AJSLP

Research Note

Standardizing Assessment of Spoken Discourse in Aphasia: A Working Group With Deliverables

Brielle C. Stark,[a,b] Manaswita Dutta,[a] Laura L. Murray,[c] Lucy Bryant,[d] Davida Fromm,[e] Brian MacWhinney,[e] Amy E. Ramage,[f] Angela Roberts,[g] Dirk B. den Ouden,[h] Kris Brock,[i] Katy McKinney-Bock,[j] Eun Jin Paek,[k] Tyson G. Harmon,[l] Si On Yoon,[m] Charalambos Themistocleous,[n] Hyunsoo Yoo,[o] Katharine Aveni,[g] Stephanie Gutierrez,[g] and Saryu Sharma[p]

Purpose: The heterogeneous nature of measures, methods, and analyses reported in the aphasia spoken discourse literature precludes comparison of outcomes across studies (e.g., meta-analyses) and inhibits replication. Furthermore, funding and time constraints significantly hinder collecting test–retest data on spoken discourse outcomes. This research note describes the development and structure of a working group, designed to address major gaps in the spoken discourse aphasia literature, including a lack of standardization in methodology, analysis, and reporting, as well as nominal data regarding the psychometric properties of spoken discourse outcomes.
Method: The initial initiatives for this working group are to (a) propose recommendations regarding standardization of spoken discourse collection, analysis, and reporting in aphasia, based on the results of an international survey and a systematic literature review and (b) create a database of test–retest spoken discourse data from individuals with and without aphasia. The survey of spoken discourse collection, analysis, and interpretation procedures was distributed to clinicians and researchers involved in aphasia assessment and rehabilitation from September to November 2019. We will publish survey results and recommend standards for collecting, analyzing, and reporting spoken discourse in aphasia. A multisite endeavor to collect test–retest spoken discourse data from individuals with and without aphasia will be initiated. This test–retest information will be contributed to a central site for transcription and analysis, and data will be subsequently openly curated.
Conclusion: The goal of the working group is to create recommendations for field-wide standards in methods, analysis, and reporting of spoken discourse outcomes, as has been done across other related disciplines (e.g., Consolidated Standards of Reporting Trials, Enhancing the Quality and Transparency of Health Research, Committee on Best Practice in Data Analysis and Sharing). Additionally, the creation of a database through our multisite collaboration will allow the identification of psychometrically sound outcome measures and norms that can be used by clinicians and researchers to assess spoken discourse abilities in aphasia.

[a]Department of Speech, Hearing and Language Sciences, Indiana University Bloomington
[b]Program in Neuroscience, Indiana University Bloomington
[c]School of Communication Sciences and Disorders, Western University, London, Ontario, Canada
[d]Graduate School of Health, University of Technology Sydney, New South Wales, Australia
[e]Department of Psychology, Carnegie Mellon University, Pittsburgh, PA
[f]Department of Communication Sciences and Disorders, University of New Hampshire, Durham
[g]Roxelyn and Richard Pepper Department of Communication Sciences and Disorders, Northwestern University, Evanston, IL
[h]Department of Communication Sciences and Disorders, University of South Carolina, Columbia
[i]Department of Communication Sciences and Disorders, Idaho State University, Pocatello
[j]Center for Spoken Language Understanding, Oregon Health and Science University, Portland
[k]Department of Audiology and Speech Pathology, University of Tennessee Health Science Center, Knoxville
[l]Department of Communication Disorders, Brigham Young University, Provo, UT
[m]Department of Communication Sciences and Disorders, University of Iowa, Iowa City
[n]Department of Neurology, Johns Hopkins University, Baltimore, MD
[o]Department of Communication Sciences and Disorders, Baylor University, Waco, TX
[p]Department of Communication Sciences and Disorders, East Carolina University, Greenville, NC

Correspondence to Brielle C. Stark: bcstark@iu.edu

Editor-in-Chief: Melissa Duff
Editor: Sarah Elizabeth Wallace

Received September 12, 2019
Revision received November 21, 2019
Accepted January 3, 2020
https://doi.org/10.1044/2020_AJSLP-19-00093
Publisher Note: This article is part of the Special Issue: Select Papers From the 49th Clinical Aphasiology Conference.

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

Discourse is a fundamental aspect of functional communication. Spoken discourse production difficulties can significantly negatively affect individuals' social communicative competence and their quality of life (Galski et al., 1998; Sim et al., 2013). Accordingly, spoken discourse analysis is a topic of increasing interest in aphasia assessment, treatment, and research (e.g., Bryant et al., 2016) and, with improved methodological rigor and analysis standardization, has the potential to serve as a primary and important outcome measure (e.g., Brady et al., 2016; Wallace et al., 2017). With respect to the International Classification of Functioning, Disability and Health model (World Health Organization, 2018), evaluation of spoken discourse provides an ecologically valid method to assess the day-to-day social participation and activity challenges faced by individuals with aphasia in social settings as a result of their communication difficulties. As such, best practice guidelines in both Australia and the United States have endorsed including spoken discourse analysis in comprehensive aphasia assessment (Clinical Centre for Research Excellence in Aphasia Research, 2014; Winstein et al., 2016).

Analyzing spoken discourse gleans microstructural (e.g., syntax, lexical-semantic structure) and macrostructural (e.g., cohesion, coherence) information in a comparatively naturalistic manner in contrast to other spoken language tasks, such as confrontation naming or repetition. To collect connected speech samples, structured and semi-structured prompts are frequently used, including single picture or picture sequence description, story retell, procedural description, and personal narratives. Increasingly, conversations with a clinician and/or familiar communication partner are also being analyzed due to their close tie to language used during activities of daily living (e.g., Armstrong, 2000; Beeke et al., 2007; Damico et al., 1999; Ulatowska et al., 1992). Language elicited during the abovementioned discourse tasks is proposed to be at least partially prompt dependent (e.g., Fergadiotis et al., 2011; Stark, 2019; Wright & Capilouto, 2009).

Despite spoken discourse analysis in aphasia gaining widespread importance in clinical and research settings, no standards exist for the most clinically useful outcome measures or data-reporting procedures, leading to inconclusive findings (Bryant et al., 2016; Dietz & Boyle, 2018). In addition to the wide variety of measures used, the heterogeneity in findings could also be due to a large proportion of aphasia studies relying on a small participant sample given the difficulty in recruiting this clinical population. As such, there is value in being able to aggregate data and protocols across sites.

Given the inconsistencies in discourse measurement and analysis procedures across aphasia studies, experts have agreed that research in this area has reached a tipping point where a more systematic approach is necessary (Dietz & Boyle, 2018; Kintz & Wright, 2018). The recently established core outcome set for aphasia treatment research demonstrate the concerted effort made by the aphasia community to adopt systematic assessment and reporting of aphasia outcomes, allowing for more robust data aggregation (e.g., meta-analyses) and reproducibility (Wallace et al., 2019). Discourse is not presently included in the core outcome set for aphasia treatment due to the scarcity of psychometric information on discourse outcome measures and vast heterogeneity in previous studies' discourse sampling and analysis procedures and, consequently, findings. Accordingly, there is a need for a multisite approach to address these issues, as collaborating and collecting spoken discourse data across multiple sites will allow for acquisition of a larger sample size that captures the variability inherent in discourse while also providing enough power to derive psychometrically sound measures.

To clarify, when referring to discourse "outcomes," we are referring to the micro- or macrostructural features extracted from the spoken sample (i.e., dependent variables). Typically, the goal of clinicians and researchers is to choose one or more discourse-derived outcomes that are representative of an element of the speech-language system. For example, one can extract information related to syntactic complexity by evaluating outcomes such as proportion of prepositions or complete sentences produced. Understandably, many outcomes can be and have been studied, resulting in a plethora of spoken discourse outcome measures. Indeed, over 536 unique discourse outcomes have been reported in the aphasia literature (Bryant et al., 2016). This heterogeneity precludes meta-analytic and systematic comparison of studies' findings, thus hindering the development of best practices in spoken discourse analysis in aphasia research and clinical practice.

Accordingly, the purposes of establishing this working group described in this research note are to (a) evaluate current practices and barriers to systematically collecting and evaluating spoken discourse outcomes in aphasia; (b) establish standards to systematically collect, analyze, and report information on spoken discourse analysis in aphasia; and (c) collect and disseminate data regarding test–retest reliability of frequently used spoken discourse outcomes for persons with and without aphasia.

Research Gap #1: No Standards for Collecting and Reporting Evidence

In addition to the large number of discourse outcome measures reported in the aphasia literature, there is a notable lack of agreement among researchers and clinicians regarding spoken discourse sampling, measurement, transcription, and analysis procedures, resulting in inconclusive findings. It must be noted that heterogeneity across reporting of spoken discourse outcomes also extends to aphasia treatment studies. For example, Richardson et al. (2016) evaluated assessment fidelity in aphasia research and noted that, across 88 treatment studies published between 2010 and 2015, less than 10% of the studies reported information on assessment instruments used and tester or rater training, approximately 35% reported information regarding tester qualifications, 37.5% reported tester or rater reliability,

and only about 27% of the studies reporting tester blinding and no studies reported information regarding assessment delivery.

Vague or inadequate descriptions of discourse elicitation and analysis procedures in addition to reporting limited participant-related information restrict comparison of outcomes across studies or translated use in clinical practice (Brookshire, 1983). Furthermore, the psychometric quality of even frequently used discourse measures (e.g., correct information units) and transcriptions are inconsistently reported, especially those concerning inter- and intrarater reliability (Pritchard et al., 2017). When interrater reliability is reported, different statistics have been used (e.g., intraclass correlation coefficient [ICC], percentage agreement) and few studies report intrarater agreement. Reporting inter- and intrarater agreement allows drawing conclusions about a study's data quality. Additionally, consistent and appropriate statistical analysis allows for comparison across studies.

Basic and recommended reporting standards have been developed to report most aspects of a research study, including study design, data collection, analysis, results, and interpretation (Gearing et al., 2011). Many fields have recognized that such reporting standards are key to research replication and robustness. For example, the Committee on Best Practice in Data Analysis and Sharing (COBIDAS) is a working group of experts on human brain mapping, who created standards for reporting methods and results in published works (Nichols et al., 2017). The stated purpose of COBIDAS was to elaborate the principles of open and reproducible research and to distill these principles in specific research practices. Studies comprise many elements, not all of which can be prescribed or restricted. However, COBIDAS and other initiatives encourage researchers to specify the information that must be reported to fully understand and potentially replicate a study and infer concrete conclusions regarding frequently used experimental measures. Across seven study areas (i.e., experimental design, acquisition of data, preprocessing, statistical modeling and inferencing, results, data sharing, and reproducibility), COBIDAS suggested best practices and reporting standards for over 100 items to help plan, execute, report, and share research in a transparent manner. Many scientific journals strongly encourage reviewers to use the COBIDAS reporting standards when evaluating the quality of a human brain mapping manuscript. Notably, COBIDAS is a living initiative, and their report continues to be updated and improved as the field grows and changes. Other similar initiatives for reporting standards include Consolidated Standards of Reporting Trials for clinical trial data and Enhancing the Quality and Transparency of Health Research network for health research.

Such reporting standards should impact both research and clinical decision making. Creating best practices for data collection, analysis, and reporting in research will directly influence clinical decision making, thus improving evidence-based practice. Accordingly, the establishment of best practice guidelines for spoken discourse analysis in aphasia will not only improve the efficiency, consistency, and quality of research but also provide well-founded recommendations for speech-language pathologists to meaningfully utilize spoken discourse assessment in guiding treatment planning and achieving optimal outcomes for individuals with aphasia, ultimately enhancing their quality of life.

Research Gap #2: Understanding Psychometric Properties of Spoken Discourse Outcomes

Due to the immense number of reported spoken discourse outcomes, very little is known about their psychometric properties (e.g., validity, reliability; Pritchard et al., 2017). Reliability is the ability to reproduce a result consistently in time and space and comprises different components, including stability, internal consistency, and equivalence (Pritchard et al., 2018). Validity refers to the property of an instrument to measure exactly what it proposes and comprises components such as content, criterion, and construct validity. An outcome's psychometric properties are important for research, clinical practice, and health assessment because they allow for identification of the best assessment tools. For spoken discourse, this means that researchers and clinicians will be empowered to select the most sensitive and robust outcomes to assess and treat aphasia.

Some studies have evaluated the psychometric properties of spoken discourse measures (e.g., Boyle, 2014; Brookshire & Nicholas, 1994; Capilouto et al., 2006; Kong, 2009; McNeil et al., 2001, 2002; Nicholas & Brookshire, 1993, 1995). However, the number of studies is limited, and the participant sample sizes on which these properties have been calculated are often small. For example, Brookshire and Nicholas (1994) evaluated the test–retest stability of two measures of connected speech (i.e., words per minute and percent correct information units) in 20 individuals with aphasia and 20 neurotypical adults. This study, while small, was critical in establishing the notion that more speech (specifically, output elicited by more than one task, and of at least 300–400 words) increased the test–retest stability of these measures. In a more recent study, Boyle (2015) examined the test–retest reliability of word retrieval measures in the narrative language samples of persons with aphasia. For the individual picture stimuli from the AphasiaBank stimuli, she found poor test–retest reliability for measures of word retrieval errors; however, combined analysis of the different narrative tasks yielded some relatively stable measures (e.g., semantic and phonological errors). Relatedly, Pritchard et al. (2018) provided select psychometric information, such as acceptability, validity, and rater reliability, on some spoken discourse outcomes in aphasia (e.g., story grammar, coherence, sentence structure). There has been a trend toward improved psychometric reporting, as highlighted by recent work (e.g., Kim et al., 2019). Future research, however, is needed to establish the psychometric quality of the existing micro- and macrolinguistic spoken discourse measures that will inform clinicians and researchers involved in spoken discourse analysis in aphasia.

Moving Forward: Establishing a Working Group

The goal of a roundtable entitled "Standardizing Assessment of Spoken Discourse in Aphasia: Directions for Future Research" at the 49th Clinical Aphasiology Conference was to identify current issues related to collecting and analyzing spoken discourse in aphasia (Dutta et al., 2019). Following discussion at this roundtable and with the coauthors, the working group *FOQUSAphasia* ("**FO**stering **QU**ality of **S**poken discourse in **Aphasia**") was created. The structure of this working group is proposed in Figure 1.

In general, the FOQUSAphasia group has a relatively flat hierarchy. Individuals will self-select to join a task force (or multiple task forces) within FOQUSAphasia ("Members-at-Large"). Within each task force are initiatives, which are the task force's main goals. Task forces are led by a Leadership Team, which is nominated and voted upon by the Members-at-Large who belong to the task force. An overall Steering Committee, nominated and voted upon by all members across task forces, guides FOQUSAphasia, keeping its task forces and initiatives on track and in line with the field's needs and wants. We envision a dynamic partnership between researchers, clinicians, and stakeholders (persons with aphasia and care providers). We foresee the Steering Committee and Leadership Team of each task force engaging with stakeholders to inform the direction of task forces and their initiatives and to brainstorm new task forces and initiatives. It is the goal of the Steering Committee to hold quarterly, open-to-all virtual meetings.

Presently, FOQUSAphasia will begin with two task forces. The first, "Best Practices," will focus on evaluating and improving field standards with respect to analyzing and reporting discourse-related data in aphasia (see Figure 2). The second, "Methodology and Data Quality," will improve the current state of data quality in spoken discourse in aphasia (see Figure 3). Within each of these task forces is an initial initiative. For the "Best Practices" Task Force, this initiative will work toward creating best practices in discourse collection, analysis, and reporting by collecting information related to current practices, with an emphasis on the usage of psychometric data. For the "Methodology and Data Quality" Task Force, this initiative will create a test–retest reliability database. Each task force and their initial initiatives are discussed below.

Joining FOQUSAphasia

Following the Clinical Aphasiology Conference roundtable discussion, many attendees self-identified as being interested in FOQUSAphasia membership. Our website (http://www.foqusaphasia.com) also encourages membership. We will advertise FOQUSAphasia and its mission to accrue members and identify their task force interest(s). Membership will be open to researchers and clinicians

Figure 1. A proposed model of the structure of the working group, Focusing on Quality of Spoken Discourse in Aphasia (FOQUSAphasia).

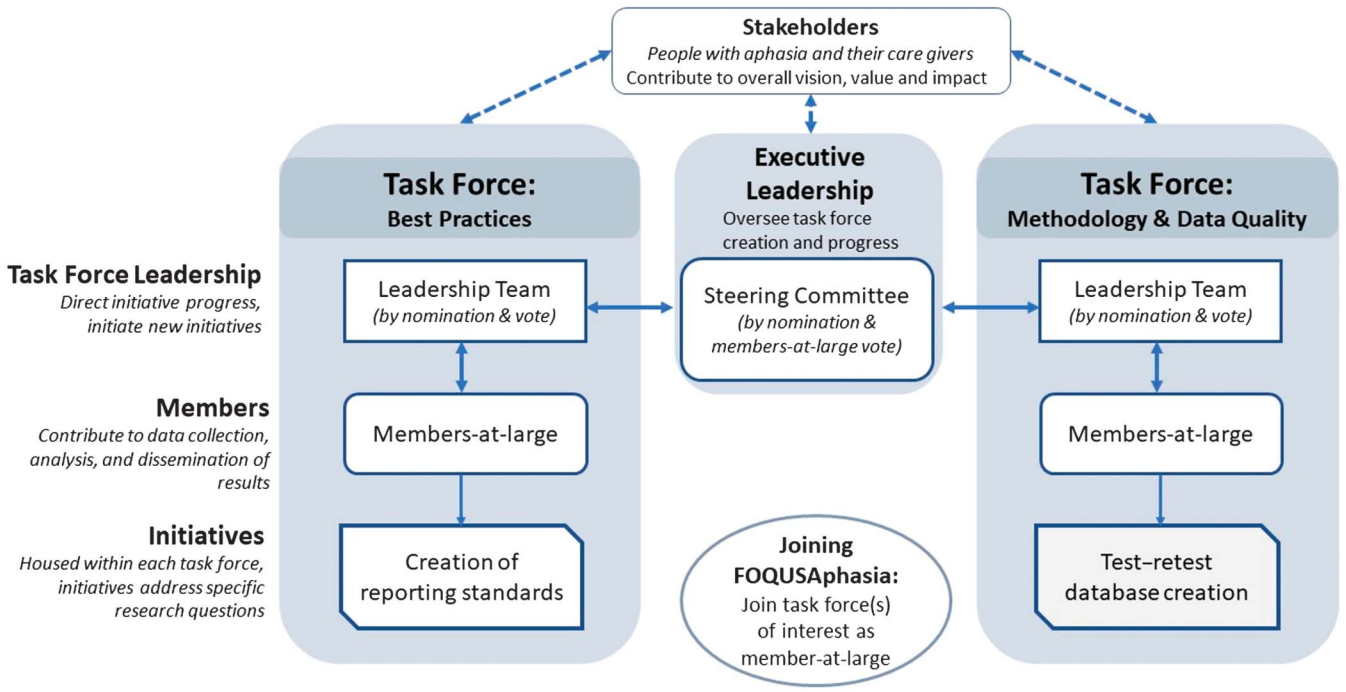
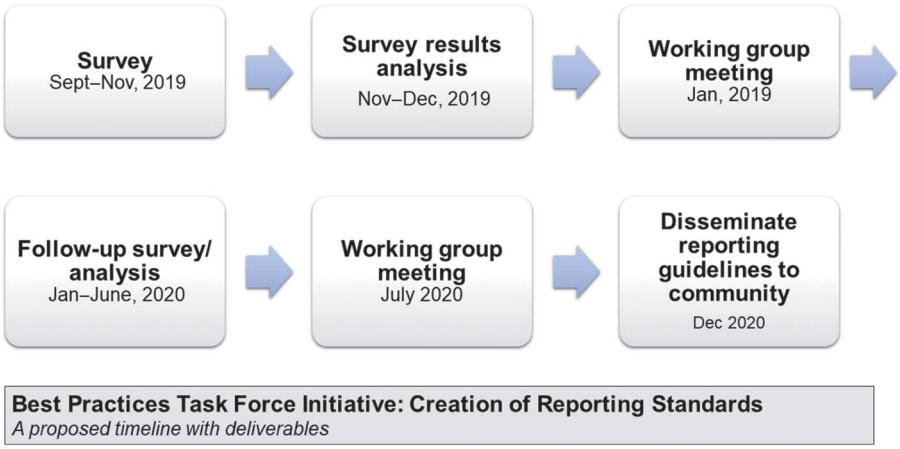


Figure 2. A proposed timeline for the Best Practices Task Force Initiative, "Creation of Reporting Standards."



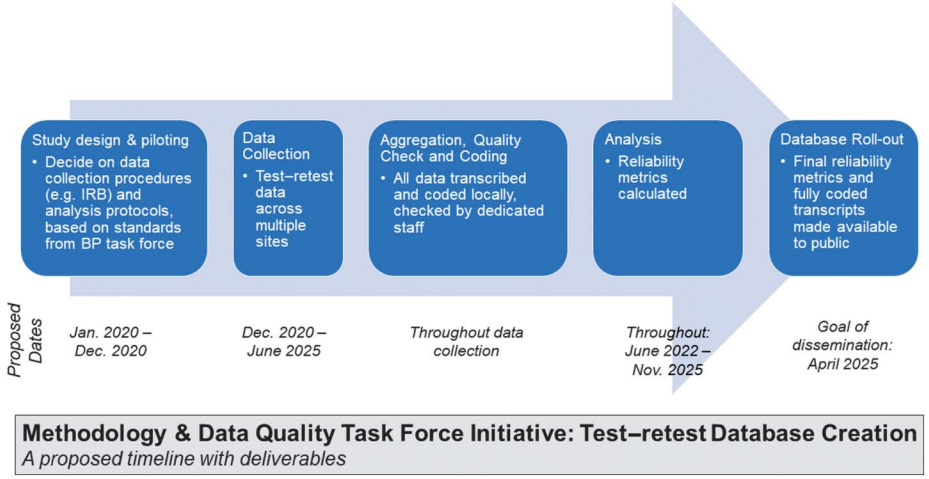
(e.g., speech-language pathologists) with expertise, experience, and interest in the assessment of spoken discourse in aphasia as the early initiatives of the group are directed toward assessing best practice, and methodological and data quality, as outlined below. Membership is likewise open to stakeholders, such as individuals with aphasia and their family members and caregivers. In later stages, as we expect the attention of FOQUSAphasia to shift to functional outcome measures, interfacing with stakeholders will be especially crucial. Procedures involving structure and voting will be available on our website (https://www.foqusaphasia.com).

Best Practices Task Force

Initiative: Creation of Reporting Standards

Reporting standards do not currently exist for the field of spoken discourse analysis in aphasia. However, an attempt to establish basic assessment and treatment fidelity guidelines has been noted in the related field of spoken language sample analysis in child language development and disorders (Finestack et al., 2014; Gearing et al., 2011) and aphasia (e.g., Richardson et al., 2016). Spoken discourse analysis is frequently conducted in both clinical and research settings. Given the heterogeneity of settings, spoken discourse analysis comes with considerations for data collection, analysis, and dissemination that are not otherwise found in guidelines currently available for clinical trials (e.g., Consolidated Standards of Reporting Trials) or health studies (e.g., Enhancing the Quality and Transparency of Health Research). Following are some examples of elements within a research report that require reporting standards: (a) specific demographics (e.g., age, education), language-specific variables (e.g., monolingual vs. bilingual, degree of proficiency), months postonset of brain injury,

Figure 3. A proposed timeline for the Methodology & Data Quality Task Force Initiative, "Creation of a Test–Retest Database." IRB = institutional review board; BP = best practices.



type and frequency of brain injury, presence/absence of aphasia, aphasia type, and severity; (b) the procedures used to examine speech-language and cognitive abilities; (c) the environment in which the spoken discourse data were collected (e.g., a sound booth, hospital room); (d) the manner of spoken discourse elicitation (e.g., training, experience, and qualifications of the person eliciting the sample, the type of sample and elicitation procedure used, the length of the sample elicited); and (e) rater and analysis information, including rater experience, training, qualifications, how spoken discourse was transcribed, segmented, and coded, what software and techniques were used, and information regarding transcription and coding reliability. These examples underscore how spoken discourse work requires certain mandatory reporting for replication of and comparison across studies. Further examples of mandatory reporting standards, organized by study section, are provided in Table 1. These examples stem from the Dutta et al. (2019) roundtable, where the italicized text are proposed additions to an existing best practice document (COBIDAS) that would be required for spoken discourse.

Given the rapid growth in research evaluating spoken discourse in aphasia and the current state of that literature, the creation of reporting standards will (a) encourage replication of studies and, thus, combat the replication crisis in the behavioral and social sciences (Dietz & Boyle, 2018); (b) ensure consistent reporting across studies; (c) recommend appropriate statistical modeling, thereby ensuring the most appropriate statistical inferences; and, (d) overall, contribute to a more homogeneous, rigorous, and standardized process by which spoken discourse research is evaluated and ultimately disseminated. This in turn will facilitate meta-analyses and lead to a higher level of evidence in the field of spoken discourse analysis. Importantly, a more homogeneous and rigorous research standard will have direct clinical implications: Creating guidelines for reporting standards will improve best practices for collecting, analyzing, and accurately interpreting changes in spoken discourse outcomes in aphasia.

Specific Task Force Objectives

The following are the objectives:

1. Acquire data on spoken discourse collection and analysis methods used by clinicians and researchers working in aphasia assessment and rehabilitation.

Goal: Collect data from professionals actively working in the area of spoken discourse assessment in aphasia regarding commonly used collection and analysis methods.

Approach: We aim to recruit at least 100 respondents from a variety of geographical locations (e.g., United States, United Kingdom, Australia), roles (e.g., speech-language pathologist, university-based researcher), and settings (e.g., hospital, outpatient clinic, university). A survey was created by the first three authors of this research note and then piloted among all co-authors. Ethics approval for survey dissemination was acquired from Indiana University, and at the end of August 2019, the survey was shared widely via social media (e.g., Facebook, Twitter), e-mail, lab webpages, and related networks (e.g., American Speech-Language-Hearing Association Special Interest Groups). The survey was distributed in English and closed mid-November 2019 for response analysis. See Table 2 for example questions.

2. Best Practice task force meeting to discuss standards for spoken discourse collection, analysis, and reporting procedures.

Goals: (a) Discuss survey results and decide on fundamental and recommended standards for reporting spoken discourse collection and analysis in aphasia and (b) identify if there is a need for a follow-up survey.

Approach: At the first virtual meeting, a leadership team (see Figure 1) for this Creation of Reporting Standards initiative will be selected. This leadership team will be in charge of aggregating data (from survey findings and a systematic literature review), formulating a plan for future Best Practices task force meetings, creating deliverables (e.g., further surveys), and writing and maintaining a best practices document (see Figure 2 for the proposed timeline).

Methodology & Data Quality Task Force

Initiative: Creation of Test–Retest Database

Certain psychometric properties are valuable and essential for quality clinical research and practice. Stability is the inherent variance (due to internal or external factors) of an outcome/measure (Tate, 2010). Establishing an outcome's degree of stability allows researchers to draw conclusions about clinically meaningful changes. Short interval sampling —that is, testing and retesting within a relatively short window of time (e.g., 2 weeks)—can determine the variability of the participant's baseline performance. A measure that varies widely within participants during a short interval is likely not stable enough to be used as a clinically meaningful outcome or assessment measure. While such short interval sampling may not be practical in clinical practice, it is necessary in a research context to determine which measures are suitably stable to be effective in the clinical assessment of aphasia. Hence, FOQUSAphasia will collect test–retest spoken discourse data in individuals with chronic nonprogressive aphasia, at an interval of 7 ± 3 days, in line with prior studies (e.g., Brookshire & Nicholas, 1994).

Assessing and reporting such reliability metrics is essential for health-related research to validate the frequently used assessment measures in clinical and research settings (Meek et al., 2000; Squires et al., 2011). Stability of a measure can be quantified absolutely (i.e., the consistency of individuals' scores across time points) or relatively (i.e., the consistency of an individual's position/rank relative to other group members). Absolute consistency is quantified most often using standard error of measurement (*SEM*), whereas relative consistency is most often quantified using ICCs (Cicchetti, 1994), Pearson *r*, and/or Cronbach's alpha (Weir, 2005).

Table 1. Sample checklist for best practices and spoken discourse measurement items to report.

Section	Item	Notes
Experimental design	Participants participated and analyzed	Provide the number of participants tested, number excluded after testing, and number included in the data analysis. If they differ, note the number of participants in each particular analysis
	Inclusion criteria and descriptive statistics	Provide age (mean, standard deviation, range), gender, sex (absolute counts or relative frequencies), education and/or socioeconomic status (specify measurement used), *presence of aphasia (type of test used to document this, test score including mean/percentile when possible), aphasia type, months postonset, type and frequency of injury or disease, native language, number of languages spoken, and proficiency in those languages. Also report demographic variables for conversational partners if applicable*
Data acquisition	Experiment preparation	*Equipment used (e.g., videography information, audio information), environment (e.g., sound booth), software information (e.g., Psychopy, Systematic Analysis for Language Transcripts)*
	Behavior acquisition	*Types of prompt used (including specific instructions from experimenter, preferably included in supplement), amount of speaking time allotted per prompt (in seconds or minutes), materials used (e.g., picture book, video clip)*
	Rationale for dependent variables	*Provide rationale for choice of dependent variables/outcomes (e.g., why was mean length of utterance evaluated?), provide psychometric properties of outcomes when available (e.g., validity, reliability)*
Preprocessing of data	Transcription	*How was transcription completed (e.g., from video, from audio, live), who did the transcription (e.g., students), specifying educational background and training, if training was provided describe the training (e.g., include supplementary training documents or refer to freely available training on a website)*
	Coding	*Specify coding system used, if any (e.g., CHAT/CLAN, Praat)*
	Reliability	*Report reliability of transcription (e.g., who, how, when), reliability of coding (e.g., who, how, when), and the statistical metrics employed to test reliability (including rationale), and specify both interrater and intrarater reliability*
Statistical modeling & inference	Mass univariate analyses	Report the number of time points and participants; specify exclusions of time points and participants, if not already specified in the experimental design; specify independent and dependent variables as well as covariates
	Multivariate modeling & predictive analysis	Specify variable type (discrete or continuous), classification settings, population stratification, and model used
Results reporting	Effects tested	Provide a complete list of tested and omitted effects, provide table of major findings
Data sharing	Material sharing	List types of data provided (e.g., audio-only data, transcripts or coded data only) and where these data are available (e.g., freely on website, by contacting author), report on completeness of data (e.g., number of participants for whom all data are available)

Note. The italicized text are proposed additions to an existing best practice document (i.e., the Committee on Best Practice in Data Analysis and Sharing) that would be required for spoken discourse. CHAT = Codes for the Human Analysis of Transcripts; CLAN = Computerized Language Analysis.

Understanding stability of outcome measures at test–retest is especially important in aphasia because it is well established that typical speakers without acquired brain injury demonstrate intra- and interindividual variability in micro- and macrostructural discourse outcomes between test and retest (Armstrong, 2000). Therefore, it is not surprising that individuals with aphasia also demonstrate such variability, given that performance variability is a hallmark of brain injury and aphasia (Goodglass, 1993). When referring to "outcome stability," we are specifically referring to the range of reliability of an outcome measure for a given group. For instance, in a large group of participants with aphasia, the outcome stability may range from very stable (high reliability metrics) to not stable (low reliability metrics) depending on the measure. We can also think of outcome stability as being variable at the group level (e.g., between speakers with aphasia and without aphasia) or, indeed, between speakers with different types or severities of aphasia. Speakers with aphasia may demonstrate lower, on average, stability of a spoken discourse outcome in comparison to individuals without aphasia.

Outcome stability is also helpful in selecting study design. As an example, let us say that the intra-individual test–retest stability of a frequently used spoken discourse outcome, words per minute, is 10–30 words per minute in a large population of speakers with aphasia. As this is quite a wide range, the researcher may note that this is not a very reliable or stable outcome. Therefore, the researcher might choose a different measure (if indeed there is a comparable, more stable metric to measure a similar language construct) or may devise a design that is more robust to less stable measures, such as a single-subject design. Furthermore, understanding an outcome's intra-individual stability can influence interpretation of an intervention study, especially at a case study level. For instance, referring back to our example of words per minute, let us suggest that the *SEM*,

Table 2. Examples of survey questions from the Best Practices Task Force.

Section	Questions
Demographic information	1. How would you describe yourself? (Mark all that apply) □ Researcher □ Academic/teacher □ Speech-language therapist/pathologist □ Student □ Other (please specify): _____ 2. In which country are you currently practicing and/or doing research? □ United States of America □ United Kingdom □ Australia □ New Zealand □ Canada □ Other (please specify): _____
Spoken discourse measurement in aphasia	1. How often do you collect and analyze spoken discourse samples in aphasia assessment and treatment? □ Never □ Rarely □ Sometimes □ Usually □ Always □ N/A 2. Why do you collect spoken discourse data? (Mark all that apply) □ To gain information regarding aphasia symptoms □ As an outcome measure for aphasia treatment □ As a part of a research study investigating language profiles in aphasia □ Other (please specify): _____
Data collection procedures	1. Where do you collect the spoken discourse samples? (e.g., quiet room, during a therapy session) (Mark all that apply) □ Sound booth □ Quiet room □ A hospital or rehab facility room with typical daily distractions (e.g., background noise) □ Participant's home □ Other (please specify): _____ 2. Who records the spoken discourse samples? (Mark all that apply) □ Researcher □ Research assistant □ Graduate student □ Undergraduate student □ Speech-language therapist/pathologist □ Not applicable □ Other (please specify): _____
Data analysis procedures	1. Once the discourse data are collected, what steps are undertaken? (Mark all that apply) □ Listen to the recorded samples □ Transcribe samples verbatim □ Code transcripts (e.g., coding for paraphasic errors) □ Perform detailed analysis of transcripts (e.g., lexical-syntactic analysis using CLAN software) □ Perceptual judgment-based analysis (e.g., rating of fluency or informativeness) □ Make clinical judgment of language ability □ Other (please specify): _____ 2. What are the barriers to collecting psychometric data? (Mark all that apply) □ Time □ Funds □ Personnel □ Knowledge and training regarding collecting psychometric information (e.g., statistical analysis to use) □ Other (please specify): _____

Note. N/A = not applicable; CLAN = Computerized Language Analysis.

derived from a large group of speakers with aphasia, was 27 words per minute. That is, for a treatment to change words per minute beyond standard error, the improvement at outcome would need to be ± 27 words compared to baseline. Similarly, statistical measures such as minimal detectable change (MDC) suggest how clinically meaningful a change may be. To determine the minimum change necessary to ensure a confidence level of 90% that a change would be unrelated to measurement error, one can calculate MDC90: $SEM \times 1.65 \times \text{sqrt}2$. MDC90 is the level recommended for decisions regarding intervention effectiveness in rehabilitation research (Donoghue et al., 2009). Continuing with our example of words per minute, if the observed outcome score was a change of 40 words per minute from

baseline, and we estimated *SEM* = 27 (again, drawn from a larger study), MDC90 = 63. Therefore, a change of at least 63 words per minute would be needed to interpret the change as clinically meaningful. When an outcome is more stable, it follows that *SEM* will be smaller and MDC90 a lower number.

Relative to a group, one would consider an outcome "stable" if ICC > .7, and optimally ICC > 0.9 for clinical, health-related outcomes at the individual level (Fitzpatrick et al., 1998). A lower ICC suggests greater relative variability and, as such, would make identifying meaningful change difficult without a sufficiently large sample size and high-quality data. Therefore, for these types of studies (e.g., cohort studies, which may measure change at the relative level), choosing an outcome measure with well-established, high stability is crucial.

Data from a large, randomly selected and representative reference population establishes a baseline distribution for a score or measurement, and a benchmark against which the score or measurement can be compared. At the moment, no such normative data exist for spoken discourse in aphasia, standing in contrast to the frequently used standardized aphasia assessments, which rely on normative information to compare an individual's scores with those from a certain population. The value of normative data of spoken discourse is potentially great. For example, they may provide more sensitive measures for mild aphasia, which may be missed by current popular aphasia tests. Fluency is multidimensional, comprising aspects of language fluency (e.g., lexical access) and motor fluency (e.g., planning and execution); unfortunately, most standardized tests do not quantify measures of language fluency, such as frequency and types of phonemic paraphasias or pause frequency, type (e.g., filled, unfilled), and duration. Standardized language tests also often underestimate connected speech capability due to their focus on relatively simple language and the use of isolated tasks such as confrontation naming, and single word and phrase repetitions (Fromm et al., 2017; Murray & Clark, 2015). Furthermore, these standardized tests tend to rely on high-frequency objects for picture naming, which may not reflect an accurate representation of lexical access in aphasia (Gagnon et al., 1997).

Therefore, a concerted effort must be made to collect test–retest spoken discourse data from speakers with and without aphasia to identify the stability of spoken discourse outcomes and how they may vary across prompt type and participant groups. Additionally, given the lack of normative data for certain spoken discourse measures, future work must focus on collecting such data at various sites to identify outcomes that are more sensitive to specific populations (e.g., stroke rehabilitation unit vs. long-term care facility). The collection of a large database of language samples will allow the calculation of multiple discourse measures from the same data set, permitting direct comparison among spoken discourse measures so that researchers are able to recommend the most valid and reliable measures for clinical use.

The scarcity of test–retest data for discourse measures is unsurprising given time constraints, limited population sizes at single sites, and costs associated with bringing participants back for testing at a later date (Pritchard et al., 2018). Having an accessible database of test–retest spoken discourse samples from speakers with and without aphasia is the logical step to address these issues. There is already a platform to host such data, AphasiaBank (MacWhinney et al., 2011), an online database of spoken discourse samples collected via a standard protocol, along with demographic and cognitive–linguistic information from individuals with and without aphasia. Currently, the database includes mostly cross-sectional data from over 300 speakers with aphasia and 181 speakers without aphasia. Transcripts, videos, and other participant-related materials are password restricted to AphasiaBank members (membership is granted upon request to licensed clinicians and faculty). Given that AphasiaBank is a resource already widely used by clinicians and researchers, it can serve as a convenient platform to host test–retest data and enhance our understanding of spoken discourse outcome stability.

The overarching aim of the Methodology & Data Quality (MDQ) task force initiative is to collect and publish test–retest data and associated stability metrics for frequently reported spoken discourse outcomes, capturing variability within and across speakers with and without aphasia. To do so, we will build an easily searchable interface of statistical metrics (e.g., ICC, Pearson *r*) for spoken discourse outcomes for use by clinicians and researchers. The intent of this database is to provide variability estimates across discourse prompts and speakers. That is, the database will be set up to filter stability metrics by prompt type (e.g., story retell, procedural), presence or absence of aphasia, aphasia severity and type, time postonset, demographics (e.g., age), and cognitive–linguistic variables (e.g., verbal and nonverbal fluency, attention scores). Specific variables will be decided upon by the initiative members (see below). Notably, while this information will help to design better studies in the future, a database of test–retest variability on frequently used outcomes will also be useful retrospectively, expounding on intervention effects in completed studies. Overall, the information collected in the database will allow for a degree of uniformity across the field, with the ultimate goal of improving evidence-based decision making for selecting psychometrically sound measures and creating normative standards for assessing spoken discourse outcomes in aphasia assessment, rehabilitation, and research.

Specific Objectives

List of objectives:

1. Design study, including data collection and analysis.

Goal: Based on recommendations and ongoing work from the Best Practices task force, members of the MDQ task force will design the test–retest study.

Approach: Group members will (a) identify data collection sites; multiple sites allow accounting for factors that may influence spoken discourse production (e.g., environmental, personal, geographic); (b) create a shared-site

institutional review board to set up a data sharing agreement with Indiana University; (c) create protocols for aggregating and sharing data with central location (Indiana University); (d) design protocols for data collection (e.g., demographics, spoken discourse tasks, other cognitive–linguistic information), including specification of inclusion/exclusion parameters for participants, and ensure consistency across sites; (e) decide on outcome measures, which will be based on the Best Practices task force's survey and previous research (e.g., Boyle, 2014; Bryant et al., 2016); (f) determine analysis procedures (e.g., analyses at the level of inter- and intra-individual) and steps to preregister analysis; and (g) design and build the database.

2. Collect test–retest spoken discourse data.

Goal: To collect a test–retest spoken discourse data set from adults with and without aphasia following the above study design.

Approach: We will collect test–retest spoken discourse data from 250 speakers without aphasia and 250 speakers with aphasia using the same discourse stimuli at both time points ("retesting" window will be within 7 ± 3 days of initial test). For the purposes of the current project, we will collect data from speakers with aphasia resulting from a stroke. As per the AphasiaBank protocol, necessary demographics from all speakers, including language status (e.g., monolingual, other languages known), chronicity (e.g., time poststroke), and injury information (e.g., type and number of strokes) will be collected.

Power Analysis for Sample Size

The rationale for the proposed sample size was threefold:

1. To identify reliability of outcomes between test–retest, sample size was based on a power analysis evaluating Cronbach's alpha. With 95% confidence and 80% power, using an acceptable Cronbach's alpha of .70 and a conservative estimation of actual measurement of Cronbach's alpha being .5, with measurement occurring at two time points and factoring in 10% attrition rate, we would need approximately 140 members per group (i.e., healthy control, aphasia). This value is in line with estimates to establish ICC (Walter et al., 1998), which would be an estimate of 70 per group when an acceptable .7 ICC, expected .5 ICC, two time points, and 10% attrition.
2. We are also interested in being able to model the extent to which there is a significant difference in outcomes between test and retest (e.g., within-individual stability). To do that, we estimated the sample size for a dependent *t* test, given two tails, effect size of 0.2, 95% confidence, and 80% power. With these parameters, factoring in 10% attrition, we would need approximately 220 members per group to identify a significant difference in outcomes at retest from test.
3. To compare score variability between healthy control and aphasia speaker groups, we would ideally compare these two groups' Pearson *r* values (i.e., between test and retest). Assuming a small effect size between groups (0.3), 95% confidence, 80% power, and two tails, we would need approximately 180 people per group to identify a significant difference. These estimations were based on a small pilot study conducted on AphasiaBank data (Dutta et al., 2019).

At the first virtual MDQ task force meeting, a leadership team (see Figure 1) will be selected. This leadership team will create a shared ethics/institutional review board template for interested data collection sites, and aggregate and analyze data. Task force members will be based at various sites and will institute their own institutional review board protocol. Members will self-identify the level of their involvement (e.g., being a data collection site for only nonaphasic speakers). Data will be transcribed and coded using Codes for the Human Analysis of Transcripts (CHAT)/Computerized Language Analysis (CLAN; MacWhinney, 2000; MacWhinney et al., 2011) at a single location (Indiana University), to facilitate reliability between and across raters and consistency of data transcription and coding. Notably, the MDQ task force will acquire data and report on outcomes based on recommendations from the Best Practices task force. This highlights the integration of task forces.

3. Dissemination of test–retest data set and open source data availability.

Goal: The test–retest data set, including audiovisual data and finalized stability metrics (e.g., ICCs, *SEM*, MDC) will be made available in a special repository hosted on AphasiaBank.

Approach: Standard metrics will be computed on the data by trained statisticians. As there are many outcomes that can be derived from this rich data set, we will compute metrics on the most frequently used micro- and macrostructural spoken discourse outcomes (Bryant et al., 2016). Because the raw data will also be made available (in CHAT/CLAN format), researchers and clinicians will be encouraged to compute statistics on outcomes of their choosing and to add these to the database. Thus, the database's growth will be driven by the needs of its users. For outcomes with statistics computed, we will also develop an online interface that allows users to filter data, such as by presence/absence of aphasia, prompt type, aphasia type, and demographics (e.g., age). Demographics and other cognitive–linguistic information will be made available for all speakers included in the database. A proposed timeline for the MDQ task force and its first initiative ("Creation of Test–Retest Database") is shown in Figure 3.

Future Directions

The FOQUSAphasia working group is meant to serve as a foundation for the creation of task forces and/or initiatives that will improve standards of research and reporting of research in spoken discourse in aphasia. As

such, this working group will be driven by the goals and needs of clinicians, researchers, and stakeholders. We envision that future FOQUSAphasia directions may include (a) collection of data to establish standardized databases for spoken discourse data in individuals with aphasia in the acute phases and due to nonstroke (e.g., traumatic brain injury) or progressive etiologies (e.g., dementia), (b) evaluation of other critical psychometric properties of spoken discourse (e.g., validity, acceptability), (c) best practices in the collection and analysis of less-structured and more complex forms of discourse (e.g., conversation), and (d) improving automatic transcription and coding of spoken discourse.

Conclusion

The goal of FOQUSAphasia is to improve the state of research in spoken discourse in aphasia and allow discourse to be added to the common outcome elements, thus improving the application of research in aphasia that goes beyond the single-word and sentence levels of processing. Findings from this research will facilitate evidence-based practice in the field of aphasia. In addition to identifying psychometrically reliable discourse outcomes, clinicians who are involved in spoken discourse measurement in aphasia will be informed about more systematic and standardized ways of assessing and analyzing spoken discourse that will allow accurately capturing the communication difficulties faced by those with aphasia and document aphasia treatment–related changes. Any interested parties are encouraged to contact this research note's first author to join one or both task forces. The intent is to report on the progress of FOQUSAphasia at each Clinical Aphasiology Conference and in relevant publications. More information about this working group can be found on our website, http://www.foqusaphasia.com.

Acknowledgments

This study was funded by the ASHFoundation New Investigator Award (awarded to B. C. Stark).

References

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

Beeke, S., Maxim, J., & Wilkinson, R. (2007). Using conversation analysis to assess and treat people with aphasia. *Seminars in Speech and Language, 28*(2), 136–147. https://doi.org/10.1055/s-2007-970571

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

Brady, M. C., Kelly, H., Godwin, J., Enderby, P., & Campbell, P. (2016). Speech and language therapy for aphasia following stroke. *Cochrane Database of Systematic Reviews, 6,* CD000425. https://doi.org/10.1002/14651858.CD000425.pub4

Brookshire, R. H. (1983). Subject description and generality of results in experiments with aphasic adults. *Journal of Speech and Hearing Disorders, 48*(4), 342–346. https://doi.org/10.1044/jshd.4804.342

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

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

Capilouto, G. J., Wright, H. H., & Wagovich, S. A. (2006). Reliability of main event measurement in the discourse of individuals with aphasia. *Aphasiology, 20*(2–4), 205–216. https://doi.org/10.1080/02687030500473122

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Clinical Centre for Research Excellence (CCRE) in Aphasia Rehabilitation. (2014). *Aphasia rehabilitation best practice statements. Comprehensive supplement to the Australian aphasia rehabilitation pathway*. CCRE in Aphasia Rehabilitation.

Damico, J. S., Oelschlaeger, M., & Simmons-Mackie, N. (1999). Qualitative methods in aphasia research: Conversation analysis. *Aphasiology, 13*(9–11), 667–679. https://doi.org/10.1080/026870399401777

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., Physiotherapy Research and Older People (PROP) group., & 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

Dutta, M., Murray, L., & Stark, B. C. (2019). *Standardizing assessment of spoken discourse in aphasia: Directions for future research*. Paper presented at the Clinical Aphasiology Conference, Whitefish, MT, United States.

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

Finestack, L. H., Payesteh, B., Disher, J. R., & Julien, H. M. (2014). Reporting child language sampling procedures. *Journal of Speech, Language, and Hearing Research, 57*(6), 2274–2279. https://doi.org/10.1044/2014_JSLHR-L-14-0093

Fitzpatrick, R., Davey, C., Buxton, M., & Jones, D. (1998). Evaluating patient-based outcome measures for use in clinical trials: A review. *Health Technology Assessment, 2*(14), 1–73. https://doi.org/10.3310/hta2140

Fromm, D., Forbes, M., Holland, A., Dalton, S. G., Richardson, J., & MacWhinney, B. (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

Gagnon, D. A., Schwartz, M. F., Martin, N., Dell, G. S., & Saffran, E. M. (1997). The origins of formal paraphasias in aphasics' picture naming. *Brain and Language, 59*(3), 450–472. https://doi.org/10.1006/brln.1997.1792

Galski, T., Tompkins, C., & Johnston, M. V. (1998). Competence in discourse as a measure of social integration and quality of

life in persons with traumatic brain injury. *Brain Injury, 12*(9), 769–782. https://doi.org/10.1080/026990598122160

Gearing, R. E., El-Bassel, N., Ghesquiere, A., Baldwin, S., Gillies, J., & Ngeow, E. (2011). Major ingredients of fidelity: A review and scientific guide to improving quality of intervention research implementation. *Clinical Psychology Review, 31*(1), 79–88. https://doi.org/10.1016/j.cpr.2010.09.007

Goodglass, H. (1993). *Understanding aphasia*. Academic Press.

Kim, H., Kintz, S., Zelnosky, K., & Wright, H. H. (2019). Measuring word retrieval in narrative discourse: Core lexicon in aphasia. *International Journal of Language & Communication Disorders, 54*(1), 62–78. https://doi.org/10.1111/1460-6984.12432

Kintz, S., & Wright, H. H. (2018). Discourse measurement in aphasia research. *Aphasiology, 32*(4), 472–474. https://doi.org/10.1080/02687038.2017.1398807

Kong, A. P.-H. (2009). The use of main concept analysis to measure discourse production in Cantonese-speaking persons with aphasia: A preliminary report. *Journal of Communication Disorders, 42*(6), 442–464. https://doi.org/10.1016/j.jcomdis.2009.06.002

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

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

McNeil, M. R., Doyle, P. J., Fossett, T. R. D., Park, G. H., & Goda, A. J. (2001). Reliability and concurrent validity of the information unit scoring metric for the story retelling procedure. *Aphasiology, 15*(10–11), 991–1006. https://doi.org/10.1080/02687040143000348

McNeil, M. R., Doyle, P. J., Park, G. H., Fossett, T. R. D., & Brodsky, M. B. (2002). Increasing the sensitivity of the story retell procedure for the discrimination of normal elderly subjects from persons with aphasia. *Aphasiology, 16*(8), 815–822. https://doi.org/10.1080/02687030244000284

Meek, P. M., Nail, L. M., Barsevick, A., Schwartz, A. L., Stephen, S., Whitmer, K., Beck, S. L., Jones, L. S., & Walker, B. L. (2000). Psychometric testing of fatigue instruments for use with cancer patients. *Nursing Research, 49*(4), 181–190. https://doi.org/10.1097/00006199-200007000-00001

Murray, L. L., & Clark, H. M. (2015). *Neurogenic disorders of language and cognition: Evidence-based clinical practice*. Pro-Ed.

Nicholas, L. E., & Brookshire, R. H. (1993). A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *Journal of Speech and Hearing Research, 36*(2), 338–350. https://doi.org/10.1044/jshr.3602.338

Nicholas, L. E., & Brookshire, R. H. (1995). Presence, completeness, and accuracy of main concepts in the connected speech of non-brain-damaged adults and adults with aphasia. *Journal of Speech and Hearing Research, 38*(1), 145–156. https://doi.org/10.1044/jshr.3801.145

Nichols, T. E., Das, S., Eickhoff, S. B., Evans, A. C., Glatard, T., Hanke, M., Kriegeskorte, N., Milham, M. P., Poldrack, R. A., Poline, J.-B., Proal, E., Thirion, B., Van Essen, D. C., 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

Richardson, J. D., Hudspeth Dalton, S. G., Shafer, J., & Patterson, J. (2016). Assessment fidelity in aphasia research. *American Journal of Speech-Language Pathology, 25*(4S), S788–S797. https://doi.org/10.1044/2016_AJSLP-15-0146

Sim, P., Power, E., & Togher, L. (2013). Describing conversations between individuals with traumatic brain injury (TBI) and communication partners following communication partner training: Using exchange structure analysis. *Brain Injury, 27*(6), 717–742. https://doi.org/10.3109/02699052.2013.775485

Squires, J. E., Estabrooks, C. A., O'Rourke, H. M., Gustavsson, P., Newburn-Cook, C. V., & Wallin, L. (2011). A systematic review of the psychometric properties of self-report research utilization measures used in healthcare. *Implementation Science, 6*(1), 83. https://doi.org/10.1186/1748-5908-6-83

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

Tate, R. L. (2010). *A compendium of tests, scales, and questionnaires: The practitioner's guide to measuring outcomes after acquired brain impairment*. Psychology Press.

Ulatowska, H. K., Allard, L., Reyes, B. A., Ford, J., & Chapman, S. (1992). Conversational discourse in aphasia. *Aphasiology, 6*(3), 325–330. https://doi.org/10.1080/02687039208248602

Wallace, S. J., Worrall, L., Rose, T., & Le Dorze, G. (2017). Which treatment outcomes are most important to aphasia clinicians and managers? An international e-Delphi consensus study. *Aphasiology, 31*(6), 643–673. https://doi.org/10.1080/02687038.2016.1186265

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

Walter, S. D., Eliasziw, M., & Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in Medicine, 17*(1), 101–110. https://doi.org/10.1002/(SICI)1097-0258(19980115)17:1<101::AID-SIM727>3.0.CO;2-E

Weir, J. P. (2005). Quantifying test–retest reliability using the intraclass correlation coefficient and the *SEM*. *Journal of Strength and Conditioning Research, 19*(1), 231–240. https://doi.org/10.1519/00124278-200502000-00038

Winstein, C. J., Stein, J., Arena, R., Bates, B., Cherney, L. R., Cramer, S. C., Deruyter, F., Eng, J. J., Fisher, B., Harvey, R. L., Lang, C. E., Mackay-Lyons, M., Ottenbacher, K. J., Pugh, S., Reeves, M. J., Richards, L. G., Stiers, W., & Zorowitz, R. D. (2016). Guidelines for adult stroke rehabilitation and recovery. *Stroke, 47*(6), e98–e169. https://doi.org/10.1161/STR.0000000000000098

World Health Organization. (2018). *International classification of functioning, disability and health*. http://www.who.int/classifications/icf/en/

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