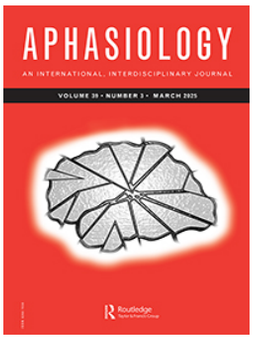

Aphasiology

ISSN: 0268-7038 (Print) 1464-5041 (Online) Journal homepage: www.tandfonline.com/journals/paph20

Reliability of the Brief Assessment of Transactional Success in Communication in Aphasia

Jacquie Kurland, Anna Liu, Vishnupriya Varadharaju, Polly Stokes & Robert Cavanaugh

To cite this article: Jacquie Kurland, Anna Liu, Vishnupriya Varadharaju, Polly Stokes & Robert Cavanaugh (2025) Reliability of the Brief Assessment of Transactional Success in Communication in Aphasia, Aphasiology, 39:3, 363-384, DOI: 10.1080/02687038.2024.2351029

To link to this article: https://doi.org/10.1080/02687038.2024.2351029

APHASIOLOGY
2025, VOL. 39, NO. 3, 363–384
https://doi.org/10.1080/02687038.2024.2351029

Routledge
Taylor & Francis Group

Reliability of the Brief Assessment of Transactional Success in Communication in Aphasia

Jacquie Kurland[a], Anna Liu[b], Vishnupriya Varadharaju[c], Polly Stokes[a] and Robert Cavanaugh[d]

[a]Department of Speech, Language, and Hearing Sciences, University of Massachusetts Amherst, Amherst, MA, United States; [b]Department of Mathematics and Statistics, University of Massachusetts Amherst, Amherst, MA, United States; [c]Department of Computer Information and Computer Science, University of Massachusetts Amherst, Amherst, MA, United States; [d]Northeastern University, Roux Institute, Boston, MA, United States

ABSTRACT

Background: While many measures exist for assessing discourse in aphasia, manual transcription, editing, and scoring are prohibitively labor intensive, a major obstacle to their widespread use by clinicians (Bryant et al. 2017; Cruice et al. 2020). Many tools also lack rigorous psychometric evidence of reliability and validity (Azios et al. 2022; Carragher et al. 2023). Establishing test reliability is the first step in our long-term goal of automating the Brief Assessment of Transactional Success in aphasia (BATS; Kurland et al. 2021) and making it accessible to clinicians and clinical researchers.

Aims: We evaluated multiple aspects of test reliability of the BATS by examining correlations between human/machine and human/human interrater edited transcripts, raw vs. edited transcripts, interrater scoring of main concepts, and test-retest performance. We hypothesized that automated methods of transcription and discourse analysis would demonstrate sufficient reliability to move forward with test development.

Methods & Procedures: We examined 576 story retelling narratives from a sample of 24 persons with aphasia and familiar and unfamiliar conversation partners (CP). Participants with aphasia (PWA) retold stories immediately after watching/listening to short video/audio clips. CP retold stories after six-minute topic-constrained conversations with a PWA in which the dyad co-constructed the stories. We utilized two macrostructural measures to analyze the automated speech-to-text transcripts of story retells: 1) a modified version of a semi-automated tool for measuring main concepts (mainConcept: Cavanaugh et al. 2021); and 2) an automated natural language processing "pipeline" to assess topic similarity.

Outcomes & Results: Correlations between raw and edited scores were excellent, interrater reliability on transcripts and main concept scoring were acceptable. Test-retest on repeated stimuli was acceptable. This was especially true of aphasic story retellings where there were actual within subject repeated stimuli.

ARTICLE HISTORY

Received 17 October 2023
Accepted 29 April 2024

KEYWORDS

Aphasia; story retelling; communication success; reliability; BATS

CONTACT Jacquie Kurland jacquie@umass.edu Associate Professor, Department of Speech, Language, & Hearing Sciences, University of Massachusetts Amherst, 358 North Pleasant Street, Amherst, MA 01003-9296, United States

Supplemental data for this article can be accessed online at https://doi.org/10.1080/02687038.2024.2351029

Conclusions: Results suggest that automated speech-to-text was generally sufficient in most cases to avoid the time-consuming, labor intensive step of transcribing and editing discourse. Overall, our study results suggest that natural language processing automated methods such as text vectorization and cosine similarity are a fast, efficient way to obtain a measure of topic similarity between two discourse samples. Although test-retest reliability for the semi-automated mainConcept method was generally higher than for automated methods of measuring topic similarity, we found no evidence of a difference between machine automated and human-reliant scoring.

Introduction

Models for assessing and treating aphasia have gradually moved from an impairment-based to a participation-based framework (Brady et al., 2016). A greater emphasis on communication has challenged the field to investigate ways of promoting aphasia recovery that can generalize to, if not focus on, real world language and communication skills (Carragher et al., 2015; deDe et al., 2019; Elman, 2007; Kagan, 1995, 1998; McVicker et al., 2009; Wilkinson & Wielaert, 2012). Unfortunately, various obstacles prevent clinicians and third-party payers from embracing a participation-based framework for assessing and treating aphasia, especially in the U.S. where healthcare is largely a for-profit enterprise. Clinicians in the U.S. and elsewhere have noted barriers in time, skill, efficiency, and confidence in the acquisition and analysis of conversation and other discourse data (Bryant et al., 2017; Cruice et al., 2020; Rose et al., 2014).

Despite an abundance of tools and procedures used in clinical research over the last 25 years to capture treatment effects in conversation (Azios et al., 2022), there are as yet no clinically convenient, psychometrically robust instruments that aphasiologists can recommend for everyday clinical use. Even in aphasia treatment research, there is only the beginning of consensus on a communication outcome measurement instrument (OMI). In the original Research Outcome Measurement in Aphasia Core Outcome Set (ROMA, COS; Wallace et al., 2019), there was not yet agreement on an OMI for the construct of communication. In the latest ROMA study, (ROMA-2; Wallace et al., 2023), one outcome measure, The Scenario Test (TST; van der Meulen et al., 2010), has now been recommended to be included in the aphasia COS. A Dutch test that has recently been translated into English (TST-UK; Hilari et al., 2018) and a few other languages, TST uses everyday scenarios (e.g., shopping, taking a taxi, visiting the doctor, etc.) to elicit verbal and non-verbal communication by asking a person with aphasia to role play being a character faced with a communicative task. By its very design, TST has excellent face validity. Pictured scenarios provide the impetus for eliciting functional communication in alignment with the "situated language use" model (Doedens & Meteyard, 2018), a model which defines language use as comprised of interactive, multimodal, context-specific joint action. Given its alignment with this theoretical model of communication, the TST was selected from among standardized and non-standardized tests, observational profiles, and linguistic and sociological analyses of connected speech and interaction as the best fit for assessing functional communicative ability in PWA in a clinical setting (Doedens &

Meteyard, 2020). However, the authors note a number of limitations of TST, chief among them that the need for role playing and lack of environmental referents may exert cognitive demands that do not reflect everyday communicative interactions, and that it is prone to ceiling effects for individuals with mild to moderate aphasia.

Nonetheless, it is clear that the field is moving at a fast pace to fill a critical void in the minimal set of treatment outcomes in aphasia, i.e., one that can assess change in real-life communication in PWA with a range of aphasia severity. In their recent comprehensive scoping review of conversation as a treatment outcome measure in aphasia, Azios and colleagues document 64 studies of 611 participants, revealing 211 different measures of conversation as a treatment outcome. Unfortunately, none of them demonstrated all the features of what may be considered an ideal measure, i.e., one that " . . . would be reliable within and across raters, achieve stability across time, be relevant and meaningful to people with aphasia, . . . feasible to administer in clinical settings . . . [and suitable] for assessing elements of conversation likely to change as a result of intervention" (p. 2936).

A new tool, the Brief Assessment of Transactional Success in aphasia (BATS; Kurland et al., 2021) aims to address this clinical-academic gap, while also providing methods for automating the labor-intensive aspects of discourse analysis. The BATS is a tool of testing "co-constructed communication" (Carragher et al., 2023; Goodwin, 1995). It includes a collection of 16 short video and audio clips from four different stimulus types, including humorous or "feel-good" stories, "how to" videos, biographical news clips, and interviews. The four types of stimuli vary along a continuum of dependency on verbal comprehension for understanding the stories.

Like the Scenario Test, the BATS is theoretically grounded in the situated language use model (Doedens & Meteyard, 2018). Unlike TST, the BATS which was inspired by Ramsberger and Rende's measure of transactional success (2002), uses story retelling before, during, and after topic-constrained conversation as the vehicle for assessing real-world communication. In their novel approach to alleviating problems associated with analyzing aphasic discourse, Ramsberger and Rende examined non-aphasic conversation partner retellings of stories co-constructed with individuals with aphasia who tried to convey a story they just watched. Analyzing the conversation partner retell provides evidence of content-related validity, i.e., that the test content is relevant to the proposed use of the test (Messick, 1995).

Another advantage is that transcription and analysis of a partner's retell is both less labor-intensive and more amenable to automated methodologies. Despite tremendous gains in automating transcription of aphasic discourse, including the new Batchalign automated "pipeline", that converts raw audio into Codes for the Human Analysis of Talk (CHAT) transcription format, the results of various measures such as word error rates from adults with language disorders vs. controls remains somewhat higher (Liu et al., 2023). Although such tools continue to improve, for the moment at least, analysis of the non-aphasic partner's retell may contribute to both a more reliable and a more clinically convenient instrument.

The reliability of a measure reflects its precision and dependability by assessing the consistency of scores of a group of test takers. This can be accomplished by evaluating scores between different raters (interrater reliability), by the same rater on different occasions (intra-rater reliability) and over repeated measurements of the same (test-retest reliability) or alternate forms of a test (internal consistency). Test-retest reliability

provides a measure of stability, repeatability, and consistency over time, and as such, is a particularly critical consideration in the design of stimuli, scoring protocols, and methods of interpreting and using scores (American Educational Research Association [AERA], American Psychological Association & National Council on Measurement in Education, 2014). Without test-retest reliability, inferences regarding treatment-induced change from repeated use of the same or equivalent stimuli may be spurious (Boyle, 2014). It is noteworthy how seldom psychometric properties such as test-retest reliability are reported on commonly used measures of discourse (Bryant et al., 2016; Pritchard et al., 2018).

The current study aims to evaluate the reliability of the BATS on traditional aspects of measurement variance in two macrostructural measures of analysis of the story retells, main concepts and topic similarity. Both measures are able to capture similar but distinct aspects of the conveyance and co-construction of information that underlies transactional success in story retelling, a context that is adjacent to conversation, especially in aphasia where the burden of communicating new information is more likely to be shared. As Ramsberger and Rende (2002) note, use of stories as "conversational topics" in aphasia can resemble natural conversation, while providing a way to measure the transfer of ideas. Thus topic-constrained, goal-oriented conversation includes shared turn-taking, conversation partners' orientation to the sequential analysis of utterances (Sacks et al., 1974), and a focus on achieving intersubjectivity, i.e., a shared interpretation of the subject (Klippi, 1996). Unlike natural conversation, which by its nature is not a replicable assessment task, story retelling, like other "co-constructed communication" tasks (Carragher et al., 2023), has enormous potential as a medium for automated discourse analysis. This is especially true for stories that have an original narration to which a story retell can be compared.

Main concepts analysis (MCA; J. D,. Richardson & Dalton, 2020; J. D. Richardson & Dalton, 2016; Nicholas & Brookshire, 1995), which measures the presence, accuracy, and completeness of relevant utterances, provides a measure of how well a speaker communicates concepts that are considered essential, i.e., the gist of a discourse. MCA begins with analysis of a non-clinical sample to develop a checklist of relevant and essential main concepts for any given discourse elicitation task. MCA has a long history of demonstrated reliability in aphasic discourse analysis and its utility, particularly in clinical research, has grown along with the AphasiaBank database (MacWhinney et al., 2011). With the recent development of the mainConcept app (Cavanaugh et al., 2021), which provides some automation of what can be a labor-intensive process of manually scoring main concepts, clinical feasibility of MCA is also likely to improve. We performed MCA using a modified version of the program developed by Cavanaugh and colleagues, i.e., using MC checklists for BATS stimuli developed in an earlier phase of test development (Kurland et al., 2021). Although the long-term goal is to obtain a fully automated tool for story retelling analysis, newer tools must be compared with existing ones to demonstrate test validity. In the current study, we focus just on reliability of MCA and topic similarity, the latter as measured by a fully automated "pipeline", or series of automated steps that transforms raw discourse data into a chosen output, in our case to assess cosine similarity between two discourse samples.

Judging topic similarity, formerly a skill unique to humans, is one among many automated language analysis tools. Whereas MCA uses human raters to judge how closely aligned a discourse sample is with the essential elements of a story as

determined by another set of human raters, cosine similarity is a natural language processing (NLP) lexical approach that assesses how closely aligned a discourse sample is with an original narration. We performed topic similarity analysis using NLP tools that assess the similarity between text documents on the basis of vector semantics. Vector semantics describes a fundamental aspect of machine learning of word meaning, in which it is hypothesized that words that occur in similar contexts are assumed to have similar meanings.

Two NLP methods were used to determine the cosine similarity between discourse samples, i.e., between the original narratives and the story retells – count vectorization and TF-IDF vectorization. Both methods convert text data into numerical vectors. Count vectorization represents text as a "bag of words" (Harris, 1954, as cited in Jurafsky & Martin, 2019), essentially counting each word in a matrix of token counts for each document. Count vectorization disregards word order and context. TF-IDF (term frequency – inverse document frequency) vectorization accounts for each word's frequency in the documents as well as the inverse document frequency or rarity of each word in the corpus. In so doing, it gives a higher rank to semantically weighted words (e.g., nouns, verbs, adjectives) than more frequently occurring functors (e.g., pronouns, prepositions, articles). Although the machine does not understand language per se, it "knows" a lot about words, having been trained on a large corpus of words, thus having learned representations of word meanings based on their distributions in texts. Obtaining the cosine similarity between two documents enables a comparison of topic similarity, a macrostructural level analysis of discourse that until recently relied on human language skills. The more similar two documents are, the higher their cosine similarity will be, with identical documents having a cosine similarity equal to 1.

We examined six aspects of test reliability, to establish that the tool is sufficiently reliable for clinical research and clinical practice, critical to our long-term goals of automating and disseminating the BATS: 1) reliability of story retell AI-generated automated transcription from speech: we hypothesized that most (raw) machine transcripts would be highly accurate, with some exceptions for participants with speech sound errors, distortions, or accents; 2) interrater reliability of edited transcripts: we expected little need for editing of automated transcripts and high interrater reliability among edited words and phrases; 3) reliability of using raw vs. edited transcripts in a measure of topic similarity: we hypothesized that there would be strong reliability between topic similarity scores of raw and edited transcripts, suggesting that AI-generated transcripts will usually suffice for fully automated analysis of topic similarity; 4) interrater reliability on scoring main concepts: we hypothesized that there would be strong reliability of main concept scoring using mainConcept; 5) within subject repeated stimuli test-retest reliability: given the well-known variability that can occur day-to-day in aphasia, we expected moderate-to-strong test-retest reliability when tested with identical stimuli 7–10 days apart; and 6) within subject repeated within stimulus type test-retest reliability: given that we hoped to avoid a learning effect in unfamiliar conversation partners by varying the stimuli (within stimulus type) that their partners with aphasia watched/listened to in the retest session with the same unfamiliar partner, we expected weaker, but acceptable, test-retest reliability in this condition.

Materials and Methods

Participants

Twenty-four persons with aphasia (PWA) and 24 non-aphasic familiar conversation partners (FCP) were recruited from aphasia centers and support groups in the U.S. The sample is a subset of an ongoing test development study of the BATS. An additional 38 non-aphasic unfamiliar conversation partners (UCP) were recruited via flyer, email, and word of mouth.

Inclusionary criteria for all participants included 18 years or older, fluent in English, with normal or corrected vision and hearing, no history of neurological conditions other than left hemisphere stroke in the aphasia group (at least three months post-onset), medically stable, willing to be videotaped retelling stories, and able to participate in study sessions via teleconference software, i.e., Zoom. Exclusionary criteria included history of significant psychiatric disease, drug or alcohol dependency, TBI with loss of consciousness and/or significant cognitive sequelae, chronic medical conditions likely to impair cognition, presence of visual field cuts or visual neglect, lack of technical skill or other resource for participating via Zoom. Screens were administered by telephone or over Zoom during the initial screening and consenting process. Participants with aphasia were screened using the Auditory Verbal Comprehension subtest of the Western Aphasia Battery (WAB-R; Kertesz, 2007), with a minimum required score used to calculate the aphasia quotient of 4.0. No one was excluded based on auditory comprehension (range: 6–10). We did not have an upper tier cutoff and anyone self-identifying as having had a stroke and living with chronic aphasia was included in the study, regardless of WAB-R aphasia quotient (WAB-AQ). We did not exclude participants with moderate-to-severe apraxia of speech. We also did not exclude participants whose first language was not English, provided they reported (and demonstrated) receptive and expressive fluency in English. The Telephone Interview for Cognitive Status (TICS; Brandt et al., 1988) was used as a cognitive screen for all non-aphasic conversation partners. Scores were all within normal limits.

All participants were pre-screened for other issues that might affect their performance on the tasks, including: 1) whether they wore glasses when using a computer; and 2) whether they wore hearing aids. Participants who responded positively to wearing glasses (n = 71; PWA = 18; CP = 53) and those who responded positively to wearing hearing aids (n = 1; PWA = 1; CP = 0) were told to wear them on each day of their participation in the study. All participants complied with this request.

The institutional review board of the University of Massachusetts Amherst approved the study, and informed consent was obtained from all participants via phone or video conferencing software (Zoom), with signatures obtained via DocuSign.

Stimuli

BATS stimuli consist of 16 short video and audio clips (mean = 2.55 minutes; *SD* = 0.50 minutes; range = 1.63–3.30 minutes). They include four non-verbal (NV) video clips, four "how to" videos in which visual and verbal (VV) information are approximately equivalent and synchronized, four short biographical video clips that are more reliant on auditory comprehension but with some visual support (VS), and four audio clips with only

Table 1. Descriptive information on video and audio stimuli.

#	Title	Condition	time (s)	Description	Discourse Genre	# of MCs
1	Bicycle Boy	NV	119	Silent video about doing good	story	10
2	Chaplin Shoe	NV	158	Silent video clip (Chaplin)	story	12
3	Share Care	NV	98	Silent video about doing good	story	10
4	Chaplin Shotgun	NV	157	Silent video clip (Chaplin)	story	15
5	Light Switch	VV	163	How to replace a light switch	procedural	9
6	Hang Blinds	VV	118	How to hang blinds	procedural	7
7	Curb Appeal	VV	150	How to improve your curb appeal	procedural	9
8	Fire Pit	VV	115	How to install a backyard fire pit	procedural	11
9	Marcus Yam	VS	198	Marcus Yam: photo journalist	autobiographical	11
10	Sylvia Earle	VS	181	Sylvia Earle: marine biologist	autobiographical	8
11	Naomi dela Rosa	VS	194	Naomi DLR: on family separation	autobiographical	8
12	Robin Steinberg	VS	128	Robin Steinberg: The Bail Project	autobiographical	7
13	Ferguson	SD	178	Ferguson protesters find friendship	interview	10
14	September 11	SD	172	Sept 11: One survivor's story	interview	12
15	Aunt Mother	SD	166	Aunt turned mother after tragedy	interview	7
16	No Handbook	SD	156	Mother/son on school shootings	interview	11

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; MCs = main concepts (Kurland et al., 2021).

one still photographic image that are mainly speech-dependent (SD) for complete understanding of the story. Descriptive characteristics of the 16 stimuli are reported in Table 1.

Data Collection Procedures

All data was acquired via Zoom video conferencing software. Most of the BATS study sessions (n = 66) were administered by the research speech-language pathologist (SLP; Stokes) with the others (n = 6) administered by the first author. Given the importance of assessment fidelity for reliability and validity of results (Richardson et al., 2016), significant time and preparation was invested in creating a testing manual, including checklists for training and ensuring consistent administration of the study protocol. All testing and study sessions were video recorded. The first twelve sessions and six additional randomly selected sessions were reviewed for adherence to the testing protocol using an administration fidelity checklist (Dekhtyar et al., 2020). Fidelity to consistent administration of the protocol was extremely high.

Discourse samples were digitally recorded in Zoom in three one-hour study sessions. Participants with aphasia first viewed and/or listened to each of four video/audio clips, including one 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 stimulus type. Participants with aphasia viewed the stimuli on a range of devices that included laptops (n = 18), PC or Mac desktops (n = 5), with no device smaller than an 8th generation iPad (n = 1). Prior to viewing the first stimulus, the participants were introduced (if unfamiliar). Familiar CP, if in the same dwelling and not connecting via separate devices, were asked to leave the room so that they could neither see nor hear the stimuli.

After viewing each stimulus, the participant with aphasia was prompted to "retell what each clip was about, in as much detail as you can remember". Immediately following this video recorded retell, the CP was brought back from the Zoom waiting room (or the FCP was texted to return if not on a separate device). They were instructed to engage in

a timed six-minute conversation with the goal of reaching a shared understanding of what the clip was about so that the CP could retell it in as much detail as possible. They were encouraged to use any verbal or nonverbal modality (gesture, writing, drawing, etc.) that might assist the person with aphasia to successfully communicate their ideas. When the timer went off, the CP was prompted to retell the story, using the same instruction as above, and the participant with aphasia was asked not to comment on the CP retelling. Both the conversation and the retell by the CP were also video recorded, and then the procedure repeated another three times. Each of the three study sessions including four BATS stimuli were administered in one hour or less.

Order of familiarity of conversation partners (CP) was counterbalanced between the first and second sessions, such that half of the time the familiar (FCP) went first, and half of the time the unfamiliar (UCP) went first. In the third session, acquired specifically to address test-retest stability in two conditions, participants with aphasia engaged in conversations with two UCP. They watched novel stimuli prior to meeting with the UCP they had met previously (UCP1). They watched stimuli they had seen before in an earlier session prior to meeting with a new UCP (UCP2). The order of meeting with UCP1 vs. UCP2 was also counterbalanced. Test-retest data collection occurred within 7 + 3 days, in accordance with standards proposed by the FOQUSAphasia Methodology and Data Quality (MDQ) Task Force (Stark et al., 2021). See Figure 1.

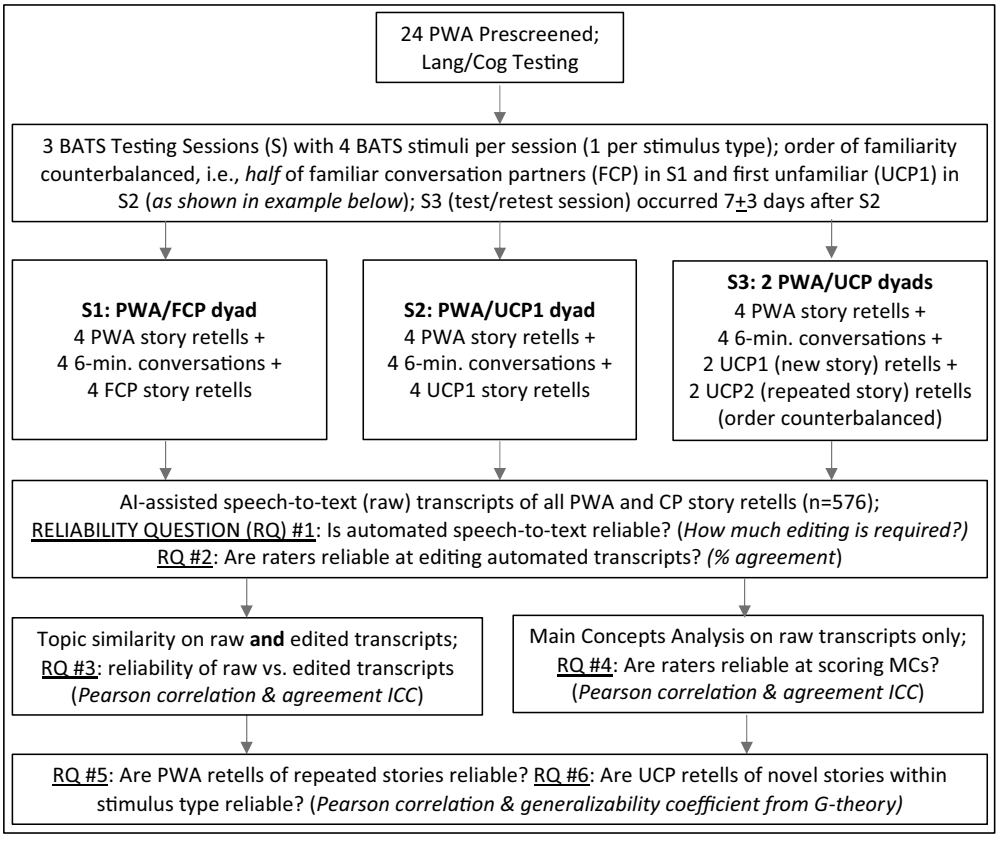

Figure 1. Schematic of experimental design and reliability research questions (RQ)

Generation and Pre-processing of Transcripts

De-identified audio files from PWA, FCP, and UCP story retells were transcribed using Python and Assembly AI's (https://www.assemblyai.com/) speech-to-text application programming interface, producing 576 transcripts. All transcripts were checked for transcription accuracy by research assistants, who were trained in a 2-hour session using transcripts and videos that had been previously checked by the first and fourth authors. All assistants reached 100% reliability during training, prior to independently checking transcripts for accuracy. The focus of editing of transcripts was on semantically weighted words and phrases, i.e., nouns and noun phrases, verbs and verb phrases, adjectives, and adverbs. Errors in transcribing fillers or part words or restarts were ignored. Only words that were missing or inaccurately transcribed and that changed the meaning of an utterance were changed in an edited text file. For example, if the program transcribed " . . . and uh, the um, aunt" as " . . . and the uh ant", only the word "ant" would be corrected to "AUNT". Corrections were capitalized for ease of counting the number of semantically-relevant editing changes. A random sample of 10% of CP transcripts were edited by a second research assistant and 10% of PWA transcripts by the first author, both blinded to the original edited transcripts for interrater reliability with respect to edited word changes. Reliability was 96% and 94% respectively. All disagreements were resolved by the first author.

Macrostructural Main Analyses

This reliability study focuses on two macrostructural main analyses that were performed on the story retells. Topic similarity is fully automated, i.e., it is scored completely by a computer program. Main Concept analysis uses a "semi-automated" application, wherein humans rate the utterances for presence, accuracy, and completeness of main concepts in the mainConcept app, which performs the scoring automatically. Automated data analysis was performed on all raw and edited transcripts. Semi-automated main concept (MC) data analysis was performed mainly on the raw transcripts. In addition, some MC analyses were performed on selected edited transcripts, i.e., for two participants with moderate-to-severe apraxia of speech.

Topic Similarity

Prior to performing automated analyses, transcripts underwent automated pre-processing steps that include removal of punctuation, stop words, and irrelevant comments. Examples of irrelevant comments, i.e., comments not germane to the story retell, included questions asked of the administrator at the beginning of a retell, e.g., "Should I go now?", or other opening comments, e.g., "Okay, let's see", and concluding comments, e.g., "I'm done". The CountVectorizer module from the machine learning Python library, scikit-learn (Bengfort et al., 2018) was used for text vectorization, i.e., to convert the original stimulus narrations and the discourse samples into numerical vectors of token counts. During text vectorization, the overall distribution of the words were captured. The cosine of the angle between two vectors, a very common metric that assesses topic similarity between two text documents (Jurafsky & Martin, 2022), was then computed. In this study, PWA and CP story retells were compared to the original narrations. Although

there are other text similarity metrics, we chose cosine similarity, given that: 1) it is suitable for handling documents of varying lengths, 2) it captures the semantic meaning of the text, instead of just capturing the frequency of words, and 3) it handles high dimensional sparse data in text well.

Main Concept Analysis

Main concept analysis was performed using a modified semi-automated application first developed by Cavanaugh et al. (2021). Using main concepts acquired from a normative sample of 96 non-aphasic story retellers (Kurland et al., 2021), we adapted their open-source web-app, mainConcept, to score presence, accuracy and completeness of main concepts (MCs) on PWA and CP story retellings of the 16 BATS stimuli. Because the stimuli have variable numbers of main concepts (see Table 1), an MC composite ratio is utilized on any comparisons between stimuli. For example, if a stimulus contains 10 MCs, the maximum composite score for a story retelling with all accurate and complete MCs would be 30. A participant scoring 15 on that story retell would thus have a MC composite ratio of 0.5.

Data Analysis

Raw (Unedited Speech-to-Text) vs. (Human) Edited Transcripts

To examine the reliability of the raw transcripts, we calculated the Pearson correlation between the cosine similarity scores of the raw and the human edited transcripts, and agreement ICC.

Interrater Reliability on Scoring Semi-Automated MCs

Ten percent of all semi-automated MCs were interrated by a second trained RA. Raters were masked to each other's results. Disagreements in MC composite scores greater than 3 points were resolved by the first author. The Pearson correlation and the agreement ICC of interrater scores are reported.

Test-Retest Reliability

Twenty-four participants with aphasia underwent three story retelling testing sessions, that included two sessions with two unfamiliar conversation partners. Stability between the first and second (test-retest) scores on topic similarity and main concepts was assessed using the Pearson correlation and the generalizability coefficient from the Generalization Theory (G-theory; Brennan, 2003). In applying G-theory, we used a linear mixed model to identify the sources of the variation in scores of topic similarity and main concepts, which includes the variations due to individuals ($\sigma^2(p)$), the stimuli ($\sigma^2(t)$) their interactions ($\sigma^2(pt)$), and the test sessions ($\sigma^2(s)$) and their interactions ($\sigma^2(ps)$, ($\sigma^2(pst)$)). The generalizability coefficient for test-retest reliability is calculated according to equation (40) in Brennan (2003), which is

$$\frac{\sigma^2(p) + \sigma^2(pt)}{\sigma^2(p) + \sigma^2(pt) + \sigma^2(ps) + \sigma^2(pst)}$$

The confidence intervals were generated based on the bootstrap technique of parametric methods with spherical random effects (Jiang et al., 2022). The calculations were carried out using the statistic software R (R Core Team, 2022) and the R package lme4 (Bates et al., 2015).

Test-retest stability was tested for PWA story retells under two conditions: 1) on repeated (identical) stimuli, and 2) on stimuli from the same stimulus types. These test-retest samples were acquired in sessions with different UCP. Test-retest reliability for UCP story retells was also assessed under these same two conditions, for UCP who participated in sessions with the same conversation partner with aphasia.

Finally, to compare the two different scoring methods, topic similarity versus main concepts, in terms of their test-retest reliability, we built the same linear mixed models as above which included scores from both methods and allowed correlation among the scores as well as method-specific variance components. Testing of the differences between the reliability coefficients was carried out using the same parametric bootstrapping technique (Jiang et al., 2022).

Results

Participants

The study included 24 PWA, 24 FCP, and 38 UCP. Most PWA were chronically aphasic, i.e., more than six months post-onset, with an average time poststroke of 6.6 years (*SD* = 5.5, range = 0.3–22). PWA were classified according to the WAB-R (Kertesz, 2007), modified for remote delivery (Dekhtyar et al., 2020). Average WAB-AQs were 79.9 (*SD* = *16.3*, range = 34.1–94.9). Most PWA (*n* = 22) were monolingual English speakers, and all 24 reported English as their primary language. Most CP (22 FCP; 34 UCP) had at least an Associate's degree. All were monolingual English speakers who had completed high school. Individual clinical characteristics for PWA, and demographic data for all participants are reported in Tables 2 and 3.

Reliability of Raw vs. Edited Transcripts

Based on cosine similarity measures (count and TF-IDF), the correlation between raw and edited scores were both 0.99 and the ICC were both 0.99 with 95% CI (0.989, 0.992).

Interrater Reliability on Scoring Semi-Automated MCs

The Pearson correlation for the two raters' overall ratings was 0.87, which remained the same for PWA retells only and CP retells only. The overall ICC was 0.86 with 95% CI (0.77, 0.92). The ICC for PWA retells was 0.86 with CI (0.7, 0.94), and the ICC for CP retells was 0.87 with 95% CI (0.72, 0.94).

Test-Retest Reliability

Table 4 reports within subject test-retest reliability measures for PWA and for UCP on both strictly matched and loosely matched stimuli on measures of cosine similarity and main concepts. Strictly matched stimuli included only those stimuli repeatedly watched and then retold (two per PWA). Loosely matched stimuli were those from the same stimulus type, but not necessarily the same stimulus. The general trend suggested that reliability was higher, or at least as high, for: 1) strictly matched vs. loosely matched stimuli,

Table 2. Clinical and demographic information for the sample of 24 persons with aphasia (PWA)

ID	Age (yrs)	Gender	Approx. time post- onset (yrs)	WAB Aphasia Classification	WAB AQ	WAB Aud. Verbal Comp.	Educ. (yrs)	Highest degree {GED,HS,A, B,M,PhD)	Race/ ethnic background {AA,W,HL)
PWA1	76	M	3	Anomic	88. 2	8.9	20	M	AA
PWA2	60	M	2.5	Anomic	90.9	9.3	12	HS	W
PWA3	70	F	3	Anomic	82.9	9.9	24	M	W
PWA4	76	F	3	Anomic	87.8	10.0	24	PhD	W
PWA6	83	F	10	Anomic	87.1	10.0	16	B	W
PWA8	75	M	9	Conduction	88. 2	9.2	24	M	W
PWA10	72	M	04	Anomic	94.9	10.0	21	M	W
PWA11	62	F	2	Anomic	93. 2	10.0	16	A	W
PWA12	69	M	17	Wernickes	54.1	6.4	16	B	W
PWA14	57	M	1	Anomic	94.9	9.9	16	B	W
PWA15	58	M	0.3	Broca's	51.7	7.2	12	B	W
PWA16	60	F	8	Conduction	77.3	9.8	12	GED	AA
PWA17	50	F	22	Anomic	81 0	9.4	18	M	W
PWA20	58	F	10	Anomic	91.7	10.0	18	M	W
PWA21	58	M	8	Broca's	62.6	7.8	16	B	W
PWA22	64	M	6	Broca's	34.1	6.0	12	HS	AA
PWA23	62	M	12	Anomic	94.5	9.7	18	M	W
PWA25	56	M	3	Conduction	63.6	6.7	16	B	W
PWA26	64	M	4	Anomic	94.0	10.0	18	M	W
PWA27	48	M	15	Anomic	80.8	7.8	16	B	W
PWA32	70	M	5	Anomic	87.8	10.0	18	M	W
PWA33	53	F	3	Conduction	69.1	7.3	16	B	W
PWA38	60	F	7	Anomic	72.5	8.3	14	A	HL
PWA44	62	F	4	Anomic	934	9.5	18+	M	W

Notes: WAB AQ = Western Aphasia Battery Aphasia Quotient (Kertesz, 2006); M = male; F = female GED = high school equivalency degree; HS = high school diploma; A = Associates degree; B = Bachelors degree; M = Masters degree; PhD = Doctoral degree; AA = African American; W = White/Caucasian; HL = Hispanic/Latinx.

Table 3. Demographic information for all groups.

Group	*N*	Age (years)	Gender	Education (years)	Race/ethnicity	TICS scores
FCP	24	50.79 *(15.61)* range: 21–76	20 female 3 male 1 non-binary	17.25 *(2.47)* range: 12–24	24 Caucasian	35.92 *(1.64)* range: 33–38
UCP	38	39.97 *(18.85)* range: 18–73	25 female 13 male	16.68 *(2.23)* range: 12–22	36 Caucasian 1 Hispanic/Latino 1 Asian (Eastern)	37.0 *(1.7)* range: 34–41
PWA	24	63.5 *(8.8)* range: 48–83	10 female 14 male	17.09 *(3.65)* range: 12–24	20 Caucasian 3 African American 1 Hispanic/Latina	n/a

Notes: FCP = familiar conversation partners; UCP = unfamiliar conversation partners; PWA = persons with aphasia; TICS = Telephone Interview for Cognitive Status (Brandt et al., 1988); Age, Education, and TICS scores are mean *(sd)*

Table 4. Pearson and G-theory measures of test-retest reliability.

Group/Measures	Reliability method	Strictly matched	Loosely matched
PWA within subject measures			
Cosine similarity (count)	Pearson correlation	0.76	0.70
	G-theory coeff. *(CI)*	0.76 *(0.60, 0.86)*	0.76 *(0.65, 0.87)*
Cosine similarity (TF-IDF)	Pearson correlation	0.78	0.69
	G-theory coeff. *(CI)*	0.74 *(0.55, 0.86)*	0.75 *(0.64, 0.86)*
MC composite ratio	Pearson correlation	0.86	0.73
	G-theory coeff. *(CI)*	0.82 *(0.68, 0.90)*	0.82 *(0.69, 0.90)*
UCP within subject measures			
Cosine similarity (count)	Pearson correlation	0.62	0.51
	G-theory coeff. *(CI)*	0.56 *(0.16, 0.78)*	0.56 *(0.38, 0.77)*
Cosine similarity (TF-IDF)	Pearson correlation	0.63	0.51
	G-theory coeff. *(CI)*	0.60 *(0.22, 0.82)*	0.56 *(0.38, 0.77)*
MC composite ratio	Pearson correlation	0.89	0.65
	G-theory coeff. *(CI)*	0.85 *(0.63, 0.94)*	0.86 *(0.73, 0.93)*

Notes: PWA = persons with aphasia; UCP = unfamiliar conversation partners; TF-IDF = term frequency-inverse document frequency; MC = main concepts;G-theory coeff.*(CI)* = Generalization theory reliability coefficient *(confidence interval)*; strictly matched refer to repeated stimuli, whereas loosely matched stimuli refer to stimuli within the same stimulus type.

regardless of the metric (cosine similarity vs. MCs) or the story reteller group (PWA vs. UCP); 2) MC composite ratio scores vs. cosine similarity scores, regardless of the degree of stimulus matching (strictly vs. loosely) or the story reteller group; and for 3) most PWA vs. UCP repeated measures, with the exception of MC composite ratio scores on strictly matched stimuli, where test-retest stability was slightly higher for UCP than PWA, even though in this condition, the comparison was always between two different UCP. The differences in the reliability coefficients were not statistically significant though. For example, the p-values were 0.66 and 0.68 when comparing the reliability coefficients of Cosine similarity (count) versus MC composite ratio, and Cosine similarity (TF-IDF) versus MC composite ratio for strictly matched PWA retells.

Comparisons of PWA, FCP, and UCP Story Retells

Although not related to reliability, we also examined story retelling performance between groups, i.e., between aphasic and non-aphasic story retells and between familiar and unfamiliar conversation partner story retells. Results can be found in Supplemental Figures 2, 3, and 4.

Discussion

The purpose of the current study was to examine various types of reliability related to semi-automated and automated measures for scoring story retelling discourse in persons with aphasia and their familiar and unfamiliar conversation partners. The results suggest that automated speech-to-text produces transcripts that are very accurate in most cases, with exceptions described below. Results also suggest that natural language processing (NLP) automated methods such as text vectorization and cosine similarity are a fast, efficient way to obtain a measure of topic similarity between two discourse samples. Even though test-retest reliability was generally higher for the semi-automated than fully-automated method, we found no evidence suggesting a statistically significant difference in reliability between the two methods.

NLP vs. Human Methods of Assessing Story Retelling

Two fully-automated NLP methods were used to determine the cosine similarity between discourse samples, i.e., by comparing aphasic and non-aphasic story retells to the original narratives, using count vectorization and TF-IDF vectorization. Obtaining the cosine of the angle between two document vectors, i.e., a story retell and the story's original narration, provides an automated method of assessing topic similarity.

In our study, when compared with the original narrative using count vectorization, story retells ranged on average between 0 and 0.63, with some as high as 0.73. Using TF-IDF, the range on average was 0 to 0.39, with some as high as 0.51. Cosine similarity scores can vary between 0.0 and 1.0, with higher scores indicating story retells that are closer to the original narrations. To obtain a score of 0.0, a story retell would have zero content in common with the original narration, while a score of 1.0 would indicate an exact copy of the original narration. For comparison, it may be helpful to look at scores we obtained in a test development study in which we acquired 768 story retells from 96 non-aphasic persons immediately after watching or listening to BATS stimuli. Using count vectorization, story retells ranged between 0.18 and 0.86 (mean = 0.64; sd = 0.11). Using TF-IDF, story retells ranged between 0.07 and 0.64 (mean = 0.39; sd = 0.10).

It may also be helpful to look at the extremes, e.g., participants with very mild and very severe expressive aphasia. Only one participant (PWA22) obtained cosine similarity scores of 0 retelling four different stories. In each of these cases, the conversation partner's retell scored higher (count range: 0.2–0.27; TF-IDF range: 0.09–0.10). A similar pattern was observed across most low-performing participants with aphasia (count: < 0.2; TF-IDF: < 0.1), such that both FCP and UCP story retells tended to obtain higher, and often much higher, cosine similarity scores than their conversation partners with aphasia, particularly when the latter were severely nonfluent. This was not surprising, but rather confirms Audrey Holland's much-quoted observation that "people with aphasia often communicate better than they talk" (Holland, 1977). Indeed, Holland's premise is the *raison d'être* for developing the BATS tool, i.e., to provide a psychometrically robust, clinically feasible tool for assessing how people with aphasia use language in everyday contexts such as story retelling. Returning to the extremes, the opposite pattern held for the highest-performing participants with aphasia (count: > 0.6; TF-IDF > 0.35), i.e., there was a tendency for both FCP and UCP story retells to obtain lower, at times significantly

lower, cosine similarity scores than their conversation partners with aphasia, especially when the latter were only mildly anomic.

Main concept composite (MCComp) ratio scores generally reflected these same two tendencies for the lowest (<0.25) and highest (>0.75) performing participants with aphasia and their conversation partners. In general, the MCComp ratio difference in scores observed between PWA and FCP, and between PWA and UCP, reflected similar trends as those obtained via cosine similarity, despite the fact that these measures capture different constructs. Whereas cosine similarity, a fully automated metric, captures the overlap in subject matter between a story retell and its original narration based on word meaning, MCComp ratio is determined by human raters judging a story retell against a set of main concepts for that story, as pre-determined by a normative sample (Kurland et al., 2021). Main concepts are thus presumed to indicate story gist, even though there is a wide range in what non-aphasic story retellers might consider to be the gist of any given story. Moreover, the scoring of the presence, accuracy, and completeness in story retells, given a pre-determined set of MCs, is not a perfectly reliable measure, either within or between raters, and raters must first be trained to a standard of reliability. Thus, there may be some layering of instability inherent in the MC measure simply because it relies on human judgment. Developing a fully automated, reliable macrostructural measure that is not reliant on human raters could make discourse analysis even more accessible to clinicians and clinical researchers.

Raw vs. Edited Transcripts

Given the long-term goal of developing automated methods for discourse analysis that would be accessible to clinicians and clinical researchers, a fundamental issue is whether the automated speech-to-text process is reliable. In other words, are automated transcripts "good enough" to use without editing, a step that makes the process less clinically feasible due to time and labor constraints. We tested both the raw and edited transcripts for cosine similarity with the original narratives and found very high correlations even before removing outliers, suggesting that the automated transcripts can be used without editing, with some caveats.

Among the few outliers, two participants with severe expressive aphasia had little to no topic-oriented content in their story retells, and two participants had moderate-to-severe distortions in their speech which contributed to their transcripts being less reliable than others. In general, in cases where participants' speech is sparse and/or mostly nonverbal, severely dysarthric or apraxic, or when a non-native English speaker's accent is pronounced, use of automated transcripts is less reliable. If the goal is analysis of those PWA retells, then it is advised that the transcripts be edited for accuracy; however, if the goal is analysis of the CP's retell, as eventual use of the BATS will enable, then aphasic speech intelligibility is less of an issue. A second caveat is that not all speech-to-text programs are equally reliable and/or affordable. For example, the Zoom transcripts that are produced along with the audio files are free, but not sufficiently accurate. In general, automated speech-to-text applications are becoming increasingly accurate, affordable, and accessible. This is good news for development of an automated pipeline for the BATS.

Interrater Reliability on MC Scoring

In prior studies of aphasic and non-aphasic discourse, MCA has been observed to have "acceptable" reliability across raters, i.e., consistently at or above 80% (Boyle, 2014; Nicholas & Brookshire, 1995). Using the AphasiaBank picture description, picture series, wordless picture book, and procedural discourse stimuli (MacWhinney et al., 2011), Richardson and Dalton observed higher interrater reliability scores, ranging between 0.90 and 1.00 (2016) and between 0.88 and 0.93 (2020) in non-aphasic samples. Richardson et al. (2018) observed 0.975 interrater reliability using a large clinical sample and three of the AphasiaBank stimuli, however the intrarater reliability was lower (0.90) in the same study.

In the current study, both overall and by participant group, interrater reliability on MC scoring was acceptable (0.87, with confidence intervals ranging between 0.71 and 0.94). This is slightly lower than interrater reliability using point-by-point comparison in our Phase I test development study (Kurland et al., 2021). However, that study sample consisted entirely of non-aphasic participants who retold stories immediately after watching/listening to them.

Another factor that may influence interrater reliability of MC scoring in our study is the nature of the BATS stimuli. BATS stimuli were purposefully selected to be relatable stories, often reflecting current events, at times controversial and often producing strong emotional responses in the narrative retells. Some of the MCs reflect this complexity and seem to elicit more judgments of implicit information than is typically elicited by the static images traditionally used to elicit narrative discourse. For example, there are fewer implicit judgments a rater needs to make if the main concepts are, "The boy was outside. He was playing soccer . . . The man looked out of the window" than if the main concepts are, "A little boy finds money stuck on his tire. He fantasizes about the sweets that he could buy . . . The video encourages acts of kindness".

It is, in fact, notable that the interrater reliability is as high as what we observed, given the number of opportunities for rater judgments that may vary with their own experience and world knowledge. Hence developing a fully automated tool that is approximately as reliable and considerably faster at analyzing story gist, without the instability of human raters, would render discourse analysis more clinically feasible.

Test-Retest Stability

Test-retest reliability is a fundamental measure of the stability, repeatability, and consistency of a test over time. We found acceptable correlation of test-retest reliability within groups (PWA and UCP), on both cosine similarity and MCs, more so with strictly matched, than loosely matched stimuli. Most PWA retells scored slightly higher on both measures when tested on the same stimuli a second time within 7 + 3 days, although there were some notable outliers. Two participants with severe aphasia scored significantly lower on cosine similarity measures, while a few outliers scored significantly higher on both cosine similarity and MC measures.

Since the the current study focused on communication success in the context of co-constructed story retelling, test-retest reliability was more critical to establish in conversation partners than in PWA. One limitation in the study was that different UCP were used in

strictly matched stimuli, to avoid a learning effect of the stimuli. Similarly, no FCP were used in the test-retest portion of the study, given the potential for learning the stimuli. Thus, returning UCP (UCP1) in their second study session co-constructed narratives with their same PWA that described different stimuli that those they had been exposed to in their first study session. New UCP (UCP2) were exposed to stimuli that the PWA were watching/listening to for a second time. In both cases, some additional variability (stimulus variability in the former, individual participant variability in the latter) was introduced. Despite this additional variability, UCP test-retest reliability was strong for the MC measure and moderate/acceptable for cosine similarity measures.

Given the tendency for participants with aphasia to recall, on average, a little more detail in a second retell of the same story, it is reasonable to assume that there might be a learning effect for multiple repeated measures. For this reason, it was our intention in selection of stimuli during test development to have equivalent sets of stimuli. We included stimuli from similar genres, e.g., NPR StoryCorps interviews, PBS "Brief But Spectacular" autobiographical stories, Loews procedural/how-to videos, Chaplin silent comic clips, and "feel good" silent stories. However, these loosely matched stimuli (within stimulus type) had acceptable, but lower test-retest reliability for both PWA and UCP. Formal investigation into alternate or equivalent forms is ongoing.

Influence of Aphasia Severity

This reliability sample, a subset of a larger test development study, was mostly a convenience sample of the first 24 PWA who agreed to undergo a third BATS study session. The sample included individuals across a broad spectrum of aphasia severity. Although aphasia severity predictably influenced CP narrative retells, with a few exceptions at the very low- and very high-performing ends of the spectrum, it did not diminish test reliability. As noted earlier, severity can influence automated transcript reliability, and therefore is a factor to consider if one objective is analysis of the PWA story retell or the dyadic conversation. Since one of the long-term objectives of BATS test development is to deliver an automated application that could uniquely focus on the CP retell, reliability of PWA speech-to-text in this instance is not an issue.

As Goodglass observed, the more severe the aphasia, the greater " . . . the need for inference, questioning, and guessing by the listener . . . [and the more] the listener carries the burden of communication" (Goodglass et al., 2001, p. 8). Moreover, as noted by Goodglass and others, aphasia severity plays a role in, but is not determinative of communication success when language is used in its everyday currency, i.e., the daily exchange of information between people. Language use, as Clark (1996) observed, is *joint action* and it matters who is on the "receiving" end of a story when assessing the narrative retells of a conversation partner. This was clearly borne out in the current study sample, in which CP demonstrated a range of knowledge about aphasia generally, and ability to support conversations with PWA. Some CP were more at ease conversing with a person with aphasia, even severe aphasia. Some were more outgoing, or more tuned in to the task of co-constructing the stories. This held for both familiar and unfamiliar partners, demonstrating that familiarity is not necessarily a key ingredient when it comes to conversation partners' skill level in acknowledging and revealing competence in conversation with individuals with aphasia (Kagan et al., 2004). The ongoing investigation of

multiple factors contributing to transactional success during the topic-constrained conversations, including how they interact with aphasia severity, is beyond the scope of the current study.

Conversation vs. Co-constructed Communication

Elsewhere, we have made an argument for including conversation in a core outcome set of discourse in aphasia (Kurland & Stokes, 2018). Story retelling is a close ally to conversation, particularly in the context of aphasia, where the original story is known and the person with aphasia is tasked with conveying new information to a "naïve" conversation partner (Ramsberger & Rende, 2002). As Carragher et al. (2015) note, there are many advantages to using story retelling over traditional language assessment, including the social perspective that is lacking in monologic discourse, the face validity of achieving success using interactive, multimodal communication, the clinical benefit of enhancing social/conversational opportunities in an arena that is often significantly diminished post-stroke, the linguistic richness of narrative production, and the methodological rigor, including reproducibility and opportunities for comparison across individuals with and without aphasia.

Story retelling is only one of numerous assessment tasks in the genre of co-constructed communication recently identified in a comprehensive scoping review (Carragher et al., 2023). Carragher and colleagues identified 37 studies in five categories of co-constructed communication, including referential communication tasks, telephone enquiry tasks, joint problem-solving tasks, a collaborative naming task, and message exchange tasks that include semi-structured, transactional conversations like in the current study. Similar to the findings of Azios et al. (2022), Carragher and colleagues found a proliferation of measures (n = 95) derived from this relatively small number of studies. Unfortunately, they also found, similar to Azios and colleagues, that most measures require further investigation of psychometric properties including validity and reliability.

We have found the story retelling paradigm using the BATS stimuli to be a reliable method of assessing communication success during the exchange of new information between persons with aphasia and familiar and unfamiliar conversation partners. Moreover, in our pursuit of developing a free, accessible, clinically practical tool for measuring communication success in aphasia, we are investigating multiple automated methods of analyzing the end product, including the story retell by both conversation partners after a timed, topic-constrained conversation with a person who has aphasia. While we acknowledge that natural conversation in aphasia is the behavior we are ultimately interested in assessing, and that natural language processing tools are rapidly evolving and may ultimately be able to handle natural conversation, in the current study we are focused on automated methods that can reliably assess co-constructed communication success in aphasia via the conversation partner's story retell.

Concluding Remarks, Limitations, and Future Directions

Strong and Shadden (2021) discuss "small and big stories" that are "at the heart of human interaction and life participation". They implore clinicians to be more attentive to story as a means of enabling successful living with aphasia. The embrace of a life participation-

based framework for assessing and treating aphasia by clinicians, insurers, grant-awarding institutions, and other stakeholders has at times been achingly slow, especially as considered in the context of Audrey Holland's persistent pleas over the last half century. Some of the resistance may be well-founded in that tools for measuring participation-based treatment outcomes, while numerous, have not always been developed with psychometrically rigorous methods (Azios et al., 2022; Carragher et al., 2023). Nevertheless, we may be on the precipice of a sea change in our ability to harness natural language processing tools that take advantage of artificial intelligence to analyze discourse accurately, reliably, and efficiently.

This study describes the first demonstration of various aspects of reliability of the BATS, a tool that will ultimately enable a psychometrically robust, clinically accessible method of measuring outcomes in participation-based therapies. The current study is limited in scope to assessing the tool's reliability by focusing on: 1) demonstrating dependability of automated vs. manual transcription and scoring; and 2) the lexical content of conversation partner story retells, an important transactional product of co-constructed communication, but one which does not directly measure it. As such, the critical interactional and transactional aspects of conversation that characterize the situated language use theoretical model on which the tool was developed are not directly instantiated in this first examination of some aspects of the tool's reliability.

Although perhaps not evident in the current study, the BATS is a tool that we envision will someday be useful, practical, and accessible to clinicians and clinical researchers who may be interested in examining a multitude of clinically relevant PWA characteristics, including but not limited to baseline monologic story retelling ability, transactional success in conveying a story to a conversation partner, characteristics of PWA and CPs that enhance vs. inhibit successful co-constructed communication, and treatment-induced change in such ability. As we continue to demonstrate both the reliability and validity of the instrument, including a checklist of dyadic behaviors, and develop the tools for automating the process of acquiring and analyzing the discourse data, it is our hope that the BATS will be adopted as an outcome measurement instrument to assess communication in a future core outcome set.

Analysis of both PWA and CP variables along with other multivariate sources of variability in communication success are ongoing as we examine the full set of narrative retells (n = 1728) and dyadic conversations (n = 864). With the exception of participants who did not consent to it, most of the discourse samples are being shared with the AphasiaBank (MacWhinney et al., 2011), in the interest of open source data sharing. In the meantime, investigations of the validity and reliability of this tool, including through the use of state-of-the-art automated tools for analysis, are ongoing and will be reported elsewhere.

Acknowledgments

Research reported in this publication was entirely supported by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under award number R21DC020265. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We would like to thank research assistants (Caroline Pare and Mia Tittmann) and all our study participants.

Disclosure statement

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

Funding

This work was supported by the National Institute on Deafness and Other Communication Disorders [R21DC020265].

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (Eds.). (2014). *Standards for educational and psychological testing*.

Azios, J. H., Archer, B., Simmons-Mackie, N., Raymer, A., Carragher, M., Shashikanth, S., & Gulick, E. (2022). Conversation as an outcome of aphasia treatment: A systematic scoping review. *American Journal of Speech Language Pathology*, *31*(6), 2920–2942. https://doi.org/10.1044/2022_AJSLP-22-00011

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. http://dx.doi.org/10.18637/jss.v067.i01

Bengfort, B., Bilbro, R., & Ojeda, T. (2018). *Applied text analysis with Python: Enabling language-aware data products with machine learning* (1st) ed.). O'Reilly Media.

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*. https://doi.org/10.1002/14651858.CD000425.pub4

Brandt, J., Spencer, M., & Folstein, M. (1988). The telephone interview for cognitive status. *Neuropsychiatry, Neuropsychology and Behavioral Neurology*, *1*(2), 111–117.

Brennan, R. L., (2003). *Coefficients and indices in generalizability theory*. (CASMA Research Report No. 1). Center for Advanced Studies in Measurement and Assessment, The University of Iowa. https://education.uiowa.edu/sites/education.uiowa.edu/files/2021-11/casma-research-report-1.pdf

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

Carragher, M., Mok, Z., Steel, G., Conroy, P., Pettigrove, K., Rose, M. L., & Togher, L. (2023). Towards efficient, ecological assessment of interaction: A scoping review of co-constructed communication. *International Journal of Language & Communication Disorders*, 1–45. Advance online publication. https://doi.org/10.1111/1460-6984.12957

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

Cavanaugh, R., Richardson, J., & Dalton, S. G. (2021). *mainConcept: An open-source web-app for scoring main concept analysis. R package version 0.0.1.0000*. https://github.com/aphasia-apps/mainConcept

Clark, H. H. (1996). *Using language*. Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511620539

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

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

Dekhtyar, M., Braun, E. J., Billot, A., Foo, L., & Kiran, S. (2020). Videoconference administration of the Western Aphasia Battery-Revised: Feasibility and validity. *American Journal of Speech Language Pathology*, *29*, 673–687. https://doi.org/10.1044/2019_AJSLP-19-00023

Doedens, W. J., & Meteyard, L. (2018, July 31). The importance of situated language use for aphasia rehabilitation. http://dx.doi.org/10.31231/osf.io/svwpf

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

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 ed., pp. 39–50). Thieme Medical. https://doi.org/10.1097/01.TLD.0000299884.31864.99

Goodglass, H., Kaplan, E., & Barresi, B. (2001). *The assessment of aphasia and related disorders*. Pro-Ed.

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

Hilari, K., Galante, L., Huck, A., Pritchard, M., Allen, L., & Dipper, L. (2018). Cultural adaptation and psychometric testing of The ScenarioTest UK for people with aphasia. *International Journal of Language & Communication Disorders*, *53*(4), 748–760. https://doi.org/10.1111/1460-6984.12379

Holland, A. L. (1977). Some practical considerations in aphasia rehabilitation. In M. Sullivan & M. S. Kommers (Eds.), *Rationale for adult aphasia therapy* (pp. 167–180). University of Nebraska Medical Center.

Jiang, Z., Raymond, M., diStefano, C., Shi, D., Liu, R., & Sun, J. (2022). A Monte Carlo study of confidence interval methods for generalizability coefficient. *Educational and Psychological Measurement*, *82*(4), 705–718. https://doi.org/10.1177/00131644211033899

Jurafsky, D., & Martin, J. H. (2019). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3rd ed.). Prentice Hall. https://web.stanford.edu/~jurafsky/slp3/

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*, 851–864. https://doi.org/10.1080/02687039808249575

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*, 67–83. http://dx.doi.org/10.1310/CL3V-A94A-DE5C-CVBE

Kertesz, A. (2007). *Western Aphasia Battery – Revised*. Harcourt Assessment, Inc.

Klippi, A. (1996). Conversation as an achievement in aphasics. *Studia Fennica Linguistica*, *6*, 201–214.

Kurland, J., Liu, A., & Stokes, P. (2021). 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*, 39–68. https://doi.org/10.1080/02687038.2017.1398808

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

Liu, H., MacWhinney, B., Fromm, D., & Lanzi, A. (2023). Automation of language sample analysis. *Journal of Speech, Language, and Hearing Research*, *66*(7), 2421–2433. https://doi.org/10.1044/2023_JSLHR-22-00642

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

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*, 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. (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*, 145–156. http://dx.doi.org/10.1044/jshr.3801.145

Pritchard, M., Hilari, K., Cocks, N., & Dipper, L. (2018). Psychometric properties of discourse measures in aphasia: Acceptabiity, reliability, and validity. *International Journal of Language and Communication Disorders*, *53*(6), 1078–1093. https://doi.org/10.1111/1460-6984.12420

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

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

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

Richardson, J. D., Dalton, S. G., Shafer, J., & Patterson, J. (2016). Assessment fidelity in aphasia research. *American Journal of Speech-Language Pathology*, *25*(4S), 788–797. https://doi.org/10.1044/2016_AJSLP-15-0146

Rose, M., Ferguson, A., Power, E., Togher, L., & Worrall, L. (2014). Aphasia rehabilitation in Australia: Current practices, challenges and future directions. *International Journal of Speech-Language Pathology*, *16*(2), 169. https://doi.org/10.3109/17549507.2013.794474

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

Stark, B. C., Dutta, M., Murray, L. L., Bryan, L., Fromm, D., MacWhinney, B., & Sharma, S. (2021). Standardizing assessment of spoken discourse in aphasia: A working group with deliverables. *American Journal of Speech Language Pathology*, *30*(1S), 491–502. https://doi.org/10.1044/2020_AJSLP-19-00093

Strong, K. A., & Shadden, B. B. (2021). Stories at the heart of life participation: Both the telling and listening matter. In *AL Holland & RJ Elman, Neurogenic Communication Disorders and the Life Participation Approach* (pp. 105–130). Plural Publishing.

van der Meulen, I., van de Sandt-Koenderman, W. M., Duivenvoorden, H. J., & Ribbers, G. M. (2010). Measuring verbal and non-verbal communication in aphasia: Reliability, validity, and sensitivity to change of The Scenario Test. *International Journal of Language and Communication Disorders*, *45* (4), 424–435. https://doi.org/10.3109/13682820903111952

Wallace, S. J., Worrall, L., Rose, T. A., Alyahya, R. S. W., Babbitt, E., Beeke, S., de Beer, C., Bose, A., Bowen, A., Brady, M. C., Breitenstein, C., Bruehl, S., Bryant, L., Cheng, B. B. Y., Cherney, L. R., Conroy, P., Copland, D. A., Croteau, C., Cruice, M., & Dorze, L. (2023). Measuring communication as a core outcome in aphasia trials: Results of the ROMA-2 international core outcome set development meeting. *International Journal of Language & Communication Disorders*, *58*, 1017–1028. https://doi.org/10.1111/1460-6984.12840

Wallace, S. J., Worrall, 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., & 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

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*, S70–S76. http://dx.doi.org/10.1016/j.apmr.2011.07.206