	7.43
	Median	21	24	21	18	21
	Range	0 to 45	3 to 45	2 to 33	2 to 33	0 to 36
	Skewness	0.07	−0.02	−0.62	0.11	−0.39
	Kurtosis	0.24	−0.04	0.20	−0.22	−0.22

Notes: NV = non-verbal; VV = visuo-verbal; VS = visually supported; SD = speech-dependent Maximum number of main concepts (MCs) per stimulus type is shown (in parentheses); AC = code for all essential MC elements scored accurate and complete; AB = code for MC was missing from narrative retell; MCComp = MC composite score

Results of the log linear model examining odds of different scores (AC/MCs, AI/MCs, IC/MCs, II/MCs, and AB/MCs) according to stimulus type demonstrated that visuo-verbal stimuli elicited higher counts of AC relative to AB than did non-verbal stimuli (β = 0.20, SE = 0.07, p_{FDR} = 0.02). The same trend occurred for the visually supported stimuli (β = 0.17, SE = 0.07, p_{FDR} = 0.07). The three other stimulus types elicited more MCs that were scored AI than AB as compared to the non-verbal stimuli (visuo-verbal: β = 1.23, SE = 0.18, p_{FDR} < .001; visually supported: β = 0.80, SE = 0.20, p_{FDR} < .001; and speech-dependent: β = 0.63, SE = 0.20, p_{FDR} < .005). While results of the linear mixed model using MComp as the response showed an effect of stimulus type for visuo-verbal, visually supported, and SD stimuli compared to non-verbal (Table 9), these results are moot, given the differences in number of MCs by stimulus type (non-verbal > speech-dependent > visuo-verbal > visually supported).

T-tests comparing group means on MCComp/MCs did not demonstrate significant differences between the non-verbal stimuli and the other three stimulus types (visuo-verbal: $t(381) = 1.97$, $p = 0.08$; visually supported: $t(381) = 1.97$, $p = 0.09$; speech-dependent: $t(381) = 1.97$, $p = 0.19$) (Figure 3).

Stimulus type influenced the number of "non-normal" low range scores. The non-verbal stimuli elicited fewer than visuo-verbal, visually supported, and speech-dependent (n = 11; 12; 13; and 14, respectively). As shown in Table 6, the bottom 5% threshold for non-normal MCComp/MCs scores is considerably lower for stimulus type visuo-verbal (0.55) than for non-verbal, visually supported, and speech-dependent stimuli (0.81, 1.0, and 0.96, respectively). On the high end of scores, the visuo-verbal stimuli elicited more "perfect" scores than visually supported, speech-dependent, and non-verbal (n = 28; 24; 21; and 11, respectively). Lowest and highest scores are shown in Figure 4.

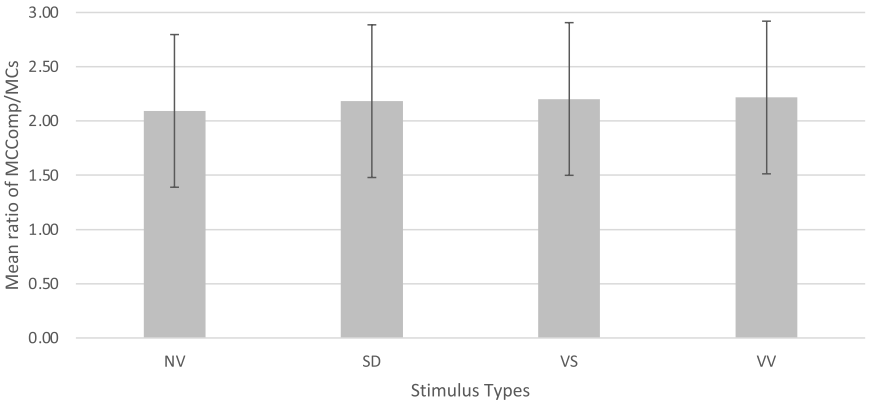

Figure 3. Mean ratio of main concept (MC) composite score to number of MCs by stimulus type.

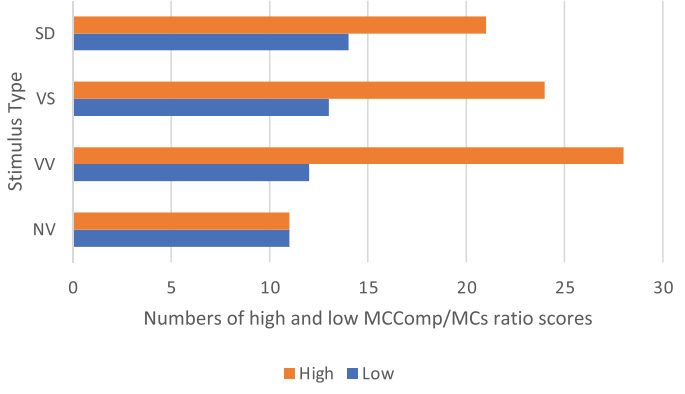

Figure 4. Numbers of lowest (bottom 5% quantile) and highest (perfect) main concept (MC) composite to number of MCs ratio scores by stimulus type.

Discussion

This study describes the acquisition and first phase of analysis of a set of story-retelling narratives for the purposes of: 1) developing normative references from a sample of non-aphasic volunteers; and 2) developing checklists of main concepts (MCs) that can be used in a new Brief Assessment of Transactional Success (BATS), a test to measure communicative success in conversation in aphasia. The notion of MCs and the method of developing them closely follows that described by Richardson and Dalton (2016); Richardson & Dalton (2020) for the five discourse tasks included in the AphasiaBank protocol (MacWhinney et al., 2011). The current study describes a new set of 16 stimuli included in the BATS, the non-aphasic sample that participated in this first phase of testing that will ultimately contribute to evidence of criterion validity in distinguishing between aphasic and non-aphasic performance. Importantly, it provides normative information regarding non-aphasic production of MCs, and checklists detailing essential

elements and alternative productions for each of the MCs. We also examine differences in performance between age groups within the sample and the influence of differences in stimuli and types of stimuli.

Story retelling variability

Many factors contribute to retelling an accurate, complete, complex, coherent story, from macrostructural elements such as coherence and cohesion, to microstructural elements that form much of the basis of the traditional structuralist approach to analyzing aphasic discourse (Armstrong, 2000). A variety of micro- and macrolinguistic measures have been studied in aphasic story retelling, e.g., number of words/min, utterances with and without mazes, mean length of utterance, CIUs/minute, story propositions, and percentage of accurate and complete independent and dependent clauses in the Story Retelling Procedure (Doyle et al., 1998), or number of main ideas (Ramsberger & Rende, 2002) or "salient content words" (Carragher et al., 2015) in measures of transactional success. Clearly, no one measure captures the inherent complexity of the uniquely human ability to create accurate, complete, coherent, rich, complex spoken narratives.

It is also clear that individual storytelling style played a role in the variability we observed in assessing the presence, accuracy, and completeness of main concepts. Even though we are reporting the results of a *non-aphasic* sample, MC analysis revealed issues that could be described as occurring at multiple levels of *aphasic* discourse processing as characterized, for instance, by the LUNA framework (Dipper et al., 2021). We observed, for example, vague or inaccurate word selection (linguistic), inclusion of too many or too few story elements that suggested poor conceptualization (propositional), poor story organization (macrostructure planning), and failure to follow the instructions (pragmatics). It is important to capture such variability in non-aphasic subjects' retelling of stories that are designed to elicit story retelling in persons with aphasia.

Although some of this variability can be expected given inherent individual differences in storytellers' style, one somewhat surprising result was the wide range of what a "normal" sample produces. We assumed that the instruction to "... retell what each clip was about, in as much detail as you can remember" would focus participants' attention on visual and/or auditory details of each clip. For most participants, the ratio of their MC composite scores to the number of MCs (MCComp/MCs) across stimuli (mean = 2.17; *SD = 0.66*) suggests that the task of producing a reasonably coherent, accurate, and complete narrative was a relatively easy task for non-aphasic persons. And yet we observed a surprising range of responses, both on the high and low ends of what comprises accuracy and completeness in conveying MCs. For example, one participant's narrative in response to the "September 11" story was entirely her own experience of the event and lacked any MCs. Although this participant (NC-67) did not personalize other stories to the same extent, she was an outlier scoring in the bottom 5% on 7 of 8 narratives. The fact that there was so little overlap between participants who produced multiple narratives that were scored very high or below threshold suggests that some characterization of individual story retelling style can be at least partially captured by MC analysis. Other intrinsic within-participant variables may have also contributed to accuracy and completeness of MCs, although the sample may not have been diverse or large enough to capture some of this variability, a study limitation that is addressed below.

Notably for future clinical studies, individual storytelling style may be a premorbid trait that we have no access to, and thus ignore completely when scoring aphasic discourse. It would be useful in the future to examine how participants rate themselves as storytellers.

Influence of types of stimuli

The experimental design of stimuli was based on an assumption that the four types of video/audio stimuli were situated along a continuum of the degree to which auditory comprehension would aid in following, retaining, and recalling the gist of the narratives. We thus expected the stimuli to facilitate comprehension of story gist in this same order of visual-to-auditory support, i.e., non-verbal > visuo-verbal > visually supported > speech-dependent for people with aphasia. This was purposeful in the design, given that the ultimate target audience is individuals with aphasia for whom visual information processing is often less challenging than auditory processing. As testing development continues, and a pool of people with aphasia of varying type and severity are tested, we expect that the range of visual and verbal support in BATS stimuli will capture sample variability with regard to individuals' skills in comprehending and integrating verbal and visual information. In the current normative study, we observed some differences on the basis of the stimulus types, but they did not conform to our expectations with regard to the experimental design of stimuli as described for people with aphasia. For example, there were higher counts of accurate and complete vs. absent MCs for the visuo-verbal and visually supported stimuli relative to the non-verbal stimuli. Nonetheless, there were no statistically significant differences in the mean ratios of MCComp/MCs across stimulus types.

One clear difference observed in our data based on stimulus types was in the bottom 5% threshold where visuo-verbal (0.55) < non-verbal (0.81) < speech-dependent (0.96) < visually supported (1.00). In this arena, visuo-verbal stimuli stand apart as having a very low threshold for "non-normal" low scores. Visuo-verbal stimuli are also distinctive from the other stimulus types in that the videos are "how to" instructional videos as opposed to "stories" per se. As such, the content may be less novel, certainly less abstract, and lacking in emotional valence compared to the other stimuli. We included them purposefully to have a range of different genres of discourse (Armstrong, 2000). The visuo-verbal stimuli and the narrative discourse they elicit fall somewhere between the typical procedural discourse task (e.g., "how do you make a peanut butter and jelly sandwich?") and a traditional story retelling task (e.g., retelling a wordless picture book story or the ubiquitous "tell me everything you see going on in this picture" tasks). Unlike the visually supported stimuli, narrated instruction in the visuo-verbal stimuli is temporally aligned with step-by-step visual instruction. And while all the stimuli elicit stories with a natural temporal order, most of the visuo-verbal stimuli include instructions which are strictly temporally ordered.

Influence of age

An abundant literature has developed over the last few decades to characterize "normal cognitive aging" including changes in sensory and motor function, attention and memory function, and the so-called executive functions (for extensive reviews, see for example,

Cabeza et al., 2005; Lezak et al., 2004). Much less is known about changes in the verbal abilities of older adults, although some studies have suggested that verbal abilities such as naming, auditory and written verbal comprehension, and spelling, when tested using a standardized language battery, remain *relatively* stable as a function of normal, healthy aging (Goulet et al., 1994; Schum & Sivan, 1997). Some of the evidence, which appears equivocal, may depend on the particular task, study design, age group comparisons, etc. With regard to naming, for example, cross-sectional studies have reported declines in naming ability on the Boston Naming Test (BNT; Kaplan et al., 1983; 2001) with age (MN Cruice et al., 2000). To control for the many possible confounding variables in cross-sectional designs, Cruice and colleagues ran a mixed longitudinal and cross-sectional study. They found mixed results, with no age-related declines in naming ability over a four-year period of the longitudinal design. In another mixed design with a larger sample (n = 236), BNT performance declined with age, but only by 2 percentage points per decade (Connor et al., 2004). In a large-scale longitudinal study (n = 541) of the effects of age over five decades on confrontation naming also using the BNT, lexical retrieval was found to be preserved with only subtle decline for individuals in their 70s and 80s (Zec et al., 2005). It has also been observed that naming in fluency tasks may be more susceptible than confrontation naming to age-related declines, although this may be an artefact of reduced processing speed (Huff, 1990).

The examination of other microlinguistic and macrolinguistic aspects of discourse production in normal aging has received less attention. Some studies have found no significant age differences in microlinguistic measures such as syntactic complexity (Nippold et al., 2013) or production of syntactic and lexical errors in connected speech between middle-aged and elderly healthy subjects (Glosser & Deser, 1992). Likewise, errors involving planning and self-monitoring were found to be common among younger and older adults (Perreira et al., 2019). One study found no age-related differences in quantifiable measures of speech production on picture description tasks, other than an increase in pauses that may reflect cognitive slowing (Cooper, 1990). Other investigations at the structural level of discourse have suggested some advantages of age, for example, in measures that tap vocabulary. In a comparison of four different types of discourse between two clearly different groups of young and old adults, the older group demonstrated significantly greater lexical diversity than the younger group, but only on tasks involving procedural discourse and personal recounts (Fergadiotis et al., 2011).

Unlike studies of age-related changes in discourse at microlinguistic levels, some evidence points to age-related declines in comprehension and/or production of discourse at macrolinguistic levels. A comparison of story comprehension and story proposition production in two cohorts of young and old adults who were also tested on measures of cognitive non-linguistic functions suggests that memory and attention contribute to declines in story processing performance in healthy aging (Wright et al., 2011). In the most direct comparison to the current study on the measure of main concept production, Richardson and Dalton found some age-related differences on some of the discourse tasks (Richardson & Dalton, 2016, 2020). In their review of studies on the influence of normal aging on global coherence, Ellis and colleagues suggested that findings of declines might be due to cognitive changes such as attention and processing speed (Ellis et al., 2016). This aligns with the hypothesis that macrolinguistic abilities, which depend on integration of linguistic and nonlinguistic cognitive functions, may be susceptible to age-related non-

linguistic cognitive declines, e.g., in processing speed or efficiency of information processing (Glosser & Deser, 1992). An important disclaimer is that older volunteers who may appear healthy and "normal", and who pass a cognitive screen, may nonetheless have early, subtle, undiagnosed symptoms of neurodegeneration (Howieson et al., 2004). For an extensive review of the influence of aging on comprehension and production in discourse, see for example, Shadden (1997).

Like earlier studies, we observed that the younger third of participants produced narrative retells that were scored significantly higher in the ratio of MCComp/MCs than the older two thirds of participants (middle-aged and older adults). Similarly, the younger group had significantly higher AC counts relative to AB counts than did the middle-aged group. In addition, the bottom 5% threshold for "non-normal" performance in terms of the ratio of MCComp/MCs was distinctively higher for young adults (1.24) than for middle-aged (0.65) and older (0.61) adults. These results are consistent with previous findings and confirm the importance of acquiring age-stratified norm-references when analyzing discourse.

Conclusions, limitations, and future directions

In this first phase of development of the BATS, we examined 768 narratives from a sample of 96 non-aphasic volunteers. The current study focuses on one macrolinguistic measure of discourse analysis, MCA, that gets at a person's ability to convey a story's gist. We generated checklists of MCs for each of the 16 stimuli that comprise the BATS, including essential elements of MCs that were produced by at least 33% of the normative sample, along with examples of alternative productions. We examined the range of performance in terms of accuracy and completeness of narrative productions for each MC for each stimulus, and developed normative reference thresholds reflecting "non-normal" scores falling below the 5% quantile. These findings were evaluated in the context of two variables known to influence narrative discourse production, age group, and stimulus type. In line with prior research, younger adults scored higher than middle-aged and older adults on the ratio of MC composite to number of MCs, and fewer of them scored at or below the "non-normal" threshold. This same ratio was less influenced by stimulus type, which did not affect production of MCs in the normative sample in the order in which we expect it to affect narratives produced by people with aphasia. Further analyses at both micro- and macrolinguistic levels of examining this first phase of discourse data are ongoing, including efforts to better understand the complex, multivariate ways in which participant and stimulus variables interact to produce a range of story retells.

We acquired and reported the following participant variables in this study: age, gender, years of education, race/ethnicity, and MMSE scores. The normative sample was gender-biased with about a 3:1 ratio of female:male volunteers. The sample was also top-heavy in self-identified Caucasian volunteers (88%) and in persons who had obtained a Bachelor's degree or higher (76%). Only three volunteers (3%) reported a high school diploma with 12 years of education. This skewed sample may account for the lack of a significant finding for the effect of years of education. Future studies examining the BATS stimuli for normative references should include more male, and a more racially/ethnically/socio-economically diverse sample of volunteers, although we are not aware of evidence suggesting that narratives would differ substantially from a more racially/ethnically

diverse group. On the contrary, one study of narratives elicited by pictorial stimuli in African American and Caucasian individuals with and without aphasia found no differences in thematic content between the two ethnic groups (Olness et al., 2002). On the other hand, robust gender differences have been observed, particularly in the context of autobiographical narratives where it has been shown that women produce longer, more detailed, more emotional, and coherent narratives than men, and that they are more likely to address others' thoughts and feelings in their narratives (Buckner & Fivush, 1998; as cited in Schulkind et al., 2012), or that women's narratives are longer, richer and more evaluative, while men's narratives contain more factual information (Schulkind et al., 2012). While these differences may be amplified in personal, autobiographical narratives, it would be worth exploring gender differences in a more balanced sample, especially given the nature of the BATS stimuli.

Discourse samples were acquired after a standard instruction to " . . . retell what each clip was about, in as much detail as you can remember". This instruction was purposeful, with the eventual clinical task in mind. We wanted one simple instruction across narrative retells, even though they differed in genre. We also hoped to acquire long enough discourse samples, given the recommendations of Brookshire and Nicholas (1994) that 300 to 400 words from aphasic speakers promote good test-retest stability, at least for the words per minute and percent CIUs measures they studied. Boyle (2014) also warned against the effect of small samples on session-to-session stability. One factor that we are still investigating in this normative sample is the organization of narrative retells. For example, we are in the process of examining order of MCs to enable a normative comparison of macrostructure planning and propositional components – what Hameister and Nickels (2018) call conceptualization – in future aphasic samples. Had we given the instruction to "retell the 'stories' with a beginning, middle, and end", though, we might not have observed the natural variability that is evident in the story grammar, including some disorganization in this non-aphasic sample. Nonetheless, it is an interesting question how the framing of instructions might influence the narrative retells, that is worthy of further investigation.

Other avenues for further investigation include a comparison of aphasic discourse samples elicited from the BATS audio/visual stimuli to those elicited from more traditional static prompts. With few exceptions (notably Carragher et al., 2015; Doyle et al., 2000, 1998; Ramsberger & Rende, 2002), most discourse elicitation prompts are static images used to elicit monologic discourse. Picture sequences have been observed to make it easier to elicit stories with some temporal-causal sequence. This may explain why they produce higher narrative levels and cohesive harmony than single pictures (Armstrong, 2000). As Doyle et al. (1998) noted, further investigation into development of stimuli that sample both discourse types and presentation modes was, and still is, warranted.

Many of the common scenes in single pictures and picture sequences reflect unusual, rather than everyday situations. The intention with selection of the BATS stimuli was to mimic real-life situations in which one attempts to retell an interesting story or describe how to do something. The stimuli do present a number of added potential challenges posed by the consequences of brain injury, including compromised visual and auditory processing, compromised auditory comprehension, and impaired memory. These tend to be less problematic with the presentation of a stationary visual stimulus, as occurs in

picture description tasks, although some of the commonly used pictures and picture sequences have their own issues, such as difficulty in perception of figure-ground (see for example, "The Painter" picture sequence; Appendix 2, Doyle et al., 1998). We also intended to provide a range of stimuli in terms of their dependence on visual and auditory stimulation for conveying story gist. In this way, the impact of impaired cognitive processing, such as auditory comprehension on story retelling in aphasia could be revealed. Once evidence is provided that the test is a valid instrument for assessing transactional success in aphasia, it will also provide clinicians and clinical researchers with choices of stimuli that can be calibrated to a client's severity level in auditory comprehension. We expect that some BATS stimuli may be more challenging than traditional discourse elicitation stimuli for many people with aphasia. However, we also anticipate that by including compelling stories and conversation partners in the co-construction of the narratives, the task will provide the kind of real-life motivation to convey information and communication support that can be missing in monologic story retells (Carragher et al., 2015).

Another line of inquiry involves what we have anecdotally observed as a propensity for some of the BATS stimuli to evoke strong emotional responses in the narrative retells. Some of the stimuli were purposefully selected to elicit emotional responses, including laughter but also sadness, awe, and inspiration. Evidence has shown a facilitatory advantage of language with emotional content on silent reading and writing (Landis et al., 1982), on auditory comprehension (Reuterskiold, 1991) and even on monosyllabic word repetition (Ramsberger, 1996). This may explain the preponderance of "chaotic" black and white images routinely used to elicit connected speech, e.g., the "Cookie Theft" scene in the Boston Diagnostic Aphasia Examination (BDAE; Kaplan et al., 1983; 2001), the "Grocery Store" scene in the Aphasia Diagnostic Profiles (ADP; Helm-Estabrooks, 1992), and the "Cat Rescue" picture (Nicholas & Brookshire, 1993).

The autobiographical stories that comprise the visually supported and speech-dependent BATS stimuli speak to contemporary issues in American life that we hypothesize are relatable, thought-provoking, and engaging in a way that is difficult to match in static black and white line drawings. Even the non-verbal stimuli evoked emotional responses that we hypothesize are less likely to occur from a similar proposition in commonly used static pictures or sequences of pictures. Although it seems obvious to us that many of the BATS stimuli are by design likely to evoke stronger emotional responses than traditional picture stimuli, this is an empirical question that requires further investigation.

Many of the BATS stories reflect topics that are current. Some are controversial, including stories of racial inequities in policing and the judicial system, immigration, protecting the oceans from pollution, and gun violence. The stories are told from the perspective of one or more persons whose life experience with these topics is presented within a relatable context of the human experience, including hope in the face of such challenging issues. Although there is always the danger of test stimuli becoming outdated, we expect the shelf life of these stories to last for a while. Although some stories may seem culturally specific to American culture, we hope that the universal nature of many of the "do good" stories, inspiring biographies, and stories of family relationships and global issues will enable the test to be used beyond the U.S. borders.

Finally, we report here only the first phase of testing the BATS stimuli on a non-clinical sample, and only in the context of one macrolinguistic measure of discourse analysis, that of main concepts. Other analyses are ongoing, including finer-grained examinations of particular stimuli that elicited outlier low scoring responses, with an eye toward their appropriateness for the clinical application of the BATS. One stimulus in particular, "Aunt Mother", elicited both the lowest and highest scoring narratives from different participants. From the speech-dependent NPR "StoryCorps" series, it is a compelling story that begins with a tragic murder and ends with a redemptive acknowledgement of family bonds strengthened by love. It is not clear why it would elicit this discrepant language behavior using MC analysis, but it also appears to be an outlier among stimuli in a natural language processing measure of topic similarity that is currently under investigation. As such, it may not be included in the eventual clinical application of the BATS.

The current study will inform Phase II of test development, i.e., the acquisition and analysis of aphasic discourse samples, temporarily suspended due to Covid-19. The ultimate intention in developing the BATS is to provide clinicians and clinical researchers with labor-saving tools for measuring a non-aphasic conversation partner's story retelling, thus circumventing the need for traditional, time-intensive analysis of aphasic discourse. This aspect of the BATS borrows from the work of Ramsberger and Rende (2002) and Carragher et al. (2015) who analyzed naïve conversation partners' story retells to assess transactional success in conversation in aphasia. Standard administration of the BATS will also create a rich database of: 1) attempts to retell stories by persons with aphasia; 2) topic-constrained conversations between persons with aphasia and familiar and unfamiliar non-aphasic conversation partners; and 3) story retellings by the conversation partners. These transcripts can, with the assistance of automatic speech recognition and natural language processing tools, be harvested to enable participation-based therapies to target and measure functional gains in aphasia therapy. These automated methods are increasingly accessible and feasible (Cho et al., 2021; Fromm et al., 2020, 2016; Le et al., 2018). Studies using these methods are currently underway and will be reported elsewhere. Finally, while the test was designed with the goal of simplifying aphasic discourse analysis and developing a more clinically feasible and accessible tool for evaluating improvements in functional discourse in aphasia, the BATS should be normed on other populations as well.

Acknowledgments

The authors wish to thank the following individuals and corporations for approving use of their video/audio stimuli: StoryCorps (produced by StoryCorps with funding provided by the Corporation for Public Broadcasting, as heard on NPR); PBS NewsHour "Brief But Spectacular" series (produced with support provided by Heising-Simons Foundation and Cambia Health Foundation, as seen on PBS); Lowes Home Improvement videos (as seen on YouTube); Naik Foundation (for "Share … Care … Joy … " seen on YouTube); and Maneesh Satheesan (for "the Bicycle Boy", A Zoola Dudes initiative seen on YouTube). We thank Jiayi Ruan for her help verifying and scoring transcripts. We are grateful to the reviewers for their thoughtful feedback and suggestions for improving the manuscript.

Disclosure statement

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

Funding

This work was supported by the University of Massachusetts Faculty Research Grant [P1FRG0000000251].

References

Armstrong, E. (2000). Aphasic discourse analysis: The story so far. *Aphasiology*, *14*(9), 875–892. https://doi.org/10.1080/02687030050127685

Armstrong, E., Ciccone, N., Godecke, E., & Kokj, B. (2011). Monologues and dialogues in aphasia: Some initial comparisons. *Aphasiology*, *25*(11), 1347–1371. https://doi.org/10.1080/02687038.2011.577204

Beeke, S., Beckley, F., Johnson, F., Heilemann, C., Edwards, S., Maxim, J., & Best, W. (2015). Conversation focused aphasia therapy: Investigating the adoption of strategies by people with agrammatism. *Aphasiology*, *29*(3), 355–377. https://doi.org/10.1080/02687038.2014.881459

Beeke, S., Wilkinson, R., & Maxim, J. (2003). Exploring aphasic grammar 2: Do language testing and conversation tell a similar story? *Clinical Linguistics & Phonetics*, *17*(2), 109–134. https://doi.org/10.1080/0269920031000061786

Boyle, M. (2014). Test-retest stability of word retrieval in aphasic discourse. *Journal of Speech, Language, and Hearing Research*, *57*(3), 966–978. https://doi.org/10.1044/2014_JSLHR-L-13-0171

Brady, M. C., Kelly, H., Godwin, J., Enderby, P., & Campbell, P. (2016). Speech and language therapy for aphasia following stroke. *Cochrane Database of Systematic Reviews*, *2016*(6). Article No. CD000425. https://doi.org/10.1002/14651858.CD000425.pub4

Brookshire, R. H., & Nicholas, L. E. (1994). Speech sample size and test-retest stability of connected speech measures for adults with aphasia. *Journal of Speech and Hearing Research*, *37*(2), 399–407. https://doi.org/10.1044/jshr.3702.399

Brumfitt, S. (1993). Losing your sense of self: What aphasia can do. *Aphasiology*, *7*(6), 569–575. https://doi.org/10.1080/02687039308248631

Bryant, L., Ferguson, A., & Spencer, E. (2016). Linguistic analysis of discourse in aphasia: A review of the literature. *Clinical Linguistics & Phonetics*, *30*(7), 489–518. https://doi.org/10.3109/02699206.2016.1145740

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. https://doi.org/10.1080/02687038.2016.1239013

Cabeza, R., Nyberg, L., & Park, D. (Eds.). (2005). *Cognitive neuroscience of aging: Linking cognitive and cerebral aging*. Oxford University Press.

Carragher, M., Sage, K., & Conroy, P. (2015). Preliminary analysis from a novel treatment targeting the exchange of new information within storytelling for people with nonfluent aphasia and their partners. *Aphasiology*, *29*(11), 1383–1408. https://doi.org/10.1080/02687038.2014.988110

Cho, S., Nevler, N., Shellikeri, S., Parjane, N., Irwin, D. J., Ryant, N., Ash, S., Cieri, C., Liberman, M., & Grossman, M. (2021). Lexical and acoustic characteristics of young and older healthy adults. *Journal of Speech, Language, and Hearing Research*, 1–13. https://doi.org/10.1044/2020_JSLHR-19-00384

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39. https://doi.org/10.1016/0010-0277(86)90010-7

Connor, L. T., Spiro, A., Obler, L. K., & Albert, M. L. (2004). Change in object naming ability during adulthood. *Journal of Gerontology*, *59*(5), PP203–209. https://doi.org/10.1093/geronb/59.5.P203

Cooper, P. V. (1990). Discourse production and normal aging: Performance on oral picture description tasks. *Journal of Gerontology*, *45*(5), PP210–214. https://doi.org/10.1093/geronj/45.5.P210

Cruice, M., Botting, N., Marshall, J., Boyle, M., Hersh, D., Pritchard, M., & Dipper, L. (2020). UK speech and language therapists' views and reported practices of discourse analysis in aphasia rehabilitation. *International Journal of Language & Communication Disorders*, *55*(3), 417–442. https://doi.org/10.1111/1460-6984.12528

Cruice, M. N., Worrall, L. E., & Hickson, L. M. H. (2000). Boston Naming Test results for healthy older Australians: A longitudinal and cross-sectional study. *Aphasiology*, *14*(2), 143–155. https://doi.org/10.1080/026870300401522

Davidson, B., Worrall, L., & Hickson, L. (2003). Identifying the communication activities of older people with aphasia: Evidence from naturalistic observation. *Aphasiology*, *17*(3), 243–264. https://doi.org/10.1080/729255457

DeDe, G., & Hoover, E. (2021). Measuring change at the discourse-level following conversation treatment: Examples from mild and severe aphasia. *Topics in Language Disorders*, *41*(1), 5–26. https://doi.org/10.1097/TLD.0000000000000243

DeDe, G., Hoover, E., & Maas, E. (2019). Two to tango or the more the merrier? A randomized controlled trial of the effects of group size in aphasia conversation treatment on standardized tests. *Journal of Speech, Language, and Hearing Research*, *62*(5), 1437–1451. https://doi.org/10.1044/2019_JSLHR-L-18-0404

Dietz, A., & Boyle, M. (2018). Discourse measurement in aphasia research: Have we reached the tipping point? *Aphasiology*, *32*(4), 459–464. https://doi.org/10.1080/02687038.2017.1398803

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. J., & Meteyard, L. (2020). Measures of functional, real-world communication for aphasia: A critical review. *Aphasiology*, *34*(4), 492–514. https://doi.org/10.1080/02687038.2019.1702848

Doyle, P. J., McNeil, M. R., Park, G., Goda, A., Rubenstein, E., Spencer, K., Carroll, B., Lustig, A., & Szwarc, L. (2000). Linguistic validation of four parallel forms of a story retelling procedure. *Aphasiology*, *14*(5/6), 537–549. https://doi.org/10.1080/026870300401306

Doyle, P. J., McNeil, M. R., Spencer, K., Goda, A., Cottrell, K., & Lustig, A. (1998). The effects of concurrent picture presentation on retelling of orally presented stories by adults with aphasia. *Aphasiology*, *12*(7–8), 561–574. https://doi.org/10.1080/02687039808249558

Ellis, C., Henderson, A., Wright, H. H., & Rogalski, Y. (2016). Global coherence during discourse production in adults: A review of the literature. *International Journal of Language &communication Disorders*, *51*(4), 359–367. https://doi.org/10.1111/1460-6984.12213

Elman, R. J. (2007). The importance of aphasia group treatment for rebuilding community and health. In L. LaPointe (Ed.), *Aphasia and related neurogenic language disorders* (3rd) (pp. 300–308). NY: Thieme Medical.

Elman, R. J., & Bernstein-Ellis, E. (1999). The efficacy of group communication treatment in adults with chronic aphasia. *Journal of Speech, Language, and Hearing Research*, *42*(2), 411–419. https://doi.org/10.1044/jslhr.4202.411

Fergadiotis, G., Kapantzoglou, M., Kintz, S., & Wright, H. H. (2019). Modeling confrontation naming and discourse informativeness using structural equation modeling. *Aphasiology*, *33*(5), 544–560. https://doi.org/10.1080/02687038.2018.1482404

Fergadiotis, G., Wright, H. H., & Capilouto, G. J. (2011). Productive vocabulary across discourse types. *Aphasiology*, *25*(10), 1261–1278. https://doi.org/10.1080/02687038.2011.606974

Ferro, J. M., Caeiro, L., & Santos, C. (2009). Poststroke emotional and behavior impairment: A narrative review. *Cerebrovascular Diseases*, *27*(1), 197–203. https://doi.org/10.1159/000200460

Fillmore, C. (1981). Pragmatics and the description of discourse. In P. Cole (Ed.), *Radical Pragmatics* (pp. 143–166). Academic Press.

Finch, E., Lethlean, J., Rose, T., Flemin, J., Theodoros, D., Cameron, A., Coleman, A., Copland, D., & McPhail, S. M. (2020). Conversations between people with aphasia and speech pathology students via telepractice: A phase II feasibility study. *International Journal of Language & Communication Disorders*, *55*(1), 43–58. https://doi.org/10.1111/1460-6984.12501

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189–198. https://doi.org/10.1016/0022-3956(75)90026-6

Fromm, D., Forbes, M., Holland, A., & MacWhinney, B. (2020). Using AphasiaBank for discourse assessment. *Seminars in Speech and Language*, *41*(1), 10–19. https://doi.org/10.1055/s-0039-3399499

Fromm, D., Greenhouse, J., Hou, K., Russell, G. A., Cai, X., Forbes, M., Holland, A., & MacWhinney, B. (2016). Automated proposition density analysis for discourse in aphasia. *Journal of Speech, Language, and Hearing Research*, *59*(5), 1123–1132. https://doi.org/10.1044/2016_JSLHR-L-15-0401

Glosser, G., & Deser, T. (1992). A comparison of changes in macrolinguistic and microlinguistic aspects of discourse production in normal aging. *Journal of Gerontology*, *47*(4), PP266–272. https://doi.org/10.1093/geronj/47.4.P266

Glueckauf, R. L., Blonder, L. X., Ecklund-Johnson, E., Maher, L., Crosson, B., & Gonzalez-Rothi, L. (2003). Functional outcome questionnaire for aphasia: Overview and preliminary psychometric evaluation. *NeuroRehabilitation*, *18*(4), 281–290. https://doi.org/10.3233/NRE-2003-18402

Goodwin, C. (1995). Co-constructing meaning in conversations with an aphasic man. *Research on Language and Social Interaction*, *28*(3), 233–260. https://doi.org/10.1207/s15327973rlsi2803_4

Goulet, P., Ska, B., & Kahn, J. H. (1994). Is there a decline in picture naming with advancing age? *Journal of Speech and Hearing Research*, *37*(3), 629–644. https://doi.org/10.1044/jshr.3703.629

Hameister, I., & Nickels, L. (2018). The cat in the tree – Using picture descriptions to inform our understanding of conceptualization in aphasia. *Language, Cognition and Neuroscience*, *33*(10), 1296–1314. https://doi.org/10.1080/23273798.2018.1497801

Helm-Estabrooks, N. (1992). *Aphasia diagnostic profiles*. PRO-ED.

Herbert, R., Best, W., Hickin, J., Howard, D., & Osborne, F. (2008). *POWERS: Profile of word errors and retrieval in speech*. J & R Press.

Howieson, D. B., Loring, D. W., & Hannay, H. J. (2004). Neurobehavioral variables and diagnostic issues. In M. D. Lezak, D. B. Howieson, & D. W. Loring (Eds.), *Neuropsychological assessment, 4th ed* (pp. 286–334). Oxford University Press.

Huff, F. J. (1990). Language in normal aging and age-related neurological diseases. In F. Boller, & J. Grafman (Eds.), *Handbook of Neuropsychology*, Vol. 4 (pp. 251-264). Amsterdam: Elsevier.

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

Kagan, A. (1995). Revealing the competence of aphasic adults through conversation: A challenge to health professionals. *Topics in Stroke Rehabilitation*, *2*(1), 15–28. https://doi.org/10.1080/10749357.1995.11754051

Kagan, A. (1998). Supported conversation for adults with aphasia: Methods and resources for training conversation partners. *Aphasiology*, *12*(9), 851–864. https://doi.org/10.1080/02687039808249580

Kagan, A., Simmons-Mackie, N., & Shumway, E. (2018). *A set of observational measures for rating support and participation in conversation between adults with aphasia and their conversation partners*. Aphasia Institute.

Kagan, A., Winckel, J., Black, S., Duchan, J., Simmons-Mackie, N., & Square, P. (2004). A set of observational measures for rating support and participation in conversation between adults with aphasia and their conversation partners. *Topics in Stroke Rehabilitation*, *11*(1), 67–83. https://doi.org/10.1310/CL3V-A94A-DE5C-CVBE

Kaplan, E., Goodglass, H., & Weintraub, S. (1983; 2001). *The Boston Naming Test*. Lea & Febiger.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). New York.
Kim, H., Kintz, S., & Wright, H. H. (2021). Development of a measure of function word use in narrative discourse: Core lexicon analysis in aphasia. *International Journal of Language & Communication Disorders*, *56*(1), 6–19. https://doi.org/10.1111/1460-6984.12567
Kim, H., Kintz, S., Zelnosky, K., & Wright, H. H. (2019). Measuring word retrieval in narrative discourse: Core lexicon in aphasia. *International Journal of Language & Communication Disorders*, *54*(1), 62–78. https://doi.org/10.1111/1460-6984.12432
Kim, H.-Y. (2013). Statistical notes for clinical researchers: Assessing normal distribution using skewness and kurtosis. *Restorative Dentistry & Endodontics*, *38*(1), 52–54. https://doi.org/10.5395/rde.2013.38.1.52
Klippi, A. (1996). Conversation as an achievement in aphasics. *Studia Fennica Linguistica*, *6*, 201–214. Helsinki: Finnish Literature Society.
Kurland, J., & Stokes, P. (2018). Let's talk real talk: An argument to include conversation in a D-COS for aphasia research with an acknowledgment of the challenges ahead. *Aphasiology*, *32*(4), 475–478. https://doi.org/10.1080/02687038.2017.1398808
Landis, T., Graves, R., & Goodglass, H. (1982). Aphasic reading and writing: Possible evidence for right hemisphere participation. *Cortex*, *18*(1), 105–122. https://doi.org/10.1016/S0010-9452(82)80022-1
Le, D., Licata, K., & Provost, E. M. (2018). Automatic quantitative analysis of spontaneous aphasic speech. *Speech Communication*, *100*, 1–12. https://doi.org/10.1016/j.specom.2018.04.001
Leaman, M., & Edmonds, L. (2019). Revisiting the correct information unit: Measuring informativeness in unstructured conversations in people with aphasia. *American Journal of Speech Language Pathology*, *28*(3), 1099–1114. https://doi.org/10.1044/2019_AJSLP-18-0268
Lezak, M. D., Howieson, D. B., & Loring, D. W. (2004). *Neuropsychological Assessment* (4th ed.). Oxford University Press.
Lomas, J., Pickard, L., Bester, S., Elbard, H., Finlayson, A., & Zoghaib, C. (1989). The communicative effectiveness index: Development and psychometric evaluation of a functional communication measure for adult aphasia. *Journal of Speech and Hearing Disorders*, *54*(1), 113–124. https://doi.org/10.1044/jshd.5401.113
MacWhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). AphasiaBank: Methods for studying discourse. *Aphasiology*, *25*(11), 1286–1307. https://doi.org/10.1080/02687038.2011.589893
McNeil, M. R., Doyle, P. J., Fossett, T. R. D., Park, G. H., & Goda, A. J. (2001). Reliability and concurrent validity of the information unit scoring metric for the story retelling procedure. *Aphasiology*, *15*(10–11), 991–1006. https://doi.org/10.1080/02687040143000348
McNeil, M. R., Doyle, P. J., Park, G. H., Fossett, T. R. D., & Brodsky, M. B. (2002). Increasing the sensitivity of the Story Retell Procedure for the discrimination of normal elderly subjects from persons with aphasia. *Aphasiology*, *16*(8), 815–822. https://doi.org/10.1080/02687030244000284
McVicker, S., Parr, S., Pound, C., & Duchan, J. (2009). The communication partner scheme: A project to develop long-term low-cost access to conversation for people living with aphasia. *Aphasiology*, *23*(1), 52–71. https://doi.org/10.1080/02687030701688783
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741
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*(2), 338–350. https://doi.org/10.1044/jshr.3602.338
Nicholas, L. E., & Brookshire, R. H. (1995). Presence, completeness, and accuracy of main concepts in the connected speech of non-brain-damaged adults and adults with aphasia. *Journal of Speech and Hearing Research*, *38*(1), 145–156. https://doi.org/10.1044/jshr.3801.145
Nippold, M. A., Cramond, P. M., & Hayward-Mayhew, C. (2013). Spoken language production in adults: Examining age-related differences in syntactic complexity. *Clinical Linguistics & Phonetics*, *28*(3), 195–207. https://doi.org/10.3109/02699206.2013.841292
Olness, G. S., Ulatowska, H. K., Wertz, R. T., Thompson, J. L., & Auther, L. L. (2002). Discourse elicitation with pictorial stimuli in African Americans and Caucasians with and without aphasia. *Aphasiology*, *16*(4–6), 623–633. https://doi.org/10.1080/02687030244000095

Perreira, N., Bresolin Goncalves, A. P., Goulart, M., Tarrasconi, M. A., Kochhann, R., & Paz Fonseca, R. (2019). Age-related differences in conversational discourse abilities: A comparative study. *Dementia & Neuropsychologia*, *13*(1), 53–71. https://doi.org/10.1590/1980-57642018dn13-010006

Prins, R., & Bastiaanse, R. (2004). Analysing the spontaneous speech of aphasic speakers. *Aphasiology*, *18*(12), 1075–1091. https://doi.org/10.1080/02687030444000534

Pulvermuller, F., & Berthier, M. L. (2008). Aphasia therapy on a neuroscience basis. *Aphasiology*, *22*(6), 563–599. https://doi.org/10.1080/02687030701612213

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ramsberger, G. (1996). Repetition of emotional and nonemotional words in aphasia. *Journal of Medical Speech-Language Pathology*, *4*(1), 1–12.

Ramsberger, G., & Rende, B. (2002). Measuring transactional success in the conversation of people with aphasia. *Aphasiology*, *16*(3), 337–353. https://doi.org/10.1080/02687040143000636

Reuterskiold, C. (1991). The effects of emotionality on auditory comprehension in aphasia. *Cortex*, *27*(4), 595–604. https://doi.org/10.1016/S0010-9452(13)80008-1

Richardson, J. D., & Dalton, S. G. (2016). Main concepts for three different discourse tasks in a large non-clinical sample. *Aphasiology*, *30*(1), 45–73. https://doi.org/10.1080/02687038.2015.1057891

Richardson, J. D., & Dalton, S. G. H. (2020). Main concepts for two picture description tasks: An addition to Richardson and Dalton, 2016. *Aphasiology*, *34*(1), 119–136. https://doi.org/10.1080/02687038.2018.1561417

Ross, K. B., & Wertz, R. T. (2003). Quality of life with and without aphasia. *Aphasiology*, *17*(4), 355–364. https://doi.org/10.1080/02687030244000716

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplist systematics for the organization of turn taking for conversation. *Language*, *50*(4), 696–735. https://doi.org/10.1353/lan.1974.0010

Sarno, M. T. (1969). *The functional communication profile*. Institute of Rehabilitation Medicine, NYU Medical Center.

Schulkind, M., Schoppel, K., & Scheiderer, E. (2012). Gender differences in authobiographical narratives: He shoots and scores; she evaluates and interprets. *Memory & Cognition*, *40*, 958–965. doi: 10.3758/s13421-012-0197-1,

Schum, R. L., & Sivan, A. B. (1997). Verbal abilities in healthy elderly adults. *Applied Neuropsychology*, *4*(2), 130–134. https://doi.org/10.1207/s15324826an0402_6

Shadden, B. (1997). Discourse behaviors in older adults. *Seminars in Speech and Language*, *18*(2), 143–157. https://doi.org/10.1055/s-2008-1064069

Simmons-Mackie, N. (2018). *The state of aphasia in North America: A white paper*. Aphasia Access.

Simmons-Mackie, N., Savage, M. C., & Worrall, L. (2014). Conversation therapy for aphasia: A qualitative review of the literature. *International Journal of Language & Communication Disorders*, *49*(5), 511–526. https://doi.org/10.1111/1460-6984.12097

Stark, B. C., Dutta, M., Murray, L. L., Bryan, L., Fromm, D., MacWhinney, B., & Sharma, S. (2020). Standardizing assessment of spoken discourse in aphasia: A working group with deliverables. *American Journal of Speech Language Pathology*, *1-12*. https://doi.org/10.1044/2020_AJSLP-19-00093

Teoh, V., Sims, J., & Milgrom, J. (2009). Psychosocial predictors of quality of life in a sample of community-dwelling stroke survivors: A longitudinal study. *Topics in Stroke Rehabilitation*, *16*(2), 157–166. https://doi.org/10.1310/tsr1602-157

Wallace, S. J., Worral, L., Rose, T., Le Dorze, G., Breitenstein, C., Hilari, K., Babbitt, E., Bose, A., Brady, M., Cherney, L. R., Copland, D., Cruice, M., Enderby, P., Hersh, D., Howe, T., Kelly, H., Kiran, S., Laska, A.-C., Marshall, J., Nicholas, M., & Webster, J. (2019). A core outcome set for aphasia treatment research: The ROMA consensus statement. *International Journal of Stroke*, *14*(2), 180–185. https://doi.org/10.1177/1747493018806200

Webster, J., Whitworth, A., & Morris, J. (2015). Is it time to stop "fishing"? A review of generalization following aphasia intervention. *Aphasiology*, *29*(11), 1240–1264. https://doi.org/10.1080/02687038.2015.1027169

Whitworth, A., Perkins, L., & Lesser, R. (1997). *Conversation analysis profile for people with aphasia*. Whurr.

Wilkinson, R. (2010). Interaction-focused intervention: A conversation analytic approach to aphasia therapy. *Journal of Interactional Research in Communication Disorders*, *1*(1), 45–68. https://doi.org/10.1558/jircd.v1i1.45

Wilkinson, R., & Wielaert, S. (2012). Rehabilitation targeted at everyday communication: Can we change the talk of people with aphasia and their significant others within conversation? *Archives of Physical Medicine and Rehabilitation*, *93*(1), S70–S76. https://doi.org/10.1016/j.apmr.2011.07.206

Wray, F., & Clarke, D. (2017). Longer-term needs of stroke survivors with communication difficulties living in the community: A systematic review and thematic synthesis of qualitative studies. *BMJ Open*, *7*(10), 1–18. https://doi.org/10.1136/bmjopen-2017-017944

Wright, H. H., Capilouto, G. J., Srinivasan, C., & Fergadiotis, G. (2011). Story processing ability in cognitively healthy younger and older adults. *Journal of Speech, Language, and Hearing Research*, *54*(3), 900–917. https://doi.org/10.1044/1092-4388(2010/09-0253)

Zec, R. F., Markwell, S. J., Burkett, N. R., & Larsen, D. L. (2005). A longitudinal study of confrontation naming in the "normal" elderly. *Journal of the International Neuropsychological Society*, *11*(6), 716–726. https://doi.org/10.1017/S1355617705050897