ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

Test–Retest Reliability of Microlinguistic Information Derived From Spoken Discourse in Persons With Chronic Aphasia

Brielle C. Stark,[a] Julianne M. Alexander,[a] Anne Hittson,[a] Ashleigh Doub,[b] Madison Igleheart,[a] Taylor Streander,[a] and Emily Jewell[a]

[a] Department of Speech, Language and Hearing Sciences, Indiana University Bloomington [b] Speech and Hearing Sciences, University of Illinois at Urbana–Champaign, Champaign

ARTICLE INFO

Article History:
Received May 12, 2022
Revision received October 17, 2022
Accepted March 16, 2023

Editor-in-Chief: Stephen M. Camarata
Editor: Sarah Elizabeth Wallace

https://doi.org/10.1044/2023_JSLHR-22-00266

ABSTRACT

Purpose: The purpose of this study was to characterize test–retest reliability of discourse measures across a battery of common tasks in individuals with aphasia and prospectively matched adults without brain damage.
Method: We collected spoken discourse during five monologue tasks at two timepoints (test and retest; within 2 weeks apart) in an aphasia group (n = 23) and a peer group with no brain damage (n = 24). We evaluated test–retest reliability for percentage of correct information units, correct information units per minute, mean length of utterance, verbs per utterance, noun/verb ratio, open/closed class word ratio, tokens, sample duration (seconds), propositional idea density, type–token ratio, and words per minute. We explored reliability's relationship with sample length and aphasia severity.
Results: Rater reliability was excellent. Across tasks, both groups demonstrated discourse measures with poor, moderate, and good reliability, with the aphasia group having measures demonstrating excellent test–retest reliability. When evaluating measures within each task, test–retest reliability again ranged from poor to excellent for both groups. Across groups and task, measures that appeared most reliable appeared to reflect lexical, informativeness, or fluency information. Sample length and aphasia severity impacted reliability, and this differed across and by task.
Conclusions: We identified several discourse measures that were reliable across and within tasks. Test–retest statistics are intimately linked to the specific sample, emphasizing the importance of multiple baseline studies. Task itself should be considered an important variable, and it should not be assumed that discourse measures found to be reliable across several tasks (averaged) are likewise reliable for a single task.
Supplemental Material: https://doi.org/10.23641/asha.23298032

Spoken discourse, which we define as language beyond a single simple clause used for a specific purpose (Armstrong, 2000), characterizes our verbal communication as humans and is part of the larger umbrella, "connected speech." Discourse reflects natural language production (in the form of monologue or conversation), requires the complex interaction of language and cognitive processes to be successful, and is commonly impaired in poststroke aphasia. Eliciting and analyzing discourse in aphasia is important for researchers and clinicians alike because doing so enables a comprehensive view of an individual's language and communication abilities. For example, discourse is an efficient means of evaluating language structure (e.g., phonology) and use (e.g., topic management, topic appropriateness), enabling researchers and clinicians to evaluate a variety of critical language processes at once. Furthermore, discourse may be particularly sensitive to picking up on subtle, though important, language weaknesses and strengths in individuals with mildest

Correspondence to Brielle C. Stark: bcstark@iu.edu. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

aphasia (Fromm et al., 2017). Clinically, this is important because analysis of discourse may be one way to advocate for continued services to the population of mildest aphasia. For example, individuals with mildest aphasia may seek to return to work and will require strong (often specialized) language skills, which discourse analysis can identify and discourse-related treatment can remediate.

Advocates for spoken discourse assessment in aphasia have called for the standardization and validation of discourse outcomes for eventual inclusion into a core outcome set (Dietz & Boyle, 2018), which is an agreed-upon minimum set of outcomes (e.g., measurements) aiming to alleviate some of the variance in the presently very heterogeneous measurement of outcomes in poststroke aphasia treatment research (Wallace et al., 2018). This is a valid concern, as a recent review cited 165 studies evaluating spoken discourse in aphasia, where the outcome measures were highly heterogeneous (536 unique outcomes discussed; Bryant et al., 2016). A major issue with including discourse into a core outcome set for aphasia research is because of impoverished psychometric data (e.g., validity, reliability) regarding discourse outcomes, and this was the issue cited as the reason for discourse's exclusion from the core outcome set (Wallace et al., 2018). Evaluating psychometric properties of discourse-extracted measures is therefore a priority, as the present state of psychometric evidence in discourse is lacking: A recent analysis of informational measures in 76 studies of discourse in aphasia showed that no study reported acceptability data (e.g., distribution of scores, missing data) and that test–retest reliability was only reported in eight studies (Pritchard et al., 2017).

Reliability and validity of a measure, like data extracted from discourse, speak to its quality. Validity is highly related to reliability, in that something that is reliable, or consistent, over time may not be an accurate or meaningfully representative of the construct of interest. For the purposes of this article, we proceed with the view that discourse has well-accepted ecological validity, in that its face value representation of everyday communication is high. This evidence comes from surveys of experts in the field regarding discourse as a valued proxy for overall communication (Bryant et al., 2017; Cruice et al., 2020; Stark, Dutta, Murray, Fromm, et al., 2021) and from individuals with aphasia, citing discourse and, more generally, conversation as a priority for improvement after stroke (Simmons-Mackie et al., 2017; Worrall et al., 2011). Recent studies evaluating psychometric properties have been promising, suggesting strong validity across discourse-derived outcomes (Bryant et al., 2016; Pritchard et al., 2018). Validity is a between-subjects construct that does not explain within-subject performance variability and is thus complemented by investigations of reliability. For the purposes of this article, we will focus on test–retest reliability of discourse-derived measures in aphasia.

According to classical test theory, there is a "true score" (not "true" in the sense of only one score being correct, but "true" in the sense of accurately representing the score over repeated testing or timepoints), and any deviation around the true score is the result of an error. As such, any increase or decrease in variance of the "true score" can be understood to accompany a corresponding decrease or increase in error variance (Traub, 2005). A reliable measure, therefore, is one with a small error, such that the deviations around the true score are small. Most reliability measures describe this on a scale of 0 to 1, where a reliability of 1 indicates that all variability is attributable to true differences and there is no measurement error, whereas a reliability of 0 means that all variability is due to measurement error (Matheson, 2019).

Potential threats to reliability come from a variety of sources (e.g., inconsistent behavior of participants, different interpretations by participants, different interpretations by raters, practice effects due to repeated administrations) and are considered contributors to error. It is also well understood that a measure demonstrated as reliable in one context (e.g., for one sample, for one measurement) may not be reliable in a different context; that is, reliability of a measure is intimately related to the inter-individual differences in that sample. In psychometrics, reliability is often assessed using internal consistency, which involves examining the similarity of the responses between individual items or scores from some scale or test and comparing this similarity to the total variability in scores within the sample (Ferketich, 1990). Internal consistency is a useful measure because the reliability of the scale/test is able to be estimated using data from a single completion of that test by each participant. However, internal consistency evaluation for discourse data is unrealistic, given that the measures extracted from discourse, such as total words or words per minute (WPM), cannot be broken into small representative parts, unlike a test with many questions. For a field that is particularly interested in assessing and identifying change with time and intervention, test–retest reliability is particularly valuable. Test–retest reliability reflects the reproducibility of the "true" score upon repeated administrations over a period of time when the respondent's condition do not change, for example, when no intervention is taking place. A highly reliable score, then, is one where the error is low over repeated administrations. A highly reliable score is more sensitive to showing intervention-related change because any deviation from the "true" score is more than likely due to a "real" intervention change and not some unexpected increase in error. Notably, reliability increases as a function of the number of observations (i.e., Spearman–Brown prophecy formula), so testing and retesting is an ideal way to explore reliability of a

discourse-derived measure, with more retests increasing reliability estimates. Test–retest reliability is particularly critical to evaluate in aphasia, given that aphasia is a population noted for its high intragroup variation in language ability and impairments (Herbert et al., 2008; Hula & McNeil, 2008).

There is a paucity of research evaluating test–retest reliability of discourse-derived measures in aphasia. Perhaps this paucity is unsurprising, given the considerable amount of time it takes to transcribe and analyze discourse (Bryant et al., 2017; Stark, Dutta, Murray, Fromm, et al., 2021), as well as the considerable cost and time barriers that arise as a result of multiple baseline designs. The earliest study on test–retest reliability of discourse-derived measures evaluated percentage of correct information units (%CIUs) and WPM across a set of 10 stimuli for 20 adults with aphasia and 20 adults with no brain damage, finding that longer sample sizes (of at least 300–400 words for individuals with aphasia) tended to associate with higher test–retest reliability CIUs and WPM but that reliability estimates and their relationship with sample size varied across individuals (Brookshire & Nicholas, 1994). Boyle (2015) examined reliability (2–7 days between testing sessions) of paraphasia production per minute in 10 individuals with aphasia during picture-description and story retell narrative prompts (Boyle, 2015). Intraclass correlation coefficients (ICCs), a measure from 0 to 1 quantifying reliability, were calculated for combined prompts. When averaging discourse measures across all tasks, only semantically related paraphasias per minute were found to be highly reliable (which Boyle defined as ICC > .9), meaning that the "true" score's error at test/retest was relatively low. Meanwhile, phonemic paraphasias, time fillers, and repetitions per minute showed adequate reliability (ICC > .7), and false starts per minute demonstrated poor reliability (ICC < .40). Another test–retest investigation was done by Boyle (2014), who examined test–retest reliability of word retrieval in 12 individuals with aphasia across three sessions (2–7 days apart) using data averaged across five discourse prompts and Pearson correlation rather than ICC (Boyle, 2014). Although several of the discourse measures demonstrated acceptable reliability ($r > .7$; e.g., CIUs; lexical diversity), she argued that few were sufficiently stable for making clinical decisions about individuals on the basis of a single administration ($r > .9$; e.g., WPM), and some demonstrated poor reliability across sessions (e.g., inaccurate main concepts; Boyle, 2014). Finally, Cameron et al. (2010) examined test–retest reliability in 11 individuals with aphasia, demonstrating considerable variability in CIUs across administration times (on average, 7 days apart, with the group range between test and retest being 1–42 days; Cameron et al., 2010). In summary, prior research suggests that at least some discourse-derived outcomes in aphasia may be reliable at the group level and within participants, but the studies suffer from limited sample sizes and lack of comparison with controls, the latter of which makes it hard to determine the extent to which variability in discourse is restricted to aphasia.

Additionally, it is unclear which subject factors (e.g., demographic, cognitive–linguistic) and task factors (e.g., genre of task, sample length) might contribute to variability of discourse performance at the single-subject level, given task effects on discourse in aphasia (Fergadiotis & Wright, 2011; Shadden et al., 1991; Stark, 2019; Stark & Fukuyama, 2021; Ulatowska et al., 1981), and known heterogeneity of language impairments in aphasia. We are particularly interested in evaluating the impact of discourse task on variability because of two recent publications from our lab, where we evaluated speech produced across different structured discourse prompts in a very large sample size of individuals with aphasia and controls (Stark, 2019; Stark & Fukuyama, 2021). Briefly, there exist discourse genres and, within those, specific tasks. Some common genres of monologic discourse that are collected include expositional, descriptive, narrative, and procedural. Within these genres are included many different tasks. Common tasks in expositional genres include pictures or picture sequences, and instructions ask participants to extrapolate on the events, rather than just describe them. Describing pictures, as well as picture sequences, falls under the descriptive genre. In narrative genres, participants are asked to expound upon personal or fictional stories. In the procedural genre, participants are asked to tell an experimenter how to do something, for example, how to make a sandwich. In both studies (using different statistical analyses), we showed that, despite impoverished output (i.e., individuals with aphasia produced ~50% less overall output than controls), language was significantly different across discourse tasks. For example, a procedural task ("Tell me how to make a peanut butter and jelly sandwich") led to the production of significantly less syntactically complex language than a narrative task ("Cinderella story") in both the aphasia and control groups. This result suggests two things: that the type of discourse task being selected has a bearing on the language being produced and that acquiring data from more than one discourse prompt is likely the most sensitive means of capturing the breadth and depth of language ability. However, whereas the latter seems the best option, time is limited in both research and clinical settings. Indeed, surveys by our group (Stark, Dutta, Murray, Fromm, et al., 2021) and other groups (Bryant et al., 2017; Cruice et al., 2020) consistently indicate that time is a barrier to discourse analysis. While discourse sampling is quick, the backend work (i.e. transcription, coding, analysis) takes, on

average, 5–12 min per 1 min of discourse sample (Boles, 1998), which poses a feasibility problem. Boyle (2015) evaluated word retrieval errors per minute (specifically, phonological errors, semantically related errors, false starts, and time fillers) across four total tasks, which were separated into three genres: sequence-picture narratives (two tasks), complex-picture narrative (one task), and story retell narrative (one task). For phonological errors per minute, ICC was highly reliable for the complex-picture sequence (ICC > .9) but moderate to good for story retell narrative (ICC = .56) and sequence-picture narratives (ICC = .64). This pattern was different for semantically related errors, which were most reliable during the story retell narrative (ICC = .94) but not complex-picture narrative (ICC = .47) or sequence-picture narratives (ICC = .52). False starts per minute were highly unreliable no matter the genre (all ICC < .22), and this was similar for time fillers per minute (all ICC < .44). This study is an ideal example of how reliability is highly task dependent and emphasizes the need to explore this in more detail. Furthermore, while Boyle (2015) describes that the tasks vary in sample length (which she defines in minutes, with the complex-picture narratives eliciting the shortest samples), she does not empirically evaluate the extent to which test–retest reliability of the discourse measures varies by task as well as by sample length.

Lastly, it is important to consider other components of reliability, such as inter- and intrarater reliability. Overwhelmingly, discourse analysis in aphasia relies on manual transcription and coding. Despite this, studies inconsistently report rater reliability and, when they do, they use unstandardized measures such as percentage agreement. Unfortunately, percent agreement was calculated differently across studies. Unlike ICC, where there are standard interpretations (e.g., ICC values of > .70 are generally considered as good and > .90 as excellent), there is no standard criterion for evaluating percentage agreement. In an analysis of information measures reported in discourse analysis in aphasia, intrarater reliability was cited in nine studies, and, in every study, rater reliability was quantified using percent agreement (Pritchard et al., 2017). Furthermore, few, if any, studies report the training protocol used nor demographic information about the raters themselves. Here, we report rater information and openly share our training protocol and data in Open Science Framework (OSF) in an effort to work toward the establishment of best practices, which we hope will greatly improve reproducibility of studies in this area.

Finally, in the spirit of this special issue, we acknowledge that best practices for data analysis and sharing in other disciplines, like neuroimaging, have been established to ensure replicability of studies (Nichols et al., 2017) and have recently been proposed for discourse analysis in aphasia (Stark et al., 2022), which guides the research presented here. Refer to Supplemental Table S1 for the best practice checklist for reporting on discourse data in aphasia.

This Study

Specifically, this article evaluates the following:

1. test–retest reliability for individuals with aphasia and adults without brain damage, across monologue discourse tasks (i.e., measures are averaged across tasks) and
2. test–retest reliability for individuals with aphasia and adults without brain damage, for each monologue discourse task.

Altogether, the proposed project has the potential to advance the discipline of communication sciences and disorders by evaluating test–retest reliability of spoken discourse measures in aphasia and by improving the collection and analysis of discourse in aphasia through transparent reporting of methods and opensource sharing of materials. The clinical implication of this project is straightforward: This study will provide data on test–retest and rater reliability of common discourse measures in aphasia, across a battery of commonly used tasks, thus providing a foundation for clinicians and researchers to identify reliable and sensitive outcome measures for use in assessment and treatment.

Method

We elaborate in greater detail on the methodological details of this study in the work of Doub et al. (2021), a technical report that we wrote to aid in research using virtual and remote designs to collect spoken discourse data from this clinical population. We received ethical approval to conduct this research (IRB #1904590484 at Indiana University). We used the best practices guidelines for reporting spoken discourse research in aphasia (Stark et al., 2022) to ensure replicability and reproducibility of our study (guidelines are also available here: https://osf.io/y48n9/; see Supplemental Table S1), and we have preregistered our hypotheses using the OSF (https://osf.io/y9qsc). R Markdown including tables, figures, and full code is available in the Files section of our OSF repository. Video and coded transcripts will be uploaded to AphasiaBank (http://aphasia.talkbank.org; MacWhinney et al., 2011) upon publication of study results.

Participants

We are interested in comparing test–retest reliability for discourse measures across the two groups (one group

with aphasia and one group without aphasia), to explore the question of whether language performance in individuals with aphasia is more variable across time and contexts. The present sample was identified based on a power analysis completed on a pilot sample of short interval test–retest reliability of discourse-derived microlinguistic variables in $n = 7$ individuals with aphasia ($M = 7.29 \pm 4.68$ days between testing) and $n = 9$ speakers with no brain damage ($M = 6.11 \pm 2.71$ days). These data were from AphasiaBank (MacWhinney et al., 2011). In the pilot sample, we evaluated test–retest measures (e.g., effect size, systematic difference, standard error of measurement) for several linguistic measures similar to this study (e.g., total tokens, mean length of utterance [MLU] in words, open/closed word class ratio). Given our pilot sample, we computed sample size using a power analysis assuming a repeated-measures analysis of variance with within-participant measures and interaction with between-participants measures assuming a small effect size (η_p^2 of .02), power of 0.80, two same-size groups, $\alpha = .05$, two measurement timepoints, and a correlation of .8 between outcome measures at test and retest (Faul et al., 2007). The sample size estimation was 42. We collected test–retest spoken discourse data from $n = 25$ persons with aphasia and $n = 24$ prospectively age- and education-matched adults without brain injury (NBD group).

Recruitment

Subject recruitment was conducted virtually (with one exception of in-person recruitment prior to onset of the COVID-19 pandemic), and potentially eligible participants were screened using an online survey hosted on REDCap (Harris et al., 2009, 2019). Our inclusion and exclusion parameters for the NBD group were to be native English speakers, 45–80 years of age with at least 10 years of education, and without a history of brain injury or neurological or developmental language disorder. Inclusion and exclusion parameters for individuals with aphasia were to be native English speakers, 18 years and older, with a diagnosis of aphasia because of an acquired brain injury that was at least 6 months prior to entrance into the study, and without any other neurological disorder or neurodegenerative disease.

If a subject was deemed eligible, a non-identifiable, unique ID was generated, which is how the subject was identified throughout the rest of the study. Informed, verbal consent was recorded using web conferencing software, and further neuropsychological tests were administered to verify eligibility. In the case of the NBD group, this included the Montreal Cognitive Assessment (Nasreddine et al., 2005), where the cutoff score of 26 was used to rule out individuals who may be experiencing cognitive decline. For our purposes, we did not want to introduce an additional variable of cognitive decline into the research. In the aphasia group, we collected the Bedside version of the Western Aphasia Battery–Revised (Kertesz, 2007) to characterize aphasia type and severity, but there was not a cutoff for inclusion. That is, if a potential subject indicated that they had received an aphasia diagnosis after an acquired brain injury but tested as clinically non-aphasic on the test, we still included them in our study. This is due to burgeoning research about latent language impairments despite scoring as clinically non-aphasic on this battery (Fromm et al., 2017). For all participants, we also collected a more detailed biographical intake form to verify that no exclusion parameters were present, for example, prior brain injuries in the NBD group or progressive neurological disorders in the aphasia group. The Apraxia Battery for Adults (Dabul, 2000) was administered at retest and only to the aphasia group. The purpose of this test was to rule out severe motor speech disorders, that is, apraxia of speech and dysarthria. We did not complete subtests related to oral apraxia or limb apraxia. If severe apraxia of speech or dysarthria was noted, data were subsequently excluded from analysis (note: no participant from the aphasia group was excluded due to this).

At the first testing session, where we obtained informed consent and verified inclusion/exclusion parameters, participants also scheduled their second session. We feel that this helped us achieve high retention, which was 100%—that is, all participants who participated in Day 1 (Test) also participated in Day 2 (Retest).

Data Collection

Virtual Methodology

We collected 2 days of data, aiming to space these 10 ± 3 days apart (see Doub et al., 2021, for full protocol). The first testing day will be referred to as Test and the second testing day as Retest. On average, participants were tested 7.79 ± 1.72 days apart.

Discourse Tasks

The main purpose of our study was to collect discourse using common elicitation protocols. We therefore used the AphasiaBank protocol (MacWhinney et al., 2011) to do so, given the prevalence of the protocol's use in aphasia research as well as the typicality of the types of discourse tasks collected (e.g., picture description being the most common assessed clinically and in research; Bryant et al., 2016). This protocol contains the following discourse assessments: (a) retelling of a personal narrative ("Tell me a story about something important that happened to you. . .."), (b) retelling of a sickness narrative (NBD group: "Tell me about a time you were ill or

injured"; aphasia group: "Tell me about your stroke and recovery"), (c) picture description (one task; cat rescue), (d) picture sequence description (two tasks; rescued umbrella and broken window), (e) fictional story retell (one task; Cinderella), and a (f) procedural narrative (one task; sandwich, "Tell me how to make a peanut butter and jelly sandwich"). For the purposes of this article, we do not analyze the personal or sickness narratives. We focus analysis on the five remaining tasks, which have easily identifiable targets of speech (i.e., we know the intended word targets, e.g., the Cinderella story, images in the picture provided).

Participants were asked to respond to the discourse prompts at both Test and Retest, in the same order both days with few exceptions (examiner error). The AphasiaBank instructions for eliciting these samples were used; it can be found in full at http://aphasia.talkbank.org. At Test, the instructions for discourse collection were as follows: "We are going to walk through several types of stories. You will tell me stories, describe some pictures, and tell me how to do something." We modified the instructions at Retest, intended to mitigate practice effects (e.g., shortening of discourse) but emphasizing to the participant to complete tasks as if it was their first time doing so: "[same instructions, then:] I want you to tell it to me as if you were telling the story to somebody for the first time."

Transcription and Attachment to Video Protocol

True blinding for transcription and analysis could not be employed, given the stark differences in language between our two participant groups (aphasia, NBD). To achieve as much blinding as possible, transcribers (authors M.I., E.J., and T.S.) were blinded to each participant's demographic and neuropsychological test scores. Transcribers used the video to transcribe all verbatim speech from experimenter and participant using orthographic transcription, and each transcript was time-locked to the video. They then manually checked their work, and author B.C.S. reviewed each transcript. The same transcriber was responsible for transcribing both Test and Retest timepoints from the same participant. All transcripts were created using the Codes for the Human Analysis of Transcripts (CHAT) coding language (MacWhinney, 2000), which is a special coding language where transcribers employ codes to characterize transcribed speech (e.g., assigning a paraphasia code to an incorrect word), and where the companion analysis software CLAN (Computerized Language ANalysis; MacWhinney, 2018) automatically tags morphological and grammatical markers of speech. Transcribers divided each transcript into utterances, which we defined based on communication units. A communication unit was defined as an utterance that cannot be further divided without the disappearance of its essential meaning, plus any subordinate clause that is part of the independent predication. At this juncture, transcribers also marked utterances for exclusion. Excluded utterances were largely commentary about the speaker's performance (e.g., "I'm not good at this"). Other CHAT codes manually employed by transcribers included marking of paraphasias (word errors) and their types, speech dysfluencies (nonlexical and lexical), gestures and non–speech sounds, repetitions, and retracings.

Dependent Variables

Our primary variables of interest were %CIUs and correct information units per minute (CIUs/min; Nicholas & Brookshire, 1993), which were coded according to Nicholas and Brookshire (1993). The rationale for including %CIUs and CIUs/min as primary variables of interest is that CIUs have long been thought to be reliable for evaluating aphasic discourse and are one of the few measures on which test–retest reliability has been computed, thus providing some level of replicability to be established across studies (Boyle, 2014, 2015; Brookshire & Nicholas, 1994). In summary, CIUs have also had more extensive psychometric evaluation than other discourse-derived outcomes, demonstrating their validity (Fergadiotis et al., 2019). Furthermore, CIUs are commonly used in aphasia treatment research (e.g., Boyle et al., 2022; Evans et al., 2020) and, as such, should be examined for test–retest reliability in order to attribute any score change to intervention and not to error.

For our secondary variables of interest, we explored linguistic measures that were automatically extracted from the transcripts and which have been widely used to characterize discourse in aphasia (Bryant et al., 2016). We used CLAN (version: February 10, 2022, Windows) to automatically extract nine variables. To do this, morphological and grammatical tiers were assigned to the transcript using the *mor* command in CLAN. Then, we used the EVAL command in CLAN to extract information about each measure from the discourse (*EVAL + t*PAR + u *.cha)*. The EVAL command automatically analyzes the transcript for specific variables that were previously identified as being useful for clinical and research purposes, defined in more detail in the CLAN manual (https://talkbank.org/manuals/CLAN.pdf; p. 131). The nine variables extracted from the EVAL output included MLU (defined in words), verbs per utterance, noun/verb ratio, open/closed class word ratio, tokens, speaking duration (in seconds), propositional idea density, type–token ratio (TTR), and WPM (Table 1 broadly categorizes these by the linguistic category that they are a proxy for, e.g., fluency).

MLU is a proxy for grammatical complexity as well as speech fluency, evaluating the MLU (in words) for

Table 1. Measures extracted from discourse.

Primary linguistic proxy	Discourse measures extracted
Lexical-semantic	Type–token ratio[b] **Tokens**[b] **Propositional idea density**[b] Verbs per utterance[b]
Fluency and efficiency	Correct information units per minute[a] Words per minute[b] **Speaking duration**[b] Mean length of utterance (words)[b]
Syntactic	**Mean length of utterance (words)**[b] **Noun/verb ratio**[b] **Open/closed class word ratio**[b] **Verbs per utterance**[b]
Informativeness	Percentage of correct information units[a]
Lexical diversity	**Type–token ratio**[b] Noun/verb ratio[b] Open/closed class word ratio[b] Propositional idea density[b]
Gross output	Tokens Speaking duration

Note. Bolded measures are those measures that fit more than one proxy, but the bolded version of the measure denotes the proxy that is most often used to describe the measure. CLAN = Computerized Language ANalysis.

[a]Primary variables of interest, hand-scored. [b]Secondary variables of interest, automatically scored using CLAN software.

utterances that contain intelligible words. Utterances containing unintelligible words (coded as xxx, yyy, or zzz in CLAN) are excluded from this calculation. Longer MLU would indicate greater speech fluency and complexity (e.g., Wilson et al., 2010), and is sensitive to the placement of utterance boundaries (hence why we demonstrate utterance-boundary rater reliability). Verbs per utterance is a proxy for verb retrieval and grammatical complexity (e.g., Thorne & Faroqi-Shah, 2016). To calculate this, CLAN evaluates the morphological tier, which assigns parts of speech to each transcript word, and calculates the average verbs (verbs, copulas, and past or present participles; does not include modals) per utterance occurring across the transcript. More verbs per utterance would indicate better verb retrieval. Noun/verb ratio is a proxy for grammatical complexity and calculates the total number of nouns divided by the total number of verbs (excluding auxiliaries and modals; e.g., Saffran et al., 1989; Thompson et al., 2013). A higher noun/verb ratio score likely indicates presence of an agrammatism. Open/closed class word ratio is a proxy for word retrieval and grammatical complexity and calculates the ratio of total open class words (a category of words that readily admits new members, e.g., verbs, adjectives, adverbs) to total closed class words (a category of words that does not readily admit new members that often serve grammatical functions, e.g., prepositions, pronouns; e.g., Thompson et al., 2013). Higher open/closed class word ratio suggests impaired closed class production, often reflective of a grammatical and/or word finding issue. Tokens are total words excluding repetitions and revisions and excluding words for which no target was given (i.e., a paraphasia with unknown target), with greater tokens indicative of greater word finding ability and speech fluency. Speaking duration (measured in seconds) is calculated based on the total duration of speaking time (easily calculated because our transcripts were linked to the video). A greater speaking duration is a proxy that is more complicated to interpret, given that longer speaking time could indicate greater or impaired fluency. Propositional idea density was adapted with permission from Computerized Propositional Idea Density Rater, third major version (Brown et al., 2007), and is a measure calculated by dividing the number of verbs, adjectives, adverbs, prepositions, and conjunctions (i.e., propositional words) by the total number of words. It is a proxy for propositionality of speech, measuring the extent to which the speaker is making assertions (or asking questions) rather than just referring to entities (Brown et al., 2008); a higher propositional idea density score would indicate more propositionality. TTR is the total number of unique words (types) divided by the total number of words (tokens), and is a crude measure of vocabulary variation, where a higher TTR would indicate a greater diversity of vocabulary (e.g., Fergadiotis & Wright, 2011). Finally, WPM is a proxy for speech fluency, where a higher WPM is indicative of greater speech fluency; this has ties to motor agility and may be reduced because of not only language impairment (i.e., aphasia) but also any concomitant motor speech disorders (e.g., apraxia of speech; e.g., Doyle et al., 2000).

Given that we collected discourse information across five tasks, the dependent variables were first extracted by task. To evaluate data across tasks, dependent variables were averaged across the five tasks.

Data Analysis

Preregistration and Planned Analyses (Primary, Exploratory)

We preregistered our study analyses using the OSF (our project is located here: https://osf.io/y9qsc) in December 2020. Note that this was a preregistration of proposed analyses and was done after data collection but prior to analyzing any data. In our preregistration, we had intended to do an analysis of covariance but, upon beginning data analysis, realized that we could do more refined analyses that were relevant for test–retest (e.g., systematic differences with paired tests, median splits, ICC). In the preregistration, we registered two hypotheses. The first hypothesis analyzed the differences in test–retest reliability of discourse measures between subject groups, across tasks

(1A) and for each task (1B). We evaluated this hypothesis in this article. We also registered a second hypothesis, which proposed evaluating the role of cognitive and linguistic factors in predicting test–retest reliability. Given the extensiveness of this article, the second hypothesis will be evaluated in a forthcoming paper. In this article, we briefly examine how test–retest is influenced by aphasia severity and presence, but the future paper will expand upon this using other neuropsychological data that we acquired (e.g., attention).

Quantifying Reliability

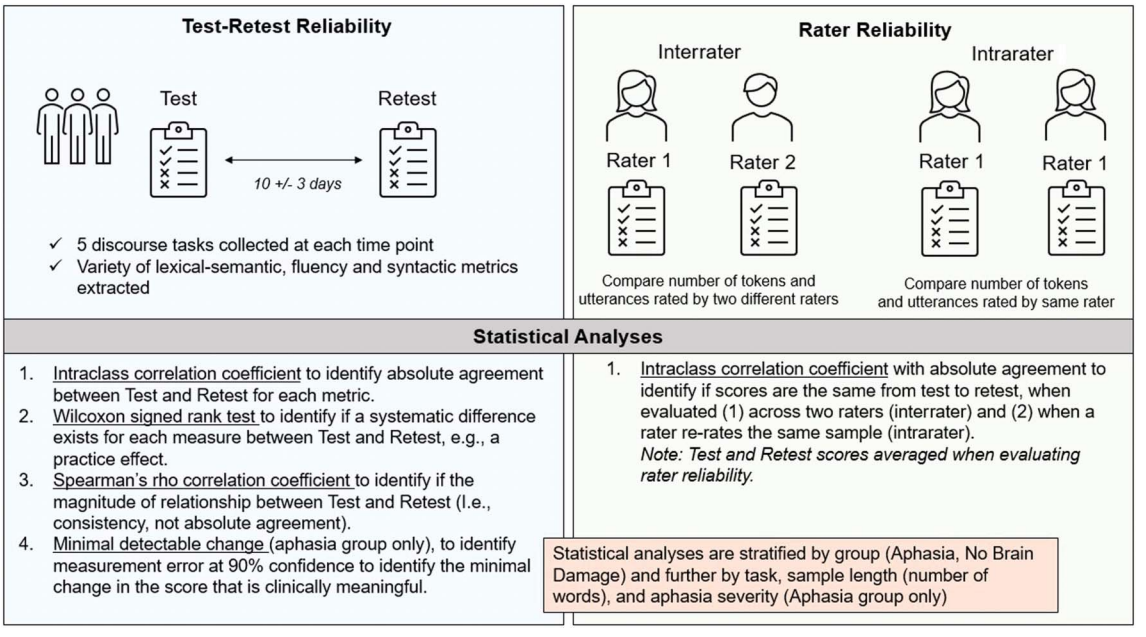
Reliability was quantified in three ways: test–retest, interrater, and intrarater reliability (see Figure 1 for schematic).

Interrater, intrarater, and test–retest reliability can be increased either by reducing the measurement error or by increasing the amount of true interindividual variability in the sample so that measurement error is proportionally smaller (Mathedson, 2019). Therefore, a high value (i.e., approaching "1") can reflect (a) low measurement error and/or (b) true interindividual, consistent, variability. A value of 0.5 could reflect equal proportions of true variability and measurement error.

Rater reliability. Raters comprised five research assistants in the NEURAL Research Lab: a PhD student with her CCC's in Speech Pathology (J.A.); a postbac student in Speech, Language, and Hearing (A.H.); and three MA students in Speech Pathology (M.I., E.J., and T.S.). A.H. and J.A. were the primary raters of the CIU data, and M.I., E.J., and T.S. were the primary coders for the CLAN-derived variables. Author B.C.S. trained raters using a shortened version of the CLAN manual (available in OSF files) and the original Nicholas and Brookshire article describing CIUs. Because we have a plethora of discourse data from individuals with aphasia collected by our lab, all raters were trained on samples outside of this study, achieving 80% agreement on coding and CIU distinction prior to coding this study's sample. Further discussion of CIUs specific to the samples of our study was done throughout the rating procedure, and our decisions on CIUs specific to the study are also available in OSF files.

Intrarater and interrater reliability for our dependent variables was computed using ICCs. Shrout and Fleiss (1979) suggest that two-way mixed-effects model is appropriate for testing inter- and intrarater reliability with multiple scores from the same rater, as it is not reasonable to generalize one rater's scores to a larger population of raters. We used the *IRR* package in R to compute two-way, absolute agreement analyses. In all cases, we use a "single"-type ICC. We used a single-type ICC to give a conservative estimate and because we were conceiving each rater as not necessarily representative, given our training and their expertise in speech-language sciences. We used the following cutoff values to interpret ICCs: excellent (ICC > .9), good (.75–.9), moderate (.5–.75), and poor (< .5; Koo & Li, 2016).

Figure 1. Schematic of reliability testing.



For interrater reliability on CIUs, nine participants were randomly selected (five from the aphasia group, four from the NBD group; 20% of the total sample transcripts; 18% of sample). Authors J.A. and A.H. independently determined CIUs for both test and retest from each of the nine participants. For intrarater reliability, authors J.A. and A.H. each rerated nine participants that they originally rated, 3–4 months after their original ratings. J.A.'s nine participants were made up of five participants from the aphasia group and four participants from the NBD group, and A.H.'s nine participants were made up of four participants from the aphasia group and five participants from the NBD group.

Across discourse measures directly extracted from the transcripts, we evaluated reliability for two critical baseline variables (total tokens, total utterances). These variables are used in calculations of dependent variables, and thus reliability for these two baseline variables is inherently important. Note that "total utterances" is a choice that is heavily dependent on the rater, as the CLAN system does not segment utterances automatically. High reliability on total utterances suggests high consistency in how utterances are distinguished at each timepoint. Tokens are calculated automatically from the transcript using the CLAN software, and reliability is reliant on accurate transcription. Participant samples selected to establish inter- and intrarater agreement were unique; that is, participant samples used for interrater agreement were not used for intrarater agreement. Participant samples were selected pseudorandomly, to ensure an equal proportion of participants from both groups (NBD and aphasia). Interrater reliability was conducted across three raters (E.J., M.I., and T.S.), for six participants (three with aphasia) for both days (Test and Retest; 12% of sample). Intrarater reliability was conducted across the same three raters for six participants (three with aphasia) for both days (12% of sample). Given the high reliability and small range of confidence intervals when examining rater reliability across tasks, we did not compute rater reliability by task.

Test–retest reliability. We completed the statistical analyses on dependent variables averaged across tasks (Research Question 1) and on dependent variables extracted from each task (Research Question 2):

1. To evaluate absolute agreement test–retest reliability of dependent linguistic variables, we used ICCs. As per recommendation for test–retest ICC, we used a two-way mixed-effects, absolute agreement, single rater/measurement that is optimal (the ICC(A,1); Koo & Li, 2016; McGraw & Wong, 1996). Koo and Li (2016) gave the following suggestion for interpreting ICC, including confidence intervals: below .50 = poor; between .50 and .75 = moderate; between .75 and .90 = good; and above .90 = excellent. Lin's concordance correlation coefficient was calculated in cases where ICC is poor, to identify if it improved the estimate. If it improved the estimate, it suggested that the low ICC was due to lack of spread (i.e., lack of true intragroup variability).
2. To measure the strength of association between two variables (i.e., consistency rather than absolute agreement), we computed Spearman rho on measures between Test and Retest. This is a complementary evaluation to ICC. One key difference between ICC and correlation is that, in the ICC, the data are centered and scaled using a pooled mean and standard deviation, whereas in the Spearman correlation, each variable is centered and scaled by its own mean and standard deviation. Correlations have a tendency to be more lenient than ICCs, in the way we have calculated them here.
3. To evaluate if there was a systematic difference between Test and Retest on our dependent variables (e.g., to identify significant changes between testing timepoints), we employed the Wilcoxon signed-ranks test (two-tailed, $\alpha = .05$), a nonparametric means of comparing two within-subject values to evaluate systematic bias. We chose the Wilcoxon signed-ranks test because evaluation of the data showed that, on the whole, data were not normally distributed (see Results section). Some data were normally distributed, but to maintain consistency, we used the Wilcoxon test throughout.
4. To demonstrate clinically meaningful change that would need to be measured at a follow-up session (after treatment), given baseline variability, we computed Minimal Detectable Change (MDC) at 90% confidence (Donoghue & Stokes, 2009). This was done only for the aphasia group. MDC90 was computed to ascertain approximate change needed to associate with a treatment effect, given variance from test–retest. MDC90 = 1.65 * *sqrt*(2) * *standard error of measurement*. Standard error of measurement, an estimate of how the repeated measure is distributed around the "true" score, is also given for both subject groups. We also calculated standard error of measurement for both groups, as this is a commonly calculated measure of reliability that complements ICC and Spearman rho.

Sample length (number of tokens elicited) and aphasia severity may influence test–retest reliability. For example, some participants produced very short samples for some tasks (which is a common issue in discourse tasks in aphasia), and short samples could lead to greater variability across tasks and sessions. Furthermore, several discourse measures are known to be influenced by sample

length (e.g., lexical diversity metrics, such as TTR, e.g., Fergadiotis et al., 2013), and it is important to assess the extent to which their test–retest reliability is likewise impacted by sample length. Thus, we stratified reliability by sample length and aphasia severity:

1. We computed a median split on tokens to identify a long and short sample length group for both subject groups (i.e., a median split was specific to each group). We chose a median split given the wide variation in sample length within and between groups, and thus this comparison would be an adequate statistical comparison for determining differences in reliability in longest versus shortest samples.
2. For aphasia severity, we used the Western Aphasia Battery (WAB) identified cutoff for mild aphasia (aphasia quotient [AQ] of 75), splitting the group into a "Mild or Latent" group (AQ > 75, n = 14) and "Moderate or Severe" group (AQ ≤ 75; n = 9). We chose to do this instead of a median split because the WAB standards are used often clinically, and therefore interpretation of differences between the two severity groups is more straightforward.
3. Because we had participants who did not test as having clinical aphasia (see Table 4), we explored how test–retest differed by clinical aphasia presence, comparing the group with clinical aphasia according to the WAB (AQ < 93.8; n = 17) with those who scored above the clinical aphasia cutoff (AQ > 93.8; n = 6).

We present results stratified by sample length and aphasia severity for data across tasks (Research Question 1) and by task (Research Question 2).

Analysis Software and Data Availability

All statistical analyses were computed using RStudio Version 1.4.1717, using R Version 4.2.1. The R Markdown document is available in our OSF project. De-identified data used in this study are also available on our OSF page and on AphasiaBank upon publication.

Results

Overview of Included Participants

While we had 100% retention (all participants showed up to both sessions), sometimes, a discourse task could not be obtained. In the aphasia group, n = 2 did not have data from at least four tasks (out of five tasks) per timepoint. Specifically: RC55 had data for all five tasks at Test, but no data at Retest (refusal) *and* RC47 had data for one task at Test and two tasks at Retest (refusal due to difficulty). We chose to remove these individuals from analyses, as there was not enough information at either Test or Retest to reliably impute their missing data. RC73, also a member of the aphasia group, was missing data for only a single timepoint for a single task (Cinderella at Test). For this reason, we imputed their data using the *imputeData* function from the *mclust* package in R. We then replaced the missing value in our data set so that RC73 had no missing data.

In the NBD group, we determined that we would exclude participants *by task* if they demonstrated a dependent variable that was > 3 *SD*s from group mean. No exclusions were necessary. Because the aphasia group had high heterogeneity, we did not further exclude any participants from this group based on standard deviations. This created a final data set of n = 23 in the aphasia group and n = 24 in the NBD group.

Distribution of Dependent Variables

To analyze normal distribution of data, we conducted Shapiro–Wilk tests (appropriate for small samples) on each dependent variable, computed as an average from test and retest, for each subject group. We considered dependent variables with p > .05 to be normally distributed (after further inspection of qq plots). For the primary variables, %CIU was not normally distributed for the NBD group (p < .001) but was normally distributed for the aphasia group (p = .059). CIUs per minute was normally distributed for both groups (p > .39). For tokens, data were normally distributed for the NBD group (p = .21) but not the aphasia group (p = .04). Speaking duration was normally distributed for both groups (p > .16), as was TTR (p > .21). MLUs was normally distributed for the NBD group (p = .34) but not the aphasia group (p = .02), and this pattern was similar for noun/verb ratio (NBD group, p = .19; aphasia group, p = .002). Open/closed class word ratio was not normally distributed for either group (p < .005). Propositional idea density was normally distributed for the NBD group (p = .89) but not for the aphasia group (p = .002), and this pattern was similar for verbs per utterance (NBD group, p = .053; aphasia group, p = .001) and WPM (NBD group, p = .45; aphasia group, p = .006). Because of the mixed normality shown across variables and groups, we opted for nonparametric statistics, as described in the Method section. Note that the AQ score (aphasia severity), which we used to explore how test–retest reliability varies by severity, was likewise not normally distributed (p = .006).

Rater Reliability of Dependent Variables

Reliability for raters was computed on data across tasks (i.e., all discourse data averaged across the five tasks

at Test and Retest; see Table 2). For %CIUs, interrater and intrarater agreement was excellent at Test and at Retest (ICC > .90). For the discourse measures extracted from the transcript, we evaluated total utterances and total tokens. Interrater reliability across the three raters was excellent at Test and Retest for both total utterances and for total tokens (ICC > .90). Intrarater reliability was on average excellent (ICC > .90), with the confidence interval containing "good" reliability, for total utterances at Test. Intrarater reliability at Test for total tokens was excellent (ICC > .90). Intrarater reliability at Retest was excellent for both total utterances and total tokens (ICC > .90).

Demographic and Linguistic Comparison Between Groups

Despite our efforts to prospectively match the NBD group to the aphasia group, the aphasia group was significantly older than the NBD group, and there were also more men in the aphasia group than the NBD group. Members of the aphasia group elected to take more days between testing sessions than the NBD group. There was no significant difference in education between the groups. We then evaluated dependent variables across tasks (i.e., averaged across tasks). Some dependent variables demonstrated a significant difference between groups when the variables were averaged across tasks. Descriptive statistics and group comparison statistics are shown in Table 3. More information on the aphasia group's neuropsychological data (i.e., WAB scores, aphasia types) can be found in Table 4. Figure 2's correlation matrices per group demonstrate the relatively strong correlation between most dependent variables for the aphasia group, compared with more limited correlation between dependent variables for the NBD group.

Research Focus 1: Assessing Test–Retest Reliability Across Tasks

For a summary table of test–retest reliability measures across task for ICC, see Table 5 as well as Figures 3, 4, and 5. Given the extensiveness of the results presented in Table 5, we chose to present a summary in the text and refer readers to exact statistics in Table 5.

Across tasks, both groups produced ICC values that ranged from poor to excellent. In general, the aphasia group tended to have discourse measures that were more reliable than the NBD group, having eight measures with confidence intervals containing "excellent" reliability standards compared to the NBD group's five measures. Seven measures had confidence intervals containing "poor" reliability for the NBD group and three for the aphasia group. The discourse measures that had the highest test–retest reliability fell broadly within lexical, informativeness, and fluency/efficiency proxies. Syntactic proxies had measures that were most commonly "poor" in reliability. Individual measures with the highest reliability regardless of subject group were %CIUs, tokens, TTR, and CIUs/min.

Identifying Systematic Differences and Magnitude of Relationship Between Test and Retest

Nearly all dependent variables, regardless of subject group, demonstrated a significant Spearman rho, suggesting high consistency of scores across timepoints for all proxies (see Table 5). An exception to this was propositional idea density in the NBD group. Regarding systematic differences in measures, only tokens was found to be significant after correcting for multiple comparisons, and this was true for both groups (where Retest tokens was higher than Test; see Table 5). Therefore, most variables did not appear to demonstrate a significant change between Test and Retest, reflecting minimal practice or repeated measure effects.

A Role for Sample Length and Aphasia Severity

Supplemental Table S2 stratifies reliability statistics (ICC, Spearman rho, and Wilcoxon signed-ranks test p value) by sample length (for both groups) and aphasia severity (two iterations [mild–latent vs. moderate–severe and latent vs. clinical aphasia] for aphasia group only)

Table 2. Rater reliability statistics.

Variable	Intrarater reliability	Interrater reliability
% Correct information units	Test, ICC = .97 [.923, .99] Retest, ICC = .998 [.995, .999]	Test, ICC = .965 [.71, .99] Retest, ICC = .979 [.91, .995]
Total utterances	Test, ICC = .961 [.75, .99] Retest, ICC = .997 [.98, 1]	Test, ICC = .997 [.98, 1] Retest, ICC = .994, [.95, .999]
Total tokens	Test, ICC = .999 [.996, 1] Retest, ICC = .997 [.98, 1]	Test, ICC = 1 [.999, 1] Retest, ICC = 1 [.999, 1]

Note. Parentheses show 95% confidence intervals around ICC. Koo and Li (2016) give the following suggestion for interpreting intraclass correlation coefficient (ICC), including confidence intervals: below .50 = poor; between .50 and .75 = moderate; between .75 and .90 = good; and above .90 = excellent. ICC = intraclass correlation coefficient.

Table 3. Demographic and dependent variable data across all five tasks.

Variable	NBD (*n* = 24)	Aphasia (*n* = 23)	Statistical test
Time between sessions (days)			
M (*SD*)	7.08 (0.504)	8.52 (2.19)	W = 153, p = .003*^ (aphasia > NBD)
Mdn [min, max]	7.00 [6.00, 8.00]	8.00 [7.00, 14.0]	
Age			
M (*SD*)	58.4 (8.74)	65.0 (9.57)	W = 153, p = .008*^ (Aphasia > NBD)
Mdn [min, max]	56.7 [45.1, 77.8]	65.9 [40.7, 79.8]	
Gender			
F	18 (75.0%)	6 (26.1%)	χ^2 = 9.37, p = .002*^ (aphasia > NBD for males)
M	6 (25.0%)	17 (73.9%)	
Race and ethnicity			
Hispanic or Latino	Yes (4.17%)	Yes (0)	Not calculated
	No (95.83%)	No (100%)	
White	24 (100%)	22 (95.65%)	
Black or African American	0	0 (1 excluded)	
Asian	0	0	
More than one race	0	1 (4.35%)	
Education	17.04 (3.16)	15.83 (3.05)	W = 343, p = .15
Brain injury etiology (may be more than one)	N/A	Stroke (100%) Brain bleed due to surgery (4.35%)	Not calculated
Years post-onset of injury	N/A	9.09 (9.73)	Not calculated
Handedness	Right (91.67%) Left (8.33%)	Right (95.65%) Left (4.35%)	Not calculated
Aphasia quotient			
M (*SD*)	Not collected	78.9 (20.1)	Not calculated
Mdn [min, max]		89.2 [30.8, 100]	
Aphasia severity			
Latent	Not collected	6 (26.1%)	Not calculated
Mild		8 (34.8%)	
Moderate		6 (26.1%)	
Severe		3 (13.0%)	
Very severe		0 (0%)	
Montreal Cognitive Assessment			
M (*SD*)	27.8 (1.28)	Not collected	Not calculated
Mdn [min, max]	28.0 [26.0, 30.0]		
Average lexical and informativeness measures			
Proportion of correct information units			
M (*SD*)	0.813 (0.074)	0.617 (0.193)	W = 489, p < .0001*^ (NBD > aphasia)
Mdn [min, max]	0.828 [0.527, 0.900]	0.679 [0.162, 0.883]	
Propositional idea density			
M (*SD*)	0.489 (0.0143)	0.425 (0.090)	
Mdn [min, max]	0.489 [0.454, 0.516]	0.454 [0.194, 0.515]	W = 443, p = .004*^ (NBD > aphasia)
Type–token ratio			
M (*SD*)	0.472 (0.057)	0.511 (0.107)	W = 221, p = .249
Mdn [min, max]	0.456 [0.375, 0.570]	0.497 [0.323, 0.735]	
Tokens			
M (*SD*)	222 (77.5)	129 (84.80)	W = 447, p = .0002* (NBD > aphasia)
Mdn [min, max]	219 [115, 426]	115 [13, 333]	

(*table continues*)

Table 3. *(Continued).*

Variable	NBD (*n* = 24)	Aphasia (*n* = 23)	Statistical test
Average fluency measures			
Correct information units per minute			
M (*SD*)	124 (23.50)	56.5 (34.4)	W = 525, $p < .0001$*^ (NBD > aphasia)
Mdn [min, max]	123 [73.6, 168]	55.5 [2.50, 120]	
Speaking duration (s)			
M (*SD*)	148 (24.20)	73.70 (37.10)	W = 220, $p = .24$
Mdn [min, max]	146 [82.90, 196]	76.1 [15, 138]	
Words per minute			
M (*SD*)	148 (24.20)	73.70 (37.10)	W = 534, $p < .0001$*^ (NBD > aphasia)
Mdn [min, max]	146 [82.90, 196]	76.10 [15, 138]	
Average syntactic measures			
Mean length of utterance (words)			
M (*SD*)	10.6 (1.49)	7.51 (2.95)	W = 467, $p < .0001$*^ (NBD > aphasia)
Mdn [min, max]	10.2 [7.64, 3.0]	8.72 [1.94, 12.30]	
Noun/verb ratio			
M (*SD*)	1.14 (0.142)	1.45 (0.862)	W = 234, $p = .381$
Mdn [min, max]	1.15 [0.926, 1.37]	1.21 [0.516, 3.64]	
Open/closed class word ratio			
M (*SD*)	0.979 (0.072)	1.02 (0.297)	W = 358, $p = .08$
Mdn [min, max]	0.966 [0.883, 1.22]	0.904 [0.747, 1.84]	
Verbs per utterance			
M (*SD*)	1.83 (0.249)	1.28 (0.578)	W = 442, $p = .0003$*^ (NBD > aphasia)
Mdn [min, max]	1.76 [1.48, 2.51]	1.52 [0.220, 1.91]	

Note. Statistical testing used Wilcoxon rank sum exact test (W = test statistic; *p* = *p* value) and chi-square test (χ^2 = test statistics; *p* = *p* value). F = female; M = male; N/A = not applicable.

*Significant at $p < .05$. ^Significant at $p < .0038$ (13 comparisons), corrected for multiple comparisons using Bonferroni. NBD = adult group without brain damage; aphasia quotient extracted for only aphasia group using Western Aphasia Battery–Revised (WAB-R) Bedside (Kertesz, 2007) version, with 0 being *no language* and 100 being *no aphasia*; aphasia severity determined using cutoffs from the WAB-R, where aphasia quotient < 25 = very severe, 25–50 = severe, 51–75 = moderate, 75–93.8 = mild, and > 93.8 = latent; Montreal Cognitive Assessment, where exclusion principles necessitated we exclude any with a score of < 26 (out of 30), only for the NBD group.

whereas Figure 4 presents a visual of test–retest reliability for primary dependent variables, stratified by sample length and aphasia severity (mild–latent vs. moderate–severe), across tasks. The other linguistic dependent variables stratified by sample length and aphasia severity, across tasks, are shown for the aphasia group in Figure 6.

Long and short sample groups tended to have similar measures of reliability across a majority of measures and subject groups. A notable exception for the NBD group was %CIUs, which had excellent reliability for the long sample group yet moderate reliability for the short sample group. Some notable exceptions for the aphasia group included tokens and speaking duration, which both showed excellent reliability in the short sample group yet moderate reliability in the long sample group. Verbs per utterance in the aphasia group demonstrated poor reliability for the long sample group and excellent reliability for the short sample group. An opposite pattern was found for noun/verb and open/closed class word ratios in the aphasia group, which had higher reliability statistics for long samples ("good" reliability) and moderate reliability for short samples. To summarize, sample length appeared to have some impact on test–retest reliability in both subject groups, more often in the aphasia group, and this occurred for a variety of language proxies (i.e., lexical and informativeness, fluency, syntactic).

We next evaluated differences in dependent variables in the aphasia group in two ways: by comparing reliability by aphasia severity, contrasting a mild or latent group with a moderate or severe group, and by clinical aphasia presence, comparing a latent group with a clinical aphasia group. For comparisons between mild or latent and moderate or severe groups, notable differences in reliability included speaking duration (moderate for mild or latent group, excellent for moderate or severe group), noun/verb ratio (good for mild or latent group, poor for moderate or severe group), and open/closed class word ratio (excellent for mild or latent group, poor for moderate or severe group). For comparisons between latent and clinical aphasia groups, notable differences in reliability were found for

Table 4. Characteristics of the aphasia group derived from the Western Aphasia Battery (n = 23).

Characteristics	Overall (n = 23)
Aphasia type	
Anomic	7 (30.4%)
Broca's	6 (26.1%)
Conduction	3 (13.0%)
Latent	6 (26.1%)
Transcortical motor	1 (4.3%)
Aphasia type, dichotomous	
Fluent	10 (43.5%)
Latent (not aphasic by WAB score)	6 (26.1%)
Nonfluent	7 (30.4%)
Aphasia severity (derived from aphasia quotient [AQ])	
Latent (AQ > 93.8)	6 (26.1%)
Mild (AQ 75–93.7)	8 (34.8%)
Moderate (AQ 50–74)	6 (26.1%)
Severe (AQ < 50)	3 (13.0%)
WAB Auditory Comprehension (max = 10)	
M (*SD*)	9.61 (0.656)
Mdn [min, max]	10.0 [8.00, 10.0]
WAB Object Naming (max = 10)	
M (*SD*)	8.07 (2.32)
Mdn [min, max]	9.00 [2.50, 10.0]
WAB Repetition (max = 10)	
M (*SD*)	7.33 (2.57)
Mdn [min, max]	8.50 [2.00, 10.0]
WAB Spontaneous Speech Content (max = 10)	
M (*SD*)	7.98 (2.89)
Mdn [min, max]	9.00 [1.00, 10.0]
WAB Spontaneous Speech Fluency (max = 10)	
M (*SD*)	6.83 (2.81)
Mdn [min, max]	8.00 [1.00, 10.0]

Note. WAB = Western Aphasia Battery–Revised, Bedside Version (Kertesz, 2007).

the majority of measures, except for WPM (both excellent), noun/verb ratio (both moderate), and CIUs/min (both excellent). Taken together, it appears that reliability of variables representative of all proxies (i.e., lexical and informativeness, fluency, syntactic, gross output) differed by presence of aphasia (though this may have been mitigated by sample size difference between clinical [n = 17] and latent [n = 6] groups) and somewhat by aphasia severity.

Measuring MDC

For the aphasia group only, we calculated MDC at 90% confidence for each measure across tasks (see Table 5). To illustrate how we would use this information, we will elaborate on %CIUs. If our aphasia group underwent some treatment after test–retest testing on %CIUs, each subject from the aphasia group would need to exceed a 10% change in %CIUs (across all tasks; see Table 5) at follow-up for us to be confident that the change was due to the treatment and not error.

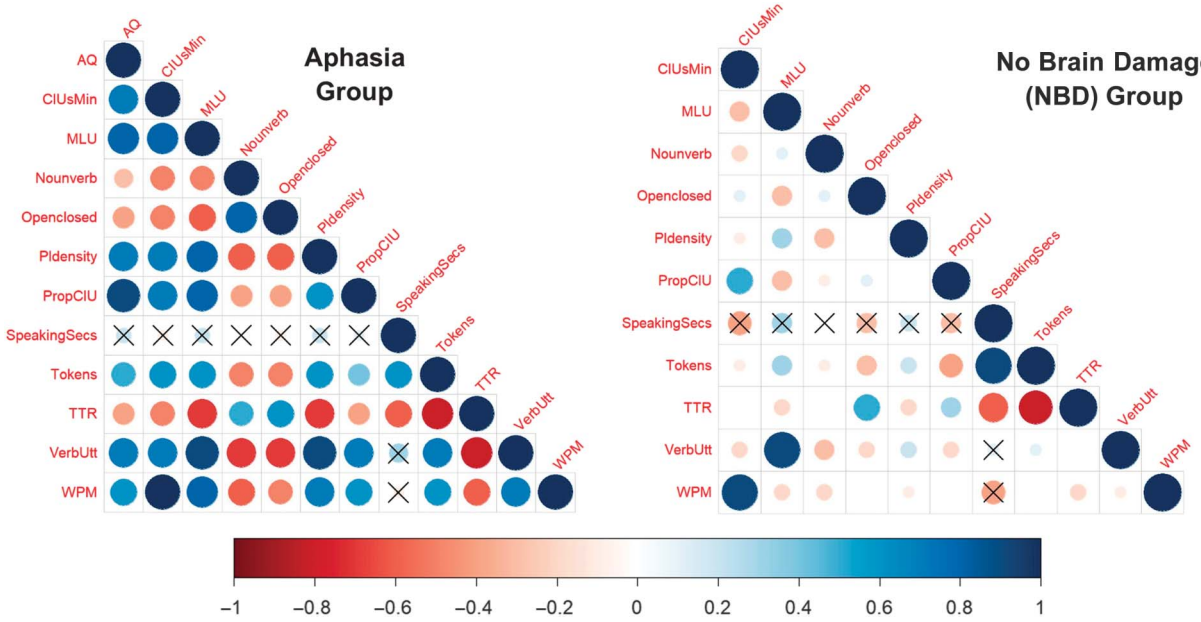
Research Focus 2: Assessing Test–Retest Reliability by Task

We evaluated test–retest reliability by task, the findings of which are summarized by ICC category in Table 6 and provided in more detail in Supplemental Table S2. For a summary table of raw data by task, see Supplemental Table S3. Test–retest relationships (i.e., ICC, systematic difference, correlation) by task and group are explored in Supplemental Tables S4–S8 and in Supplemental Figures S1–S5. Below, we briefly describe differences by task.

For the cat rescue (picture description) task, the ICC values for the NBD group ranged from poor to good: poor (noun/verb ratio, open/closed class word ratio, propositional idea density), moderate (%CIU, CIUs/min, MLU, verbs per utterance, WPM), and good (TTR, tokens, speaking duration). For the aphasia group, the ICC values ranged from moderate to excellent: moderate (noun/verb ratio, propositional idea density, open/closed class word ratio, CIUs/min), good (%CIU, speaking duration, TTR, verbs per utterance, tokens), and excellent (WPM, MLU). Across most measures and for both groups, there was high consistency (as measured by correlation). There was also only one measure that showed a significant systematic difference between Test and Retest (speaking duration in the aphasia group), but this was not significant when controlled for multiple comparisons.

For the Cinderella story (fictional story retell), the ICC values for the NBD group ranged from poor to good: poor (noun/verb ratio, open/closed class word ratio, propositional idea density), moderate (%CIU, CIUs/min, MLU, verbs per utterance, WPM), and good (tokens, speaking duration, TTR). For the aphasia group, the ICC values ranged from moderate to excellent: moderate (CIUs/min, noun/verb ratio, open/closed class word ratio, propositional idea density), good (%CIU, verbs per utterance, tokens, speaking duration, TTR), and excellent (MLU, WPM). Across most measures and for both groups, there was high consistency (as measured by correlation). There were several measures that showed a significant systematic difference between Test and Retest and which were significant when controlled for multiple comparisons. These included tokens, %CIUs, and WPM for the aphasia group (all increasing at Retest). There were several measures that showed a significant difference between Test and Retest for the NBD group, but no comparison survived multiple comparison correction.

Figure 2. Correlation matrices of dependent variables for aphasia group and NBD group, including covariates of interest (tokens, aphasia quotient [AQ] in aphasia group). Correlations that are crossed-out are those that are not significant, $p < .05$ (uncorrected). Size of circle is indicative of correlation magnitude; for example, larger circle is a greater correlation value (Pearson r here). CIUsmin = correct information units per minute; MLU = mean length of utterance (in words); Nounverb = noun-to-verb ratio; Openclosed = open-to-closed class word ratio; PIdensity = propositional idea density; PropCIU = ; SpeakingSecs = speaking duration in seconds; TTR = type–token ratio; VerbUtt = verbs per utterance; WPM = words per minute.



For the sandwich task (procedural narrative), the ICC values for the NBD group ranged from poor to good: poor (%CIU, CIUs/min, MLU, verbs per utterance, open/closed class word ratio, propositional idea density), moderate (noun/verb ratio, TTR, WPM, speaking duration), and good (tokens). For the aphasia group, the ICC values ranged from poor to excellent: poor (noun/verb ratio), moderate (%CIU, verbs per utterance, tokens, speaking duration, TTR), good (MLU, open/closed class word ratio, propositional idea density), and excellent (CIUs/min, WPM). Across most measures and for both groups, there was high consistency (as measured by correlation). There were no measures that showed a significant systematic difference between Test and Retest when controlled for multiple comparison, for either subject group.

For the broken window task (picture sequence exposition/description), the ICC values for the NBD group ranged from poor to good: poor (MLU, verbs per utterance, noun/verb ratio, open/closed class word ratio, propositional idea density), moderate (TTR, tokens, speaking duration), and good (%CIU, CIUs/min, WPM). For the aphasia group, the ICC values ranged from poor to excellent: poor (open/closed class word ratio), moderate (%CIU, noun/verb ratio, speaking duration, propositional idea density), good (CIUs/min, MLU, verbs per utterance, tokens, TTR), and excellent (WPM). Like the sandwich task, across most measures and for both groups, there was high consistency (as measured by correlation). There were no measures that showed a significant systematic difference between Test and Retest when controlled for multiple comparisons, for either subject group.

Finally, for the refused umbrella task (picture sequence exposition/description), the ICC values for the NBD group ranged from poor to good: poor (MLU, verbs per utterance, noun/verb ratio, open/closed class word ratio, propositional idea density), moderate (CIUs/min, TTR, WPM), and good (%CIU, tokens, speaking duration). For the aphasia group, the ICC values ranged from poor to excellent: poor (noun/verb ratio, open/closed class word ratio), moderate (tokens, speaking duration, TTR), good (%CIU, MLU, verbs per utterance, WPM), and excellent (CIUs/min, propositional idea density). Across most measures and for both groups, there was high consistency (as measured by correlation). Two measures showed a significant systematic difference between Test and Retest when controlled for multiple comparisons, which were tokens and speaking duration in the NBD group (both increasing at Retest).

To summarize, fluency measures seemed to be most reliable for both subject groups, followed by lexical and informativeness measures, and then by syntactic measures. The NBD group did not demonstrate a measure that

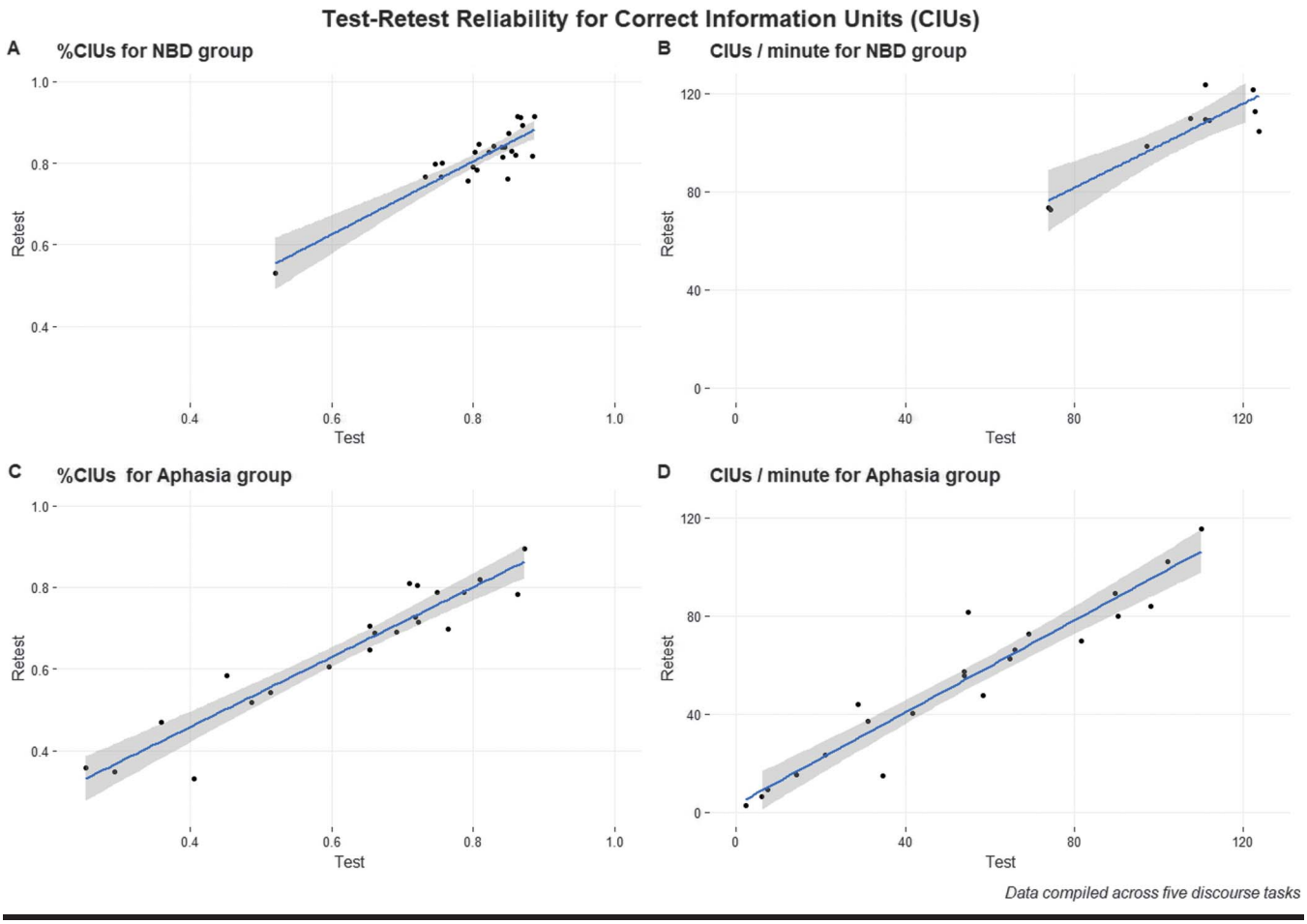
Table 5. Summary of test–retest results across tasks.

Primary Proxy	Measure	Group	ICC (CCC)	95% ICC CI (95% CCC CI)	Koo and Li (2016) ICC quality (CI quality)	Spearman rho (*p* value)	Systematic difference	*SEM*/MDC90
Lexical and informativeness	%CIU	NBD Aphasia	.89 .95	.76, .95 .86, .98	Good (good–excellent) Excellent (good–excellent)	.72 (.0001)*^ .92 ($p < .0001$)*^	V = 122, $p = .44$	0.03 0.04/0.10
	PI density	NBD Aphasia	.22 (.21) .94	–.21, .57 [–.18, .54] .86, .97	Poor (poor–moderate) *CCC remains poor* Excellent (good–excellent)	.27 (.21) .89 ($p < .0001$)*^	V = 137.5, $p = .073$ V = 71, $p = .04$*	0.02 0.02/0.05
	TTR	NBD Aphasia	.76 .91	.49, .90 .79, .96	Good (poor–good) Excellent (good–excellent)	.79 ($p < .0001$)*^ .91 ($p < .0001$)*^	V = 227, $p = .03$* V = 144, $p = .87$	0.03 0.03/0.08
	Tokens	NBD Aphasia	.82 .83	.50, .93 .53, .93	Good (moderate–excellent) Good (moderate–excellent)	.87 ($p < .0001$)*^ .92 ($p < .0001$)*^	V = 48, $p = .003$*^ V = 31, $p = .001$*^	34.35 38.94/90.87
Fluency/efficiency	CIUs/min	NBD Aphasia	.85 .96	.69, .93 .91, .98	Good (moderate–excellent) Excellent (excellent)	.83 ($p < .0001$)*^ .96 ($p < .0001$)*^	V = 170, $p = .58$ V = 115, $p = .50$	9.36 6.77/15.79
	SpeakingSecs	NBD Aphasia	.80 .75	.49, .92 .47, .89	Good (poor–excellent) Moderate (poor–good)	.83 ($p < .0001$)*^ .85 ($p < .0001$)*^	V = 57, $p = .007$* V = 65, $p = .03$*	16.996 26.63/62.14
	WPM	NBD Aphasia	.79 .97	.57, .90 .93, .99	Good (moderate–excellent) Excellent (excellent)	.76 ($p < .0001$)*^ .97 ($p < .0001$)*^	V = 177, $p = .46$ V = 83, $p = .10$	11.64 6.24/14.55
Syntactic	MLU	NBD Aphasia	.66 .94	.36, .83 .86, .97	Moderate (poor–good) Excellent (good–excellent)	.66 ($p = .0006$)*^ .80 ($p < .0001$)*^	V = 116, $p = .35$ V = 146, $p = .82$	0.96 0.73/1.70
	Noun/verb	NBD Aphasia	.38 (.37) .59	–.03, .68 [–.02, .67] .26, .80	Poor (poor–moderate) *CCC remains poor* Moderate (poor–good)	.43 ($p = .04$)* .79 ($p < .0001$)*^	V = 162 $p = .75$ V = 99, $p = .25$	0.13 0.61/1.43
	Open/closed	NBD Aphasia	.41 .70	.03, .69 .41, .86	Poor (poor–moderate) Moderate (poor–good)	.36 ($p = .09$) .87 ($p < .0001$)*^	V = 227, $p = .03$* V = 141, $p = .94$	0.07 0.17/0.41
	VerbUtt	NBD Aphasia	.56 .91	.21, .78 .79, .96	Moderate (poor–good) Excellent (good–excellent)	.57 ($p = .004$)*^ .75 ($p < .0001$)*^	V = 98, $p = .14$ V = 143, $p = .89$	0.19 0.18/0.42

Note. Koo and Li (2016) give the following suggestion for interpreting intraclass correlation coefficient (ICC), including confidence intervals: below .50 = poor; between .50 and .75 = moderate; between .75 and .90 = good; and above .90 = excellent. Lin's concordance correlation coefficient (CCC) is given in cases where ICC is poor, to identify if this improves the estimate. If it does improve the estimate, it suggests that test–retest of the low ICC is due to lack of spread (i.e., lack of true intragroup variability). Systematic difference estimated by Wilcoxon's signed-ranks test (statistic V) for paired data. CCC = Lin's Concordance Correlation Coefficient; CI = confidence interval; *SEM* = standard error of measurement (both subject groups); MDC90 = minimal detectable change at 90% confidence (aphasia group only); %CIU = percentage (proportion) of correct information units; NBD = group with no brain damage; PI density = propositional idea density; TTR = type–token ratio; CIUs/min = correct information units per minute; SpeakingSecs = speaking duration in seconds; WPM = words per minute; MLU = mean length of utterance (in words); VerbUtt = verbs per utterance; Noun/verb = noun-to-verb ratio; Open/closed = open-to-closed class word ratio.

*Significant. ^Significant after Bonferroni correction (11 row-wise within-group corrections; new $p < .0045$).

Figure 3. Test–retest reliability for percentage of correct information units (CIUs) and CIUs per minute across tasks, per subject group. The blue line represents the linear correlation, with the shading representing 95% confidence around the trend. The wider the light blue shade, the larger the variance. NBD = group with no brain damage.



achieved an excellent ICC for any task. Together with ICC and systematic difference calculations by task, these results suggest a high degree of intragroup variability in test–retest reliability by task and by group.

A Role for Sample Length and Aphasia Severity by Task

Supplemental Tables S9–S13 summarize the impact of sample length (for both groups) and aphasia severity and presence (aphasia group only) on ICC, Spearman rho, and Wilcoxon signed-ranks test for each task. Given that the results vary by task, we have elected to non-exhaustively contrast test–retest reliability of measures within and between subject groups. Two different tasks are elaborated on below, and we direct the reader to the supplemental tables for further statistics.

In the cat rescue task (picture description), CIUs/min showed high reliability for both subject groups, with little variation between long and short samples (NBD group, good reliability for both; aphasia group, excellent reliability for both). For the other primary variable, %CIUs, there was an impact of sample length of reliability, though it differed by subject group. For the aphasia group, long samples had excellent reliability for %CIUs, whereas short samples had moderate reliability. For the NBD group, long samples had moderate reliability for %CIUs, and short samples had poor %CIUs. In both cases, shorter samples had lower %CIU reliability. Other metrics' reliability appeared to also differ by sample length. For the NBD group, MLU and verbs per utterance showed higher reliability for shorter samples. For the aphasia group, propositional idea density had moderate reliability for long samples and excellent reliability for short samples, and this pattern (shorter samples having higher reliability) was the same for tokens, speaking duration, MLU, and verbs per utterance. The pattern was opposite for noun/verb ratio and open/closed class word ratio, where reliability was greater for short samples.

Figure 4. Visualizing test–retest reliability for primary dependent variables in the aphasia group, stratified by sample length and aphasia severity, across tasks. (A) Violin plots for low and high average tokens, for proportion CIUs (left) and CIUs/Min (right), with aphasia severity denoted by points of differing color and shape; (B) scatter plots between test and retest, arranged by aphasia severity (columns), with each individual datapoint labeled with aphasia quotient from the Western Aphasia Battery–Revised. Avg = average; CIUs = correct information units.

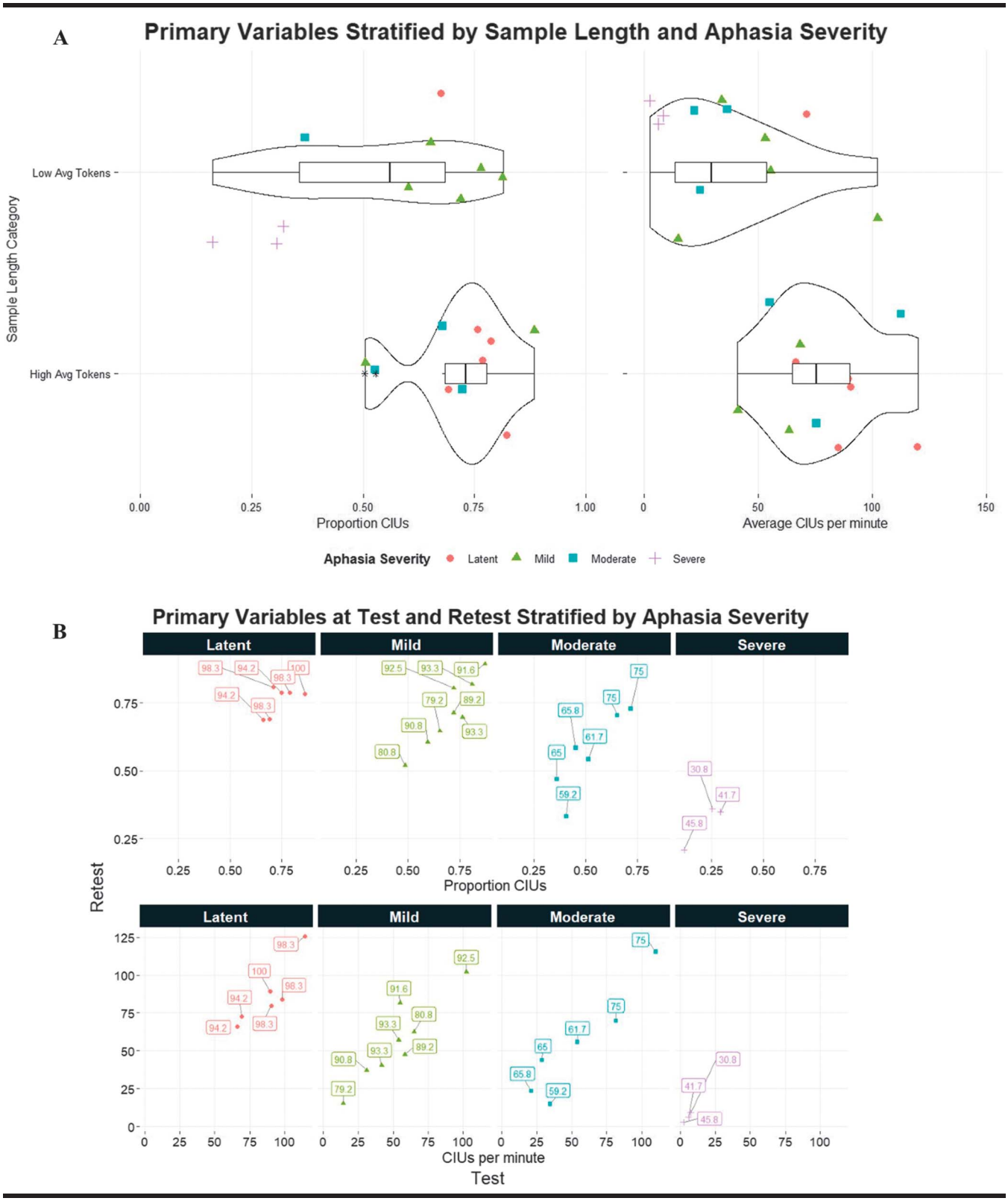
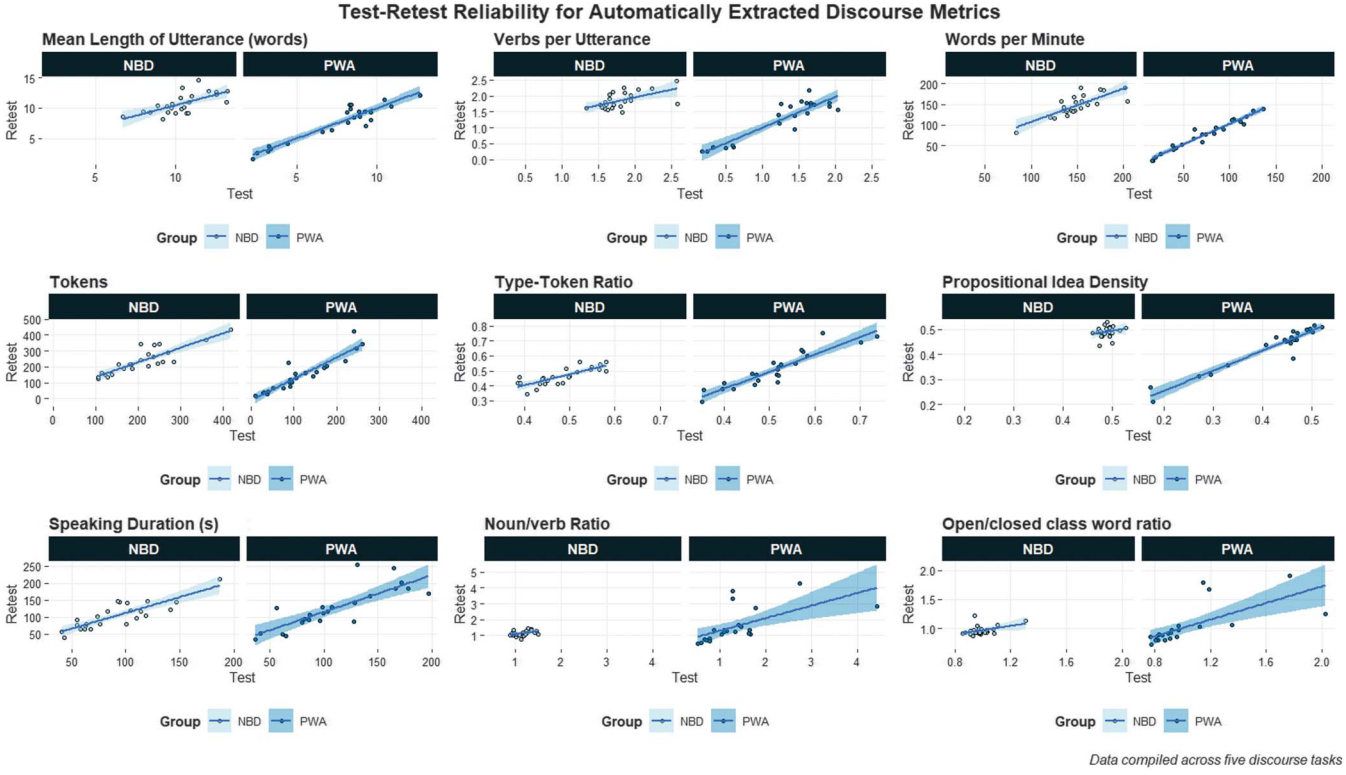


Figure 5. Test–retest reliability for microlinguistic variables derived from the discourse across tasks using CLAN, per subject group. Note that, due to each measure, the axes will have different scales. The blue line represents the linear correlation, with the shading representing 95% confidence around the trend. The wider the shade, the larger the variance. NBD = group with no brain damage; damaged group. CLAN = Computerized Language ANalysis; PWA = persons with aphasia.

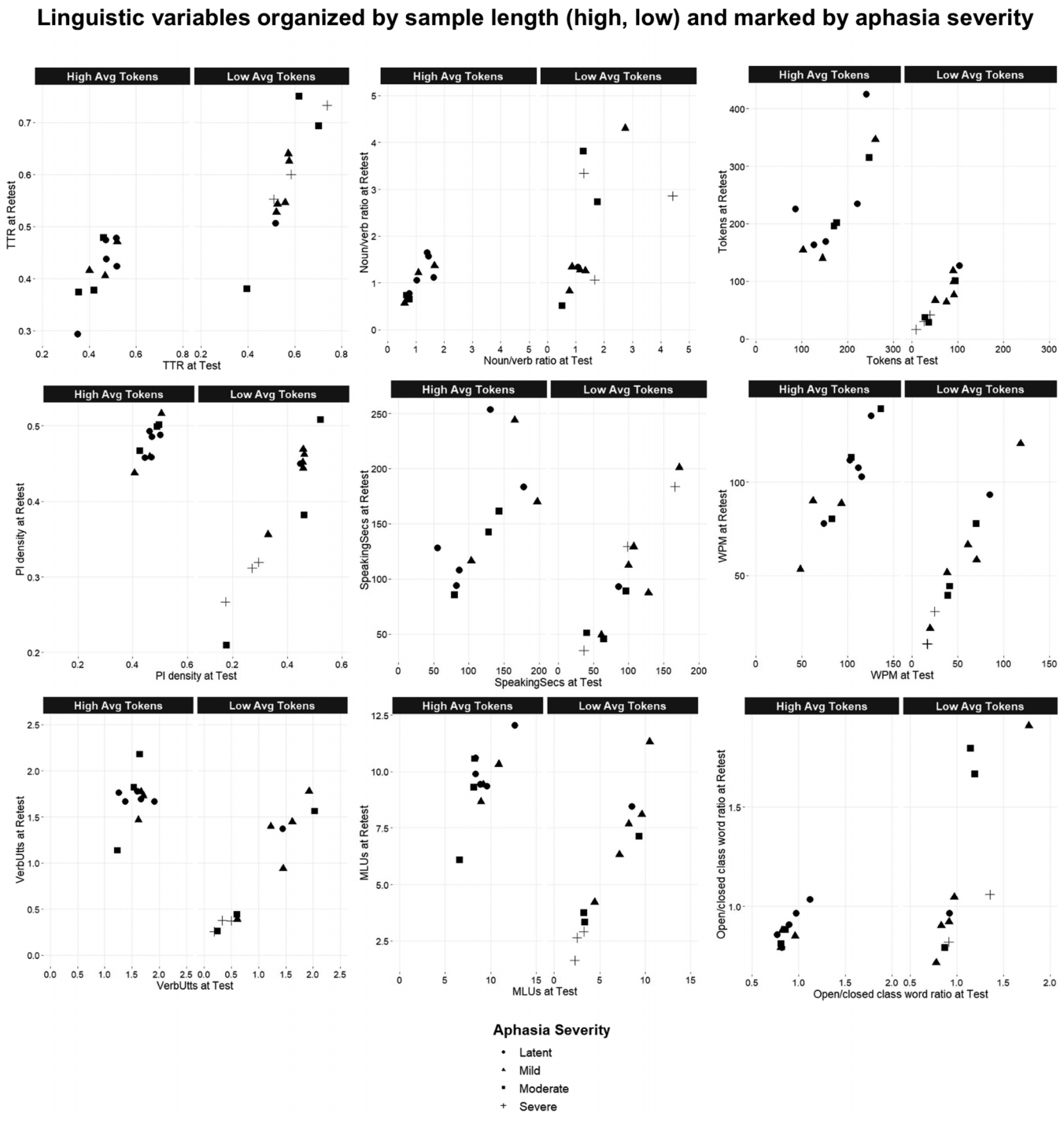


Within the aphasia group, there were notable differences in measure reliability during the cat rescue story by aphasia severity. %CIUs were more reliable in the mild or latent group than the moderate or severe group, a pattern also found for speaking duration, noun/verb ratio, and verbs per utterance. Other measures that differed by aphasia severity all differed in the opposite direction, where the moderate or severe group demonstrated higher reliability (for CIUs/min, speaking duration, WPM, noun/verb ratio, and verbs per utterance). Regarding presence of aphasia, for almost every measure (except open/closed class word ratio and MLU), the clinical aphasia group showed higher test–retest reliability (always higher reliability except for noun/verb ratio reliability). It is a reminder that it is difficult to make strong comparisons with the latent group considering the large difference in sample size between the two groups (latent group, $n = 6$; clinical aphasia, $n = 17$). They should be interpreted with caution.

We compared the cat rescue findings with the Cinderella (story retell) task, given that the Cinderella task produced many more tokens (this was true in both subject groups) and may therefore show a different pattern from what we found from the cat rescue. Regarding sample length, the NBD group demonstrated more reliable CIUs/min, WPM, noun/verb ratio, open/closed class word ratio, MLU, verbs per utterance, and speaking duration for longer samples. That is, most measures showed a sample length effect for Cinderella in not cat rescue task for the NBD group. For the aphasia group, there was a sample length impact on reliability for %CIUs, TTR, speaking duration, MLU, and verbs per utterance in the opposite direction, where short samples had higher reliability. The aphasia group also demonstrated the same pattern as the controls, where longer samples produced more reliable data, for CIUs/min, noun/verb ratio, and open/closed class word ratio. The cat rescue and Cinderella tasks had a similar number of measures experiencing sample length effects on their reliability for the aphasia group.

Within the aphasia group, there were notable differences in measure reliability during the Cinderella story by aphasia severity. %CIUs, propositional idea density, TTR, tokens, speaking duration, MLU, and verbs per utterance were all more reliable in the moderate or severe group. The opposite pattern was seen for CIUs/min and noun/verb ratio, which had higher reliability in the mild or latent group. In the clinical aphasia presence comparison, there

Figure 6. Test–retest reliability for microlinguistic variables derived from the discourse across tasks using CLAN for the aphasia group, faceted by sample length and marked by aphasia severity. Avg = average; CLAN = Computerized Language ANalysis; TTR = type–token ratio.



were notable differences for all measures between groups, with the majority demonstrating more reliability for the clinical aphasia group (except CIUs/min and open/closed class word ratio). In comparing cat rescue and Cinderella for aphasia severity and clinical aphasia presence groups, the Cinderella story had more measures with differing reliability between severity groups than the cat rescue.

The most notable pattern across tasks appeared to be the sample length effect on syntactic variables that

Table 6. Summary of test–retest reliability intraclass correlation interpretations by task by group.

Proxy	Measure	Group	Cat rescue	Cinderella	Sandwich	Broken window	Refused umbrella
Lexical and informativeness	%CIU	NBD Aphasia	Mod. (poor–good) Good (mod.–good)	Mod. (poor–good) Good (mod.–exc.)	Poor (poor) Mod. (poor–good)	Good (mod.–exc.) Mod. (poor–good)	Good (mod.–exc.) Good (mod.–exc.)
	PI density	NBD Aphasia	Poor (poor–mod.) Mod. (poor–good)	Poor (poor–mod.) Good (mod.–exc.)	Poor (poor) Good (mod.–exc.)	Poor (poor–mod.) Mod. (poor–good)	Poor (poor–mod.) Good (good–exc.)
	TTR	NBD Aphasia	Poor (poor–mod.) Mod. (poor–good)	Good (mod.–exc.) Good (mod.–exc.)	Mod. (poor–good) Mod. (poor–good)	Poor (poor–mod.) Good (mod.–exc.)	Mod. (poor–good) Mod. (poor–good)
	Tokens	NBD Aphasia	Mod. (poor–good) Good (mod.–exc.)	Good (mod.–exc.) Good (mod.–exc.)	Good (mod.–exc.) Mod. (poor–good)	Mod. (poor–good) Good (mod.–good)	Good (poor–exc.) Mod. (poor–good)
Fluency/efficiency	CIUs/min	NBD Aphasia	Good (mod.–exc.) Exc. (good–exc.)	Mod. (poor–good) Mod. (poor–good)	Poor (poor–mod.) Exc. (good–exc.)	Good (mod.–good) Good (good–exc.)	Mod. (poor–good) Exc. (good–exc.)
	Speaking Secs	NBD Aphasia	Good (mod.–exc.) Mod. (poor–good)	Good (mod.–exc.) Good (mod.–exc.)	Good (mod.–exc.) Mod. (poor–good)	Mod. (poor–good) Mod. (poor–good)	Good (poor–exc.) Mod. (poor–good)
	WPM	NBD Aphasia	Mod. (poor–good) Exc. (good–exc.)	Mod. (poor–good) Exc. (good–exc.)	Mod. (poor–good) Exc. (good–exc.)	Good (mod.–good) Exc. (good–exc.)	Mod. (poor–good) Good (good–exc.)
Syntactic	MLU	NBD Aphasia	Mod. (poor–good) Good (mod.–exc.)	Mod. (poor–good) Exc. (good–exc.)	Poor (poor–mod.) Good (mod.–exc.)	Poor (poor–mod.) Good (mod.–good)	Poor (poor–mod.) Good (mod.–exc.)
	Noun/verb	NBD Aphasia	Poor (poor–mod.) Mod. (poor–good)	Poor (poor–mod.) Mod. (poor–good)	Mod. (poor–good) Poor (poor–mod.)	Poor (poor) Mod. (poor–good)	Poor (poor–mod.) Poor (poor)
	Open/closed	NBD Aphasia	Poor (poor–mod.) Good (mod.–exc.)	Poor (poor–mod.) Mod. (poor–good)	Poor (poor–mod.) Good (mod.–good)	Poor (poor–mod.) Poor (poor)	Poor (poor) Poor (poor–mod.)
	VerbUtt	NBD Aphasia	Mod. (poor–good) Good (mod.–exc.)	Mod. (poor–good) Good (mod.–exc.)	Poor (poor) Mod. (poor–good)	Poor (poor) Good (mod.–good)	Poor (poor–mod.) Good (mod.–good)

Note. Koo and Li (2016) give the following suggestion for interpreting intraclass correlation coefficient, including confidence intervals: below .50 = poor; between .50 and .75 = moderate ("mod"); between .75 and .90 = good; and above .90 = excellent. ICC itself is listed first, with confidence interval interpretation in parentheses. Refer to Supplemental Tables S4–S8 for intraclass correlation coefficients with 95% confidence intervals, Spearman rho, and minimal detectable change at 90% confidence. See Supplemental Table S2. %CIU = percentage (proportion) of correct information units; NBD = no brain damage; Mod. = moderate; exc. = excellent; PI density = propositional idea density; TTR = type–token ratio; CIUs/min = correct information units per minute; SpeakingSecs = speaking duration in seconds; WPM = words per minute; MLU = mean length of utterance (in words); Noun/verb = noun-to-verb ratio; Open/closed = open-to-closed class word ratio; VerbUtt = verbs per utterance.

were ratios, especially noun/verb ratio and open/closed class word ratio, for both subject groups. This is discussed further in the Discussion section.

Discussion

This research responded to the need for robust psychometric data to improve the use of discourse as a reliable outcome measure (e.g., for experiments, speech therapy efficacy) by evaluating short interval test–retest reliability of a multitude of discourse measures, in individuals with aphasia and a set of adult individuals without brain damage. We hypothesized a difference in reliability measures when evaluated by group (aphasia, NBD) and by task. We demonstrated excellent inter- and intrarater reliability across discourse measures. We also demonstrated 100% retention, in that all participants attended both Test and Retest sessions.

The main takeaways of this project are that group and task differences exist in discourse measure reliability. There was a noticeable difference in the reliability of discourse measures by group (aphasia, NBD), with the aphasia group most often demonstrating better discourse measure reliability. Discourse measure reliability differed when evaluated by task for both groups. Aphasia severity as well as sample length impacted discourse measure reliability when measures were averaged across tasks and for each task. Finally, that even with short interval spacing (10 ± 3 days) between tasks, there can be a systematic difference in discourse measures at Retest, making it important to establish a double baseline to ensure accurate representation of each individual's variability for each measure and task.

Discourse Measures' Reliability

Across tasks, both subject groups produced ICC values for discourse measures that included poor, moderate, and good, with only the aphasia group producing measures with excellent test–retest reliability. Our primary variables of interest, %CIUs and CIUs/min across tasks, were both found to have excellent reliability when evaluated across tasks for the aphasia group. These measures were both found to have good reliability for the NBD group. This suggests, on the whole, that these primary measures are relatively reliable for both groups, when sampled across several tasks. As Figures 3 and 4 demonstrate, while the intragroup variability for the aphasia group was higher (i.e., individuals performing in a much wider range, but with higher absolute agreement between Test and Retest), the NBD group performed within a much smaller range, and with lower absolute agreement between Test and Retest on the CIU measures. Given that ICC is a measure that takes into account *both* absolute agreement and true intragroup variability (Koo & Li, 2016), it is perhaps not surprising to see such a difference in CIU performance between the groups. It is also telling that %CIUs and CIUs/min across tasks demonstrated the best test–retest reliability, as these measures were hand scored (authors A.H. and J.A.) and required manual decision making (see OSF).

Human-scored measures, like CIU, may produce the most reliable discourse measures in the aphasia group across tasks, at least at this juncture in time. This also brings to light the importance of demonstrating inter- and intrarater agreement for this human scoring, which we have done here and encourage others to critically examine. The issue with autoscoring (like what was done using CLAN for the other variables, e.g., MLU) is that no automatic tool can yet recognize the accuracy and informativeness related to semantic information, necessitating hand scoring CIUs. Because of this hand scoring, there is unavoidable conferring between raters as a specific rulebook is created (for instances where the original Brookshire & Nicholas, 1994, rules are not as straightforward to follow; our specific decisions/rulebook can be found on OSF in the Files section). It may be the conferring between raters as well as the general specificity of the CIU measures (i.e., evaluating multiple components, such as accuracy and informativeness) that has it exceeding reliability compared with the other metrics. CIUs have long been thought to be reliable for evaluating aphasic discourse, likely because they are one of the most evaluated measures in overall discourse literature. In the Brookshire and Nicholas (1994) and Boyle (2014) articles, they focused on test–retest reliability of %CIUs (and CIUs/min). While there were some methodological differences between their procedures and statistical analyses and ours, we find complementary results: %CIUs and CIUs/min tended to be the most reliable measures for both the aphasia and NBD groups. Altogether, there is somewhat strong support for use of %CIUs and CIUs/min across several discourse tasks for aphasia, though the findings for reliability of these measures in NBD are dampened slightly by our results (albeit, "good" reliability still may be adequate for larger group studies if evaluating NBD groups is of interest). A main takeaway is that, overall, the ICC values for %CIUs (and usually for CIUs/min and WPM) were > .70 in the aphasia group across tasks and by task, suggesting that these are the measures (if any) that might be most reliable across these types of monologue discourse tasks in wider aphasia populations.

There was also the trend that lexical and fluency measures appeared most reliable across tasks, within tasks, and for each group. Boyle (2014, 2015) also found fluency

measures (specifically, CIUs/min and WPM) to have greater test–retest reliability compared with most other discourse measures, like total words, main concept correctness, and word retrieval measures (e.g., word-finding errors). By contrast, other ratio variables like noun/verb and open/closed class word were least reliable across tasks, within tasks, and for each group (we discuss that this is likely related to sample length, below). One potential reason for this limited reliability for ratio terms could be that the denominator was not present in high enough quantity (e.g., verbs in noun/verb ratio), which makes these ratio measures more sensitive to fluctuations in the said denominator. To identify whether this was the case, we calculated the correlation between nouns and verbs, and between open and closed class words, which had a tendency to be the least reliable measures for most tasks. A positive, linear relationship suggests that the denominator was likely present in high enough quantity (Yoder & Symon, 2018). We identified positive, linear relationships between nouns and verbs and between open and closed class words for both subject groups, across tasks, and per each task (nouns and verbs: aphasia, across tasks, $r = .88$, by task, $r > .76$; NBD, across tasks, $r = .99$, by task, $r > .95$; open and closed class words: aphasia, across tasks, $r = .967$, by task, $r > .83$; NBD, across tasks, $r > .99$, by task, $r > .94$). It seems that, when sampled across tasks, there is enough sampling of the denominator to be confident that the test–retest measure is more "true" and less due to a sampling issue. We also noted that aphasia severity (particularly, moderate or severe severity) tended to associate with poor ratio measure reliability (Supplemental Table S2). We correlated nouns with verbs and open with closed class words again separately for mild or latent group and for moderate or severe group. We found that correlations were overall small yet positive for the moderate–severe group for each task (broken window, $r = .45$; cat rescue, $r = .18$, Cinderella, $r = .24$, sandwich, $r = .21$; refused umbrella, $r = .28$) and, surprisingly, were remarkably smaller for the mild or latent group (broken window, $r = -.10$; cat rescue, $r = -.12$; Cinderella, $r = .03$; sandwich, $r = .52$; refused umbrella, $r = .07$). Therefore, there may be some contribution of a lacking denominator to poor reliability of noun/verb ratio for specific tasks in the aphasia group, regardless of severity. When we evaluated the correlation between open and closed class words for severity groups, all tasks and both groups had positive correlation values (for moderate or severe group, all $r > .48$; for mild or latent group, all $r > .31$). Altogether, this emphasizes the importance of choosing variables that occur with enough frequency in the sample of interest (e.g., aphasia), which necessitates careful evaluation of the data (e.g., correlation of nouns/verbs to identify a strong positive, linear relationship) before committing to a variable serving as an outcome measure.

Aphasia Severity As Well As Sample Length May Negatively Impact Discourse Measure Reliability

Sample length appeared to have some effect on reliability measures, but this differed by sample group (aphasia, NBD) and by task. The other studies that have evaluated discourse measure test–retest reliability (e.g., Boyle, 2014, 2015; Brookshire & Nicholas, 1994) have attempted to correct for sample length by using percentage variables wherever possible (e.g., percentage of speech errors in Boyle, 2014). However, there has been no systematic investigation of the relationship between sample length and reliability, which we have done here. Altogether, we did not identify a consistent pattern of longer sample length associating with greater reliability and/or systematic difference when we evaluated across tasks. There was more support for the impact of sample length at the task level, which may have been due to differences in sample length per task (e.g., in the cat rescue, when the average number of total words produced by the aphasia group was 95.6 compared with the Cinderella story, which was 325). Interestingly, though, the NBD group showed more sample length impacts of reliability measures for the Cinderella story (the longer task in general), whereas the aphasia group showed a relatively similar number of variables impacted by sample length for both cat and Cinderella tasks. Therefore, it is important to evaluate the impact of sample length on reliability measures if you are evaluating a group with wide sample length ranges (as we have done here and as is classic in aphasia sampling) because the sample length affects may not be straightforward and may differ for the specific group being investigated, the specific participants in those groups, and the task cognitive demands.

Boyle (2015) cites a future direction as investigating the relationship of aphasia severity with discourse measure test–retest reliability, which is what we chose to do in this article. As with sample length, we did not identify a consistent pattern across tasks where aphasia severity predicted lower reliability and/or systematic difference. Tokens and speaking duration (both proxies often evaluated as "gross output" in studies, but also representing lexical and fluency variables, respectively) showed a difference between milder aphasia and more severe aphasia (in this particular sample, the reliability was higher for the more severe group). This pattern was switched for noun/verb ratio and open/closed class word ratio when comparing aphasia severity, with the milder group having a higher reliability for these. This may, again, have been due to some sampling bias for the ratio measures, as discussed earlier, and should be taken with a grain of salt. However, note in Supplemental Tables (S9–S13), which are by task, that it was not always the case that tokens and speaking duration were more reliable in

the more severe aphasia group (e.g., sandwich task). Therefore, we want to echo the supposition that we put forward in evaluating sample length and its relationship to reliability: It is important to evaluate the impact of aphasia severity on measures of reliability and difference across tasks and by task within your specific sample.

Systematic Differences Between Test and Retest

There were few systematic differences (as measured by Wilcoxon signed-ranks test) for discourse measures when evaluated across tasks, for both subject groups. Systematic differences that survived after multiple comparison corrected included tokens (for both subject groups) and speaking duration (for the NBD group only). In the case of tokens and speaking seconds, there was an increase at Retest. Despite our directions to participants (that, at Retest, they should answer the prompts as if they are speaking to a new person), there are undeniable sources of variance that come into play, such as familiarity and practice effects. We kept the same experimenter for both sessions, which was a conscious choice because we did not want to introduce undue variance into the test/retest environment. However, this may have influenced familiarity, where participants felt that they could speak more freely or openly at Retest. Furthermore, it is well known that these effects exist even with a short window of time between testing sessions (hence one of the major reasons for pursuing test/retest research), yet this has received less attention in the aphasic discourse research. Note, too, that practice effects are not restricted to only the picture tasks, but also to the narrative tasks, as practice effects can affect cognitive test performance due to repeated evaluation with the same procedures. For example, the narratives (Cinderella, sandwich) were elicited at Retest using the same procedures (i.e., set of instructions), which can influence output in a similar fashion as a picture might. Others have looked at the impact of instructions on causing a difference in lexical variables (e.g., Wright & Capilouto, 2009), and it follows that similar instructions across two timepoints may cause a practice effect just as the retelling based on a picture might. Practice effects are not "deal breakers" and are part of any testing environment. For this reason and others, multiple baselines become important because they establish typical variation no matter the number of testing times (see the single-subject research body for more information, e.g., Beeson & Robey, 2006; Cameron et al., 2010; Robey et al., 1999).

Should We Derive "Normative" Psychometric Data From a Control Group?

One such idea in doing psychometric research comparing clinical and nonclinical populations is that the nonclinical population may help to establish a "normative" standard to which the clinical group can be compared. Normative data are data from a population of interest that establishes a baseline distribution to which other, new samples can be compared. In some prior iterations of test–retest work in aphasia (e.g., Nicholas and Brookshire's research), discourse measure reliability in aphasia has been directly compared to discourse measure reliability in a "control" population, with the general thought process to compare and contrast measure reliability between the groups, and with the general consideration that discourse measure reliability from the control group was a "gold standard" to compare to (e.g., Brookshire & Nicholas, 1994).

At least from this (small) sample, it seems unlikely that deriving discourse measure test–retest data from a prospectively matched adult control group will establish "normative" data. This is because of the stark differences we identified between group performances and reliability dependent on discourse measure and task. This supposition is made most obvious from glancing at Table 5, where the NBD group's ICC differs greatly from that of the aphasia group, across nearly every discourse measure that was evaluated. As can be identified from Figures 4 and 5, this is likely due to differences in absolute agreement as well as a smaller intragroup range within the NBD group, which are both factors that adversely the ICC. As such, the NBD group's test–retest reliability for discourse measures cannot be considered the "standard" group to which the aphasia group is evaluated.

While some of these issues may be mitigated by increasing sample size of the NBD group and by better matching the NBD and aphasia group (which we acknowledge as a limitation of our study), there are some reasons why discourse measure test–retest reliability should always be evaluated in the aphasia sample of interest. Aphasia is a disorder characterized by heterogeneity, and therefore the group make-up (no matter how large the sample size) is likely to greatly influence test–retest reliability calculations. Our study and the others cited above (e.g., Boyle, 2014) emphasize the necessity of collecting double baselines (or to employ single-subject designs) to establish test–retest reliability specific to the study.

Task Differences Are Important to Consider When Evaluating Discourse Measure Reliability

One thing that neither Brookshire and Nicholas (1994) nor Boyle (2014) teased apart was the extent to which test–retest reliability was related to single tasks, for example, a picture description versus a procedural narrative. Our study makes clear that test–retest reliability is important to evaluate for each discourse measure across

tasks and for each task, given that a discourse measure's reliability when averaged across tasks is not necessarily representative of its task-specific reliability. Our results argue that it is critically important to evaluate the average of tasks as well as *task-specific reliability,* especially if accurate estimates about change (e.g., due to therapy) are to be made from these data and especially because many researchers and clinicians most often use a single task to measure discourse (Bryant et al., 2016). For example, our study suggests that %CIUs cannot be assumed to be clinically reliable for NBD or aphasia groups for every task (Supplemental Tables S4–S8). During the picture description (cat rescue), the NBD group and the aphasia group produced moderate and good reliability on %CIUs, respectively. This was similar for the Cinderella fictional story retell for %CIUs. However, for the procedural task (sandwich), the NBD produced poor test–retest reliability for %CIUs, but this was not the case for the aphasia group, who produced good test reliability for %CIUs. ICC values for the NBD group during the broken window and refused umbrella were quite similar to ICC values for the aphasia group during the broken window and refused umbrella.

Language undeniably varies between discourse tasks because each task requires specific cognitive, contextual, and linguistic components (see the following for in-depth discussions of this phenomenon: Fergadiotis & Wright, 2011; Leaman & Edmonds, 2021; Stark, 2019; Stark & Fukuyama, 2021). Many complementary studies report variability of language production between narrative, expository, and procedural subtypes of monologue discourse in aphasia (e.g., Armstrong, 2000; Boyle, 2011; Conroy et al., 2009; Fergadiotis & Wright, 2011; Kim et al., 2022; Linnik et al., 2016) and, more recently, in unstructured conversation in aphasia (Leaman & Edmonds, 2021). For example, it is thought that narrative tasks draw more upon complex grammatical processes as well as coordinate episodic memory with language in a way that tasks relying on an available picture (e.g., picture description) do not. We have previously shown how narrative tasks tend to elicit the most lexically diverse and grammatically complex discourse in aphasic and non-aphasic discourse compared with procedural and picture descriptions (Stark, 2019; Stark & Fukuyama, 2021), emphasizing that language structure varies by task and that it is theoretically important to select which discourse measure to extract from task by evaluating discourse measures that are adequately represented in the task. For example, from our prior articles and this article, it seems unwise to extract ratio, grammatical variables like noun/verb and open/closed class word ratio from samples that are short in length and samples that elicit simple grammar (e.g., picture descriptions, procedural tasks). When wanting to evaluate whether noun/verb ratio and/or open/closed class word ratio improves after some therapy, it would be wise to collect longer, more complex narrative samples using multiple baselines for these very reasons. In conclusion, we encourage researchers and clinicians to make educated choices about discourse measures that are representative (e.g., you would not choose a sample that elicits limited syntax to evaluate a syntactic measure) and sensitive to measuring change (e.g., a measure that occurs often enough at baseline to be able to measure a change at follow-up).

Our lab's work has also highlighted the importance of considering genre as well as specific task. A genre, such as a picture sequence description or a narrative, contains a variety of tasks (e.g., picture description genre can contain tasks like picnic description [from WAB] or cookie theft description [from the Comprehensive Aphasia Test; Swinburn et al., 2004]). In the work of Stark and Fukuyama (2021), we demonstrated that language structure is most similar for tasks within the same genre, for both aphasic and non-aphasic discourse samples. The current project expands on this finding by demonstrating that a discourse measure, which is reliable in one genre, may likely be reliable for a variety of tasks within that genre. In the specific tasks acquired in this project, both the broken window and the refused umbrella fall under the same genre (a picture sequence description). We see some consistency in measure reliability between broken window and refused umbrella tasks (e.g., poor for MLU and verb utterance, which are measures that tend to have higher test–retest reliability in different genres, such as picture description [cat rescue] and fictional narrative [Cinderella]). Therefore, collecting several tasks within a genre may be one way of examining and validating the test–retest reliability of a specific measure of interest.

Calculating Change Scores for Discourse

The aphasia group demonstrated considerable intragroup variability, and this high variability tends to inherently limit identification of statistical differences at the group level. Therefore, computing complementary statistical measures like MDC may capture clinically relevant change despite this variability. Here, we evaluated MDC at 90% confidence, motivated by Boyle (2014) and Boyle (2015)'s studies. MDC90 may be particularly helpful for demonstrating change when the discourse measure does not have particularly high reliability because it would identify single individuals who made an MDC even if the study were to fail to find a significant group-level trend. In Table 5, %CIUs are shown to have excellent reliability averaged across studies, but a significant change at the group level in %CIUs after therapy may still be difficult to identify given the wide variability in this measure within the aphasia group. Therefore, Table 5 also lists MDC90 of %CIUs as

0.10 (i.e., 10%), enabling a researcher/clinician to identify single participants from within the aphasia group who exceed a 10% change post-intervention. Despite a potentially not significant group effect of therapy, one can identify single participants who benefit. This type of change score—which is a type of reliability measure that requires test/retest methodology—is particularly beneficial in groups that are highly heterogeneous, like in aphasia. Other similar measures exist, such as minimal clinically important difference, which represents the smallest amount of change in an outcome that is considered important by patient and/or clinician (though this requires consensus prior to a study being performed; Fitzpatrick et al., 1998).

It is important to calculate MDC specific to one's sample and by task. For example, there are clear differences in the statistics reported by Boyle (2014) and our study, which both evaluated %CIUs in a sample of individuals with aphasia using similar discourse tasks. As a reminder, Boyle (2014) elicited discourse samples from five tasks that occurred 2–7 days apart without intervening treatment. For %CIUs in the work of Boyle (2014), MDC at 90% confidence ranged from 9% to 23%, suggesting that, dependent on the session, a difference needed to be at least 9% greater than baseline and, in some cases (Sessions 1–3, in her case), up to 23% greater than baseline to be considered a clinically meaningful change. In our sample, we found an MDC at 90% confidence to be 10% (across tasks), with individual tasks showing ranges of 21% (Cinderella task) to 32% (broken window task). Therefore, we encourage evaluating MDC as a complementary measure of reliability and emphasize that the MDC will be highly dependent on the unique study sample.

Conclusions

We evaluated the reliability of discourse measures in two groups (aphasia, no brain damage) averaged across five tasks and by task, hypothesizing broadly that there would be a difference by group and by task. To summarize main findings, the aphasia group tended to have discourse measures that were more often reliable (good or excellent reliability) across and by tasks, and these measures tended to represent lexical, informativeness, and/or fluency constructs. Some measures' reliability appeared to be influenced by both sample length and aphasia severity, which again varied across and by tasks. The aphasia group demonstrated a wider intragroup variance (i.e., wider spread of scores across the different measures), which may have improved the reliability scores given that some ways of quantifying reliability, like ICC, are influenced by not only absolute agreement but also the spread of scores within the group (i.e., true variability). The NBD group demonstrated more restricted variability on many discourse measures as well as on average less reliable discourse measures across tasks and by task. Like the aphasia group, sample length seemed to influence reliability to some extent, but this was measure specific, as well as task specific.

Summary of Recommendations

Throughout this article, we identified several discourse measures that would be reliable to use as outcome measures for our aphasia and NBD groups, but this depended on whether we pooled discourse measures across tasks or evaluated discourse measures by task. We encourage collecting multiple discourse samples across a variety of genres/tasks and not only computing mean and summative calculations on discourse measures across tasks but also reporting on discourse measures for each task. This is because we demonstrated here that test–retest reliability of discourse measures is intimately linked to the specific sample, sample length, and aphasia severity. As we and others have emphasized, the task itself should be considered an important variable, and it should not be assumed that discourse measures found to be reliable across tasks are likewise reliable for each task. This is due to the different cognitive processes required by tasks, as well as the different linguistic information (and quantity) produced for each task.

We also encourage calculation of not only statistical significance to measure test–retest reliability (e.g., ICC) but also complementary measures like MDC, which may be a more sensitive and appropriate way to understand change that occurs after some therapy. When deciding on discourse measures that are appropriate to use as outcome measures of treatment or similar studies, we encourage the use of both theory and data. That is, what measure might we expect a task to elicit in an adequate amount (theory) and do we confirm that in our data? For example, it is well known that narratives produce more complex syntax than other tasks (e.g., Stark, 2019), and your specific data support this (e.g., a syntactic measure like MLU or verbs per utterance is appropriate to evaluate in your sample because of a high-enough occurrence of utterances and/or verbs), so it is reasonable and valid to evaluate this syntactic measure after treatment during narrative samples.

Finally, we emphasize the importance of transparent reporting of discourse-related information, drawing from FOQUSAphasia's recent working group guidelines (Stark, Dutta, Murray, Bryant, et al., 2021; Stark et al., 2022). Future work via large repositories such as AphasiaBank should focus on collecting double baseline data to evaluate test–retest for a variety of discourse measures in larger samples. A larger sample is critical if there are solid conclusions to be drawn about test–retest reliability and its

relationship to stimuli/task, subject group, and biographical/cognitive variables. In this study, we suggest that using a control group's test–retest data as a normative or comparative sample may not be valid for making choices about potential discourse measure reliability demonstrated by an aphasia group. Larger samples will confirm whether this is the case.

Limitations and Future Directions

Despite attempts to prospectively match the non–brain-damaged control group, there were differences in age, sex, and days between samples. In our Markdown, we ran correlations between age and days between samples with average discourse measures (calculated across tasks and calculated across NBD and aphasia samples together), finding all correlations to be $p > .05$ when corrected for multiple comparisons. This suggests no systematic relationship with these demographic variables and our discourse measures of interest. Chi-square tests similarly indicated no systematic gender differences in average discourse measures.

We have discussed future work involving larger sample sizes. As some authors have stated, it is not appropriate to give one number for sample size in all such cases, and starting with a sample of 30–50 subjects (as we have done here) is a reasonable first evaluation of psychometric properties through classical test theory (Cappelleri et al., 2014). One such reason why a larger sample size is an ideal next step is due to the wide variation in the amount of data available in discourse samples, especially in individuals with more severe aphasias (e.g., some individuals with more severe aphasia may only be able to produce a few words during a sample). However, another way of going about this is to create an even level of observations across categories (Cappelleri et al., 2014; e.g., aphasia severities) by designing studies that sufficiently sample individuals that have more severe language impairments. A focus on discourse sampling from individuals with more severe aphasia would be highly beneficial, as individuals in this study tend to be milder, and that is also the trend of a majority of data available in AphasiaBank (http://Aphasia.talkbank.org), a large repository of aphasic discourse data. A further interesting next step would be to use the generalizability theory to systematically explore contributors to reliability (e.g., sample length) in a larger sample size and/or a more evenly distributed sample (Monteiro et al., 2019; Webb et al., 2006)

In this particular study, we focused on linguistic structure rather than use measures (e.g., use measures such as topic adherence, story grammar). Furthermore, our linguistic variables were largely lexicosemantic and fluency oriented in nature, with some evaluating morphosyntax. Future work should evaluate the test–retest reliability of more fine-grained measures of syntax, for example, morphological impairments specific to agrammatism and/or paragrammatism, as well as instances of language use.

A reviewer astutely noticed that we did not evaluate measurement invariance (MI), which is a psychometric property that evaluates the extent to which using the same test (e.g., discourse elicitation) in different groups (e.g., test/retest, control vs. aphasia) measures the same construct (e.g., some latent construct) in the same way. One of the biggest reasons we did not evaluate MI is because of an issue with latent factor identification in discourse. MI, in theory, would be ideal to evaluate, because it evaluates invariance across group comparisons as well as across measurement occasions, and is a complementary (validity) measure when conducted alongside test–retest reliability. MI is far easier (and more theoretically sound) when evaluating tests with set outcomes, for example, naming tests where we know the correct set of answers, or tests with a finite set of answers, like multiple choice. The issue specific to discourse, especially in aphasia, is that, while theoretical distinctions can be made about "latent" factors in discourse (e.g., our Table 1, where we explicitly call them "proxies" rather than "latent factors"), in reality, the factoring of discourse data is difficult. In our early statistical analysis of the data, we ran exploratory factor analyses to identify whether it would be reasonable to assess the test–retest reliability of latent factors, which, in retrospect, seems to get at the idea of test–retest MI instead. This analysis remains in our Markdown in OSF. According to Kaiser's (1974) guidelines, a suggested cutoff for determining the factorability of the sample data is Kaiser–Meyer–Olkin (KMO) Measurement Sampling Adequacy (MSA) ≥ 0.60. When we combined aphasia and control group data (across test and retest) to evaluate this, the total MSA was 0.75 across the group, suggesting "middling" factorability. Note, though, that the aphasia group was driving this factorability (KMO) estimate, as the NBD group's estimate was poor (around MSA = 0.30). Within all groups, speaking duration had a very low KMO (~0.3), suggesting limited factorability of that variable in particular. Therefore, we used a parallel analysis with principal factors (excluding speaking duration) to try to identify possible factor solutions. When we did this, the models suggested that only one factor could be identified from the aphasia data. When we evaluated that one factor solution for its properties, its Tucker Lewis Index of factoring reliability was ~0.50, which is far below the acceptable number (> 0.90). Altogether, we are not confident that any "latent" factor(s) could be identified in our data set, and that further evaluation of MI would be inherently flawed. There are a few theoretical reasons for difficulty identifying latent factors from discourse. The first is that the construct of "latent" does not quite transfer to discourse as it does to

types of tests, such as questionnaires or, more commonly in the aphasia language literature, confrontation naming tests (e.g., Fergadiotis & Wright, 2015). This is why, in Table 1, we made the decision to call the language variables that were being represented by the individual discourse measures "proxies," rather than factors, constructs, or similar. They are theoretical properties, which others have before evaluated (as described in the Method section where we discuss the rationale for selecting these). All this to say—we love the idea of MI and hope to see it explored more when evaluating properties of other tests (e.g., naming), but at the current moment, it does not make sense for this present work. We intend to explore it as the test–retest data set in AphasiaBank grows, especially if larger numbers enable identification of latent factors (see Gordon, 2020, for factor analysis of discourse in aphasia, which explained ~50% of variance in the discourse).

This study was administered virtually. Because no other study (to our knowledge) has evaluated test–retest reliability in this way, we encourage replication of our methods and results in a new group of individuals with aphasia and neurotypical peers.

Data Availability Statement

Audiovisual, transcript, demographic, and neuropsychological data are available on AphasiaBank (http://aphasia.talkbank.org) under Protocol section, within the NEURAL Research Lab page. Interested parties are advised to become members of AphasiaBank to access these data.

Aphasia data: https://aphasia.talkbank.org/access/English/Aphasia/NEURAL.html

Control data: https://aphasia.talkbank.org/access/English/Control/NEURAL.html

Acknowledgments

This research was supported by an American Speech-Language-Hearing Foundation New Investigator Award to Brielle C. Stark. The authors would like to thank their two reviewers for thoughtful, constructive comments.

References

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

Beeson, P. M., & Robey, R. R. (2006). Evaluating single-subject treatment research: Lessons learned from the aphasia literature. *Neuropsychological Review, 16*(4), 161–169. https://doi.org/10.1007/s11065-006-9013-7

Boles, L. (1998). Conversational discourse analysis as a method for evaluating progress in aphasia: A case report. *Journal of Communication Disorders, 31*(3), 261–274. https://doi.org/10.1016/S0021-9924(98)00005-7

Boyle, M. (2011). Discourse treatment for word retrieval impairment in aphasia: The story so far. *Aphasiology, 25*(11), 1308–1326. https://doi.org/10.1080/02687038.2011.596185

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

Boyle, M. (2015). Stability of word-retrieval errors with the AphasiaBank stimuli. *American Journal of Speech-Language Pathology, 24*(4), S953–S960. https://doi.org/10.1044/2015_AJSLP-14-0152

Boyle, M., Akers, C. M., Cavanaugh, R., Hula, W. D., Swiderski, A. M., & Elman, R. J. (2022). Changes in discourse informativeness and efficiency following communication-based group treatment for chronic aphasia. *Aphasiology, 37*(3), 563–597. https://doi.org/10.1080/02687038.2022.2032586

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

Brown, C., Snodgrass, T., & Covington, M. (2007). *Computerized Propositional Idea Density Rater 3 (CPIDR 3)*. Artificial Intelligence Center: CASPR Project.

Brown, C., Snodgrass, T., Kemper, S. J., Herman, R., & Covington, M. A. (2008). Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods, 40*(2), 540–545. https://doi.org/10.3758/BRM.40.2.540

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

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

Cappelleri, J. C., Lundy, J. J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics, 36*(5), 648–662. https://doi.org/10.1016/j.clinthera.2014.04.006

Conroy, P., Sage, K., & Ralph, M. L. (2009). Improved vocabulary production after naming therapy in aphasia: Can gains in picture naming generalise to connected speech? *International Journal of Language & Communication Disorders, 44*(6), 1036–1062. https://doi.org/10.1080/13682820802585975

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

Dabul, B. (2000). *Apraxia Battery for Adults*. Pro-Ed.

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

Donoghue, D., & Stokes, E. K. (2009). How much change is true change? The minimum detectable change of the Berg Balance Scale in elderly people. *Journal of Rehabilitation Medicine, 41*(5), 343–346. https://doi.org/10.2340/16501977-0337

Doub, A., Hittson, A., & Stark, B. C. (2021). Conducting a virtual study with special considerations for working with persons with aphasia. *Journal of Speech, Language, and Hearing Research, 64*(6), 2038–2046. https://doi.org/10.1044/2021_JSLHR-20-00392

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

Evans, W. S., Cavanaugh, R., Quique, Y., Boss, E., Starns, J. J., & Hula, W. D. (2020). *Playing with BEARS: Balancing Effort, Accuracy, and Response Speed in a semantic feature verification anomia treatment game*. Abstract for Platform Presentation, Annual Clinical Aphasiology Conference (Conference Cancelled).

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

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. (2011). Lexical diversity for adults with and without aphasia across discourse elicitation tasks. *Aphasiology, 25*(11), 1414–1430. https://doi.org/10.1007/s11103-011-9767-z.Plastid

Fergadiotis, G., & Wright, H. H. (2015). Modelling confrontation naming and discourse performance in aphasia. *Aphasiology, 30*(4), 364–380. https://doi.org/10.1080/02687038.2015.1067288

Fergadiotis, G., Wright, H. H., & West, T. M. (2013). Measuring lexical diversity in narrative discourse of people with aphasia. *American Journal of Speech-Language Pathology, 22*(2), S397–S408. https://doi.org/10.1044/1058-0360(2013/12-0083)

Ferketich, S. (1990). Internal consistency estimates of reliability. *Research in Nursing & Health, 13*(6), 437–440. https://doi.org/10.1002/nur.4770130612

Fitzpatrick, R., Davey, C., Buxton, M., & Jones, D. (1998). Evaluating patient-based outcome measures for use in clinical trials. *Health Technology Assessment, 2*(14). https://doi.org/10.3310/hta2140

Fromm, D., Forbes, M., Holland, A., & Dalton, G. (2017). Discourse characteristics in aphasia beyond the Western Aphasia Battery cutoff. *American Journal of Speech-Language Pathology, 26*(3), 762–768. https://doi.org/10.1044/2016_AJSLP-16-0071

Gordon, J. K. (2020). Factor analysis of spontaneous speech in aphasia. *Journal of Speech, Language and Hearing Research, 63*(12), 4127–4147. https://doi.org/10.1044/2020_JSLHR-20-00340

Harris, P. A., Taylor, R., Minor, B. L., Elliott, V., Fernandez, M., O'Neal, L., McLeod, L., Delacqua, G., Delacqua, F., Kirby, J., & Duda, S. N. (2019). The REDCap consortium: Building an international community of software platform partners. *Journal of Biomedical Informatics, 95,* 103208. https://doi.org/10.1016/j.jbi.2019.103208

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics, 42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

Herbert, R., Hickin, J., Howard, D., Osborne, F., & Best, W. (2008). Do picture-naming tests provide a valid assessment of lexical retrieval in conversation in aphasia? *Aphasiology, 22*(2), 184–203. https://doi.org/10.1080/02687030701262613

Hula, W. D., & McNeil, M. R. (2008). Models of attention and dual-task performance as explanatory constructs in aphasia. *Seminars in Speech and Language, 29*(03), 169–187. https://doi.org/10.1055/s-0028-1082882

Kertesz, A. (2007). *Western Aphasia Battery–Revised*. The Psychological Corporation.

Kim, H., Berube, S., & Hillis, A. E. (2022). Core lexicon in aphasia: A longitudinal study. *Aphasiology*. Advance online publication. https://doi.org/10.1080/02687038.2022.2121598

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Leaman, M. C., & Edmonds, L. A. (2021). Assessing language in unstructured conversation in people with aphasia: Methods, psychometric integrity, normative data, and comparison to a structured narrative task. *Journal of Speech, Language, and Hearing Research, 64*(11), 4344–4365. https://doi.org/10.1044/2021_JSLHR-20-00641

Linnik, A., Bastiaanse, R., & Höhle, B. (2016). Discourse production in aphasia: A current review of theoretical and methodological challenges. *Aphasiology, 30*(7), 765–800. https://doi.org/10.1080/02687038.2015.1113489

MacWhinney, B. (2000). *The CHILDES project: Tools for analysing talk* (3rd ed.). Erlbaum.

MacWhinney, B. (2018). *Tools for analyzing talk part 2: The CLAN program* (issue 2000, p. 179).

MacWhinney, B., Forbes, M., & Holland, A. (2011). AphasiaBank: Methods for studying discourse. *Aphasiology, 25*(11), 1286–1307. https://doi.org/10.1080/02687038.2011.589893.AphasiaBank

Matheson, G. J. (2019). We need to talk about reliability: Making better use of test-retest studies for study design and interpretation. *PeerJ, 7,* Article e6918. https://doi.org/10.7717/peerj.6918

McGraw, K., & Wong, S. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Monteiro, S., Sullivan, G. M., & Chan, T. M. (2019). Generalizability theory made simple(r): An introductory primer to G-studies. *Journal of Graduate Medical Education, 11*(4), 365–370. https://doi.org/10.4300/JGME-D-19-00464.1

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatric Society, 53*(4), 695–699. https://doi.org/10.1111/j.1532-5415.2005.53221.x

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

Nichols, T., Das, S., Eickhoff, S., Evans, A., Glatard, T., Hanke, M., Kriegeskorte, N., Milham, M., Poldrack, R., Poline, J.-B., Proal, E., Thirion, B., Van Essen, D., White, T., & Yeo, B. T. T. (2017). Best practices in data analysis and sharing in neuroimaging using MRI. *Nature Neuroscience, 20*(3), 299–303. https://doi.org/10.1038/nn.4500

Pritchard, M., Hilari, K., Cocks, N., & Dipper, L. (2017). Reviewing the quality of discourse information measures in aphasia. *International Journal of Language & Communication Disorders, 52*(6), 689–732. https://doi.org/10.1111/1460-6984.12318

Pritchard, M., Hilari, K., Cocks, N., & Dipper, L. (2018). Psychometric properties of discourse measures in aphasia:

Acceptability, reliability, and validity. *International Journal of Language & Communication Disorders, 53*(6), 1078–1093. https://doi.org/10.1111/1460-6984.12420

Robey, R., Schultz, M., Crawford, A., & Sinner, C. (1999). Review: Single-subject clinical-outcome research: Designs, data, effect sizes, and analyses. *Aphasiology, 13*(6), 445–473. https://doi.org/10.1080/026870399402028

Saffran, E. M., Berndt, R. S., & Schwartz, M. F. (1989). The quantitative analysis of agrammatic production: Procedure and data. *Brain and Language, 37*(3), 440–479. https://doi.org/10.1016/0093-934X(89)90030-8

Shadden, B. B., Burnette, R. B., Eikenberry, B. R., & DiBrezzo, R. (1991). All discourse tasks are not created equal. *Clinical Aphasiology, 20,* 327–342.

Shrout, P., & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Simmons-Mackie, N., Worrall, L., Murray, L. L., Enderby, P., Rose, M. L., Paek, E. J., & Klippi, A. (2017). The top ten: Best practice recommendations for aphasia. *Aphasiology, 31*(2), 131–151. https://doi.org/10.1080/02687038.2016.1180662

Stark, B. C. (2019). A comparison of three discourse elicitation methods in aphasia and age-matched adults: Implications for language assessment and outcome. *American Journal of Speech-Language Pathology, 28*(3), 1067–1083. https://doi.org/10.1044/2019_AJSLP-18-0265

Stark, B. C., Bryant, L., Themistocleous, C., den Ouden, D.-B., & Roberts, A. C. (2022). Best practice guidelines for reporting spoken discourse in aphasia and neurogenic communication disorders. *Aphasiology, 37*(5), 761–784. https://doi.org/10.1080/02687038.2022.2039372

Stark, B. C., Dutta, M., Murray, L. L., Bryant, L., Fromm, D., MacWhinney, B., Ramage, A. E., Roberts, A., den Ouden, D.-B., Brock, K., McKinney-Bock, K., Paek, E. J., Harmon, T. G., Yoon, S. O., Themistocleous, C., Yoo, H., Aveni, K., Gutierrez, S., & 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

Stark, B. C., Dutta, M., Murray, L. L., Fromm, D., Bryant, L., Harmon, T. G., Ramage, A. E., & Roberts, A. C. (2021). Spoken discourse assessment and analysis in aphasia: An international survey of current practices. *Journal of Speech, Language, and Hearing Research, 64*(11), 4366–4389. https://doi.org/10.1044/2021_JSLHR-20-00708

Stark, B. C., & Fukuyama, J. (2021). Leveraging big data to understand the interaction of task and language during monologic spoken discourse in speakers with and without aphasia. *Language, Cognition and Neuroscience, 36*(5), 562–585. https://doi.org/10.1080/23273798.2020.1862258

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

Thompson, C. K., Meltzer-Asscher, A., Cho, S., Lee, J., Wieneke, C., Weintraub, S., & Mesulam, M.-M. (2013). Syntactic and morphosyntactic processing in stroke-induced and primary progressive aphasia. *Behavioural Neurology, 26*(1–2), 35–54. https://doi.org/10.1155/2013/749412

Thorne, J., & Faroqi-Shah, Y. (2016). Verb production in aphasia: Testing the division of labor between syntax and semantics. *Seminars in Speech and Language, 37*(01), 023–033. https://doi.org/10.1055/s-0036-1571356

Traub, R. E. (2005). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice, 16*(4), 8–14. https://doi.org/10.1111/j.1745-3992.1997.tb00603.x

Ulatowska, H. K., North, A. J., & Macaluso-Haynes, S. (1981). Production of narrative and procedural discourse in aphasia. *Brain and Language, 13*(2), 345–371. https://doi.org/10.1016/0093-934X(81)90100-0

Wallace, S. J., Worrall, L. E., Rose, T., & Le Dorze, G. (2018). Discourse measurement in aphasia research: Have we reached the tipping point? A core outcome set … or greater standardisation of discourse measures? *Aphasiology, 32*(4), 479–482. https://doi.org/10.1080/02687038.2017.1398811

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). 4 reliability coefficients and generalizability theory. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 81–124). Elsevier. https://doi.org/10.1016/S0169-7161(06)26004-8

Wilson, S. M., Henry, M. L., Besbris, M., Ogar, J. M., Dronkers, N. F., Jarrold, W., Miller, B. L., & Gorno-Tempini, M. L. (2010). Connected speech production in three variants of primary progressive aphasia. *Brain, 133*(7), 2069–2088. https://doi.org/10.1093/brain/awq129

Worrall, L., Sherratt, S., Rogers, P., Howe, T., Hersh, D., Ferguson, A., & Davidson, B. (2011). What people with aphasia want: Their goals according to the ICF. *Aphasiology, 25*(3), 309–322. https://doi.org/10.1080/02687038.2010.508530

Wright, H. H., & Capilouto, G. J. (2009). Manipulating task instructions to change narrative discourse performance. *Aphasiology, 23*(10), 1295–1308. https://doi.org/10.1080/02687030902826844

Yoder, P., & Symon, F. (2018). *Observational measurement of behavior*. Springer Publishing Company. https://www.springerpub.com/observational-measurement-of-behavior-9780826137975.html