ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

How Do Clinicians Judge Fluency in Aphasia?

Jean K. Gordon[a] and **Sharice Clough**[b]

[a] Department of Communication Sciences and Disorders, The University of Iowa, Iowa City [b] Department of Hearing and Speech Sciences, Vanderbilt University Medical Center, Nashville, TN

ARTICLE INFO

Article History:
Received September 9, 2021
Revision received December 12, 2021
Accepted December 17, 2021

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

https://doi.org/10.1044/2021_JSLHR-21-00484

ABSTRACT

Purpose: Aphasia fluency is multiply determined by underlying impairments in lexical retrieval, grammatical formulation, and speech production. This poses challenges for establishing a reliable and feasible tool to measure fluency in the clinic. We examine the reliability and validity of perceptual ratings and clinical perspectives on the utility and relevance of methods used to assess fluency.
Method: In an online survey, 112 speech-language pathologists rated spontaneous speech samples from 181 people with aphasia (PwA) on eight perceptual rating scales (overall fluency, speech rate, pausing, effort, melody, phrase length, grammaticality, and lexical retrieval) and answered questions about their current practices for assessing fluency in the clinic.
Results: Interrater reliability for the eight perceptual rating scales ranged from fair to good. The most reliable scales were speech rate, pausing, and phrase length. Similarly, clinicians' perceived fluency ratings were most strongly correlated to objective measures of speech rate and utterance length but were also related to grammatical complexity, lexical diversity, and phonological errors. Clinicians' ratings reflected expected aphasia subtype patterns: Individuals with Broca's and transcortical motor aphasia were rated below average on fluency, whereas those with anomic, conduction, and Wernicke's aphasia were rated above average. Most respondents reported using multiple methods in the clinic to measure fluency but relying most frequently on subjective judgments.
Conclusions: This study lends support for the use of perceptual rating scales as valid assessments of speech-language production but highlights the need for a more reliable method for clinical use. We describe next steps for developing such a tool that is clinically feasible and helps to identify the underlying deficits disrupting fluency to inform treatment targets.
Supplemental Material: https://doi.org/10.23641/asha.19326419

The fluency of verbal expression is commonly assessed in individuals with aphasia, both to provide a description of spontaneous speech difficulties and to facilitate the diagnosis of aphasia subtype. As defined by Clough and Gordon (2020), fluency in language production arises from the ability to smoothly coordinate linguistic subtasks, including the formulation of a syntactic framework, the timely retrieval and integration of words into the emerging framework, and the seamless programming of the formulated message for articulation. However, it has long been recognized that the measurement of fluency has poor reliability (Kerschensteiner et al., 1972; Poeck, 1989), particularly when used to make dichotomous judgments about diagnostic category (i.e., fluent aphasia vs. nonfluent aphasia). One of the main reasons for this lack of reliability is the complexity of fluency as a construct—there are a number of spontaneous speech dimensions that can affect how fluently language is produced, including word retrieval difficulties, grammatical formulation difficulties, and problems with phonological encoding and articulation. These difficulties may result in slowed and/or reduced speech production; increased (longer and/or more frequent) pausing; repetitions, repairs, and abandoned utterances; effortful speech production, sometimes with disrupted prosody; and telegraphic syntactic structures. The particular underlying impairments and the way in which they manifest vary widely in different

Correspondence to Jean K. Gordon: jean-k-gordon@uiowa.edu. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

people with aphasia (PwA). Furthermore, individual clinicians may have different conceptions about which variables are most salient to fluency (Holland et al., 1986).

Because the assessed degree of fluency is an integral step in determining aphasia subtype and overall aphasia severity, this lack of reliability has implications for both the accuracy of diagnosis and the specificity of treatment. For example, the classification of conduction and anomic aphasia as "fluent" aphasias may overlook the extent to which phonological encoding deficits and anomia, respectively, can disrupt the fluency of output. If two clinicians consider fluency to depend primarily on different underlying skills—for example, agrammatism or motor speech impairments—their ability to effectively communicate about the fluency of a given client is reduced. Developing a more consistent and reliable method of determining fluency can help avoid these interpretive issues. As a step in this direction, this study examines factors contributing to clinical impressions of fluency in individuals with aphasia by comparing clinical ratings and objective measures of spontaneous speech and by explicitly asking clinicians about their fluency measurement methods.

Clinical assessment of fluency by standardized means typically takes one of two approaches. The first involves combining multiple dimensions that contribute to fluency. In the Boston Diagnostic Aphasia Examination (BDAE; Goodglass et al., 2001b), this is accomplished by generating a profile of ratings along six relevant dimensions (melodic line, phrase length, articulatory agility, grammatical form, paraphasia, and word finding) and matching the profile of ratings to prototypical profiles for different subtypes of aphasia. Although no fluency score per se is generated, the classification of subtype aids in identifying whether the aphasia is a "fluent" or "nonfluent" subtype. In the Western Aphasia Battery–Revised (WAB-R; Kertesz, 2006), multiple dimensions (including utterance length, prosody, effort, hesitations, aspects of grammatical form, paraphasias, and word finding) are combined into one 11-point "Fluency" scale (actually labeled the "Fluency, Grammatical Competence, and Paraphasias" scale). For a given PwA, a score along the scale is assigned according to the best-fitting description corresponding to each point on the scale. Although the consideration of multiple dimensions lends validity to this approach, the methods by which dimensions are combined result in a great deal of subjectivity and a noncontinuous scale (Gordon & Clough, 2020). Fluency ratings have been shown to have poor reliability whether based on the WAB scale (Trupe, 1984) or BDAE parameters (Gordon, 1998), and syndrome diagnoses using the WAB and the BDAE are often discrepant (Wertz et al., 1984).

The second approach, evident in the Aphasia Diagnostic Profiles (ADP; Helm-Estabrooks, 1992), is to rely on a single dimension that can—at least in theory—be measured more objectively. In the ADP, fluency is calculated based on phrase length. Relying on a single quantitative dimension is likely to be more reliable than the multidimensional ratings described above, but possibly at the expense of the validity of the measurement, because a single measure may not reflect all the relevant contributors to fluency. Two of the most commonly used quantitative measures, however—phrase length and speech rate—are useful, in that they have been shown to reflect the influence of multiple underlying aspects of spontaneous speech (Gordon & Clough, 2020).

Our previous work has examined the characteristics of spontaneous speech that underlie impressions of fluency in narrative retellings of the Cinderella story. Clough and Gordon (2020) compared two sets of binary fluency classifications for 254 PwA in AphasiaBank (MacWhinney et al., 2011), one based on WAB-R scores (as described above) and the other based on clinical impression. Logistic regressions showed that WAB-R classifications were primarily dependent on aphasia severity, as well as a combination of lexical (type–token ratio [TTR], empty speech, and semantic errors) and grammatical (sentence complexity) variables, whereas clinical judgments were primarily affected by the presence of apraxia, as well as aphasia severity and lexical retrieval measures (TTR and empty speech). This finding indicates that even multidimensional measures of fluency, such as the WAB-R Fluency scale, may miss dimensions that clinicians deem to be important, such as apraxia of speech.

A companion paper (Gordon & Clough, 2020) examined contributors to three continuous measures commonly used as proxies of fluency—the WAB-R Fluency scale, utterance length (mean length of utterance [MLU] in words), and speech rate (words per minute [WpM]). As with binary fluency classifications (Clough & Gordon, 2020), aphasia severity was the strongest predictor of WAB-R Fluency scores, but lexical diversity (TTR), grammatical complexity, the presence of dysarthria, and the frequency of semantic errors also contributed. Utterance length and speech rate were also predicted by grammatical complexity and lexical diversity, as well as propositional density and content/function word ratio, but grammatical complexity was the strongest predictor of both. Predictors of utterance length (but not speech rate) also included aphasia severity; predictors of speech rate (but not utterance length) included pitch variability and apraxia of speech. Together, these results highlight considerable overlap in measures of fluency, along with some important differences. The WAB-R scale primarily reflects aphasia severity; utterance length reflects linguistic aspects of expression, both grammatical and lexical, and speech rate reflects motor speech as well as linguistic dimensions.

A limitation of this prior work is the lack of a continuous measure of fluency itself (rather than a proxy measure) analogous to the dichotomous fluency classifications

examined by Clough and Gordon (2020) and others in previous studies (e.g., Kerschensteiner et al., 1972; Park et al., 2011; Swindell et al., 1984). It is clear from previous work that considering fluency as a dichotomy is a flawed approach (Clough & Gordon, 2020; Feyereisen et al., 1991; Gordon, 1998; Trupe, 1984), because it overlooks important variation in degree of fluency and because PwA judged to be fluent by one dimension may be nonfluent by another. However, identifying a valid and reliable continuous measure of fluency is difficult. The WAB-R Fluency scale is intended to serve as just such a continuous measure, but the extensive descriptions at each anchor effectively result in a set of categories that are roughly ordered by severity, rather than a truly continuous measure (see Gordon & Clough, 2020, for details). A clearer understanding of what influences impressions of fluency among clinicians requires a continuous rating scale that can capture whatever dimensions are considered to be important predictors of fluency for a particular PwA in a particular context. This study addressed this need by collecting, in an online survey format, ratings of fluency and related dimensions of spontaneous speech in a range of PwA and comparing these ratings with objectively measured characteristics of the speech samples. Several analyses were conducted to examine the reliability, validity, and clinical relevance of methods used to assess fluency.

Method

The project was approved by the Institutional Review Board at the University of Iowa. Survey respondents were paid $25 in the form of a gift certificate if they completed the survey and were entered into a drawing for a $100 gift certificate.

Samples

Of the 278 unique English-speaking individuals with aphasia who had completed the AphasiaBank protocol at the time the survey was developed, the set was filtered to include those who (a) had continuing aphasia at the time, according to the WAB-R severity cutoff score of 93.8, (b) had completed the Cinderella story retell task, and (c) produced Cinderella stories that included at least three spontaneous (i.e., uncued) utterances but did not exceed 6 min in length. We chose to examine fluency in the Cinderella story retelling task because it standardizes the content somewhat across PwA (unlike, for example, describing an important life event) but represents a more ecologically valid communicative task than, for example, describing a sequence of pictures.

Video recordings of the AphasiaBank protocol for these 191 PwA were downloaded from AphasiaBank and trimmed using Avidemux 2.6 (2017), a free video editing tool, to include only the Cinderella story, excluding initial experimenter prompts. These video files were then converted to audio (.WAV) files by a batch processor. The purpose of using audio-only samples was to focus the clinicians' perceptions on *spoken* speech-language dimensions related to fluency, providing a more consistent basis for their judgments. This allowed us to measure the reliability and validity of judgments of verbal output without the influence of nonverbal cues. The audio files were edited using GoldWave 6.31 (2017) to improve the quality of the sound: The "maximize volume" effect was applied to increase the volume of the voice signal without clipping distortion. "Noise reduction" was applied to remove consistent background noise by a scale of 30% to improve the signal-to-noise ratio of the file while maintaining the naturalness of the speech. The audio files of the Cinderella story were then listened to by two research assistants for sound quality. Six samples were judged to have problems (e.g., low voice volume and significant background noise) that might interfere with judgments of the speech and language and were removed from the set, leaving samples from 185 PwA.

AphasiaBank includes aphasia syndrome classifications by clinical impression and by WAB-R scale scores. Because this study examines clinical perceptions (and because of known problems with the WAB-R scale cutoff; Clough & Gordon, 2020; Trupe, 1984), the clinician syndrome classifications were determined to be the most relevant for this study. These were used except when unavailable (n = 19), in which case the WAB-R syndrome classifications were used. In addition, three of the clinical categories were very small—global aphasia (n = 4), transcortical sensory aphasia (n = 1), mixed transcortical aphasia (n = 1), and optical aphasia (n = 1). To allow for a more robust analysis, these were recategorized according to their WAB classifications—global and mixed transcortical as Broca's aphasia and transcortical sensory and optical as anomic aphasia. Following the (re)classification of these 26 PwA, the set of 185 samples consisted of 64 individuals with anomic aphasia (35%), 70 with Broca's aphasia (38%), 32 with conduction aphasia (17%), 12 with Wernicke's aphasia (6%), and seven with transcortical motor (TCM) aphasia (4%). Seventy-nine were women, and 106 were men. Their ages ranged from 25 to 90 years, with a mean of 62 years. They ranged from 0 to 9 on the WAB Fluency scale (M = 6.1) and had aphasia quotient (AQs) ranging from 10.8 to 93.4 (M = 70.2).

Objective Measures

Transcripts of the PwA were analyzed using EVAL and other commands in Computerized Language ANalysis (CLAN; MacWhinney, 2000) to generate a range of measures characterizing the samples. On the basis of the prior work

described above, 18 variables (16 continuous and two categorical) were selected as having potential impacts on the ratings of the fluency dimensions. These are listed in Table 1.

Survey

The survey was administered online. The text of the survey is provided in Supplemental Material S1, and additional information about the design and administration of the survey (following guidelines from Eysenbach, 2004) is shown in Supplemental Material S2. Prior to presenting the survey questions, a consent document was presented, which explained the study. Respondents provided consent by clicking to the next page. Next, responses to six questions[1] about the respondent were elicited: their age, level of education, work setting(s), years of experience as a speech-language pathologist (SLP), proportion of caseload consisting of PwA, and number of PwA interacted with professionally. Each question was in multiple-choice format with an opportunity to decline to answer (age and education) or to provide an alternative text response.

Next, an instruction slide informed respondents about the format of the ratings. They were encouraged to listen to the audio samples over headphones in a quiet setting and were told that they could play the sample as many times as they liked. A practice audio sample was followed by eight perceptual rating questions about the sample, as listed below. The first question asked them to rate the overall fluency of the speaker; the remaining seven questions asked them to rate specific speech-language dimensions[2] hypothesized to contribute to impressions of fluency: speech rate, pausing, effort, melodic line, phrase length, grammaticality, and lexical retrieval.

a) FLUENCY: *How fluent is the speaker during this sample?*
b) SPEECH RATE: *How slow is the speaker's rate of speech during this sample?*
c) PAUSING: *How much of the sample consists of pauses?*
d) EFFORT: *How effortful is the speaker's articulation during this sample?*
e) MELODY: *How restricted is the speaker's melodic line or intonational contour during this sample?*
f) PHRASE LENGTH: *How restricted is the speaker's typical phrase length during this sample?*
g) GRAMMATICALITY: *How grammatical is the speaker during this sample?*
h) LEXICAL RETRIEVAL: *How limited is the speaker's word retrieval during this sample?*

Rating responses were recorded using a slider bar along a visual analogue scale (VAS) with text anchors at either end. A VAS was considered preferable to a scale with discrete measurement points to reflect the assumption that fluency varies continuously. Furthermore, by not including intermediate anchor points, a VAS makes fewer assumptions about the distance between points. Research suggests that VAS methods generate responses that are as reliable and valid as discrete-point rating scales but with greater sensitivity (Nguyen & Fabrigar, 2018). In this study, the anchors corresponded to less fluent output at the left end and more fluent output at the right end. The fluent end of the scale represented normal speech production, except for the scales for fluency and rate, which can deviate from normal in both directions. For these scales, the right-hand anchors indicated "hyperfluent" and "abnormally fast rate," respectively. The resulting differences in effective length of the scales were dealt with in the analyses by normalizing the scales, as described below. Respondents were also given an "unable to rate" (UR) option for each fluency dimension.

After the practice sample was rated, 20 experimental samples (or 10, depending on the version of the survey—see below) were randomly selected from the set of 185 PwA. For each trial, respondents listened to the audio sample and rated the eight fluency dimensions described above, which were always presented in the same order. Following the rating of the samples, respondents were asked to answer four further questions about how they measured fluency in the clinic, what dimensions were considered most important, whether they thought a more reliable measure was needed, and any additional comments that they wanted to add.

Procedure

The programming and dissemination of the survey, collation of responses, and disbursement of remuneration to respondents were managed by The University of Iowa Social Science Research Center (https://ppc.uiowa.edu/isrc), in part to ensure anonymity of the responses. A link to the online survey was initially disseminated through the listserve of ASHA's Special Interest Group 2 (Neurogenic Communication Disorders), with over 4,000 members, and through word of mouth (e.g., at conferences). The return rate was very low, however, so the survey instrument was modified to present only 10 audio samples instead of 20. The link to the revised survey was disseminated to the Google Group of AphasiaBank (over 700 members), again through word of mouth (conferences and e-mailing larger

[1]An additional question (Q3) asked for the respondent's email address, which was used for the purpose of providing reimbursement. This question is not shown in Supplemental Material S1, because the responses were not made available to the research team to preserve respondents' anonymity.
[2]To differentiate between the objective measures and the rating scales, the ratings will subsequently be referred to in small caps.

Table 1. List of objective measures, codes, and descriptions.

Objective measure	Code	Description
Speech rate	WpM	Words per minute, not including retraced or repeated words
Utterance length	MLU	Mean length of utterance, not including nonwords or unintelligible words
Retracing	Retrace	Number of reformulated and repeated words, calculated as a proportion of total words, i.e., tokens
Content/function ratio	Con Fun	Ratio of content words to function words
Complex grammar[a]	Gram Com	Proportion of utterances containing embeddings
Verb inflection	Vb Inflect	Total verb inflections divided by total verbs
Propositional density	Prop Dens	Propositional density: number of proposition-forming words (verbs, adjectives, adverbs, prepositions, and conjunctions) as proportion of total words
Lexical diversity[a]	MATTR	Moving-average type–token ratio, generated by counting the ratio of types to tokens in a succession of windows of fixed length (here, we used the average TTR using windows of five, 10, and 20 words)
Grammatical errors	Gram Err	Proportion of utterances containing one or more grammatical errors
Morphological errors[a]	Morph Err	Proportion of tokens containing morphological errors
Neologistic errors[a]	Neo Err	Proportion of tokens consisting of neologistic errors
Phonological errors[a]	Phon Err	Proportion of tokens consisting of phonological errors
Semantic errors[a]	Sem Err	Proportion of tokens consisting of semantic errors
Circumlocution	Circum	Proportion of utterances containing circumlocutions
Empty speech	ES	Proportion of utterances containing empty speech
Pitch variation[b]	Pitch Var	Standard deviation of fundamental frequency
Apraxia of speech	AoS	Presence or absence, as documented in AphasiaBank
Dysarthria	Dys	Presence or absence, as documented in AphasiaBank

Note. All measures were obtained using the EVAL command in Computerized Language Analysis (CLAN), except where noted.

[a]Grammatical complexity, MATTR, and lexical-level error proportions were generated using the FREQ command in CLAN (see CLAN manual for details). [b]Pitch variability was calculated using Praat from a 60-s excerpt of each audio file edited to exclude examiner speech and background noise. The analysis window was narrowed to include values just above and below the speaker's maximum and minimum and fundamental frequency.

SLP departments in rehabilitation facilities) and by postal mail to a list of 3,714 addresses (generated by Dynata, a marketing research company) associated with Standard Industrial Codes of "speech specialists," "speech therapists," or "speech pathologists."

Analyses

Responses on the visual analog scale were recorded as numbers ranging from 0 to 100. For statistical analysis, these raw scores were transformed to z scores calculated across all individual ratings but separately for each rating dimension. Similarly, the objective measures obtained from AphasiaBank were also standardized, putting them all on the same scale. We conducted three types of analysis, which aimed to investigate the reliability of perceptual ratings relevant to fluency (Analysis 1), the validity of the ratings as they pertain to more objective measures and aphasia subtypes (Analysis 2), and the opinions of clinicians regarding methods of assessing fluency (Analysis 3).

Analysis 1: Interrater reliability

To assess interrater reliability of ratings, we calculated intraclass correlation coefficients (ICC; Bartko, 1966; McGraw & Wong, 1996; Shrout & Fleiss, 1979) using the iccNA() function from the irrNA package in R (Brueckl & Heuer, 2021). Samples were rated by different (but overlapping) random subsets of respondents, resulting in a varying number of ratings per PwA (see the Respondents section below). The irrNA package provides interrater reliability coefficients for data sets that are randomly incomplete (i.e., unbalanced) without imputing missing values or omitting available data. So as not to make a priori assumptions about contributing sources of variance, we followed recommendations to report all relevant forms of ICC and their associated confidence intervals (e.g., Liljequist et al., 2019; Shrout & Fleiss, 1979). This included one-way and two-way models with single raters as the unit (i.e., each rating corresponds to a single measurement rather than an average measurement) as well as ICCs of averaged ratings for comparison.

One-way models using single-rater analysis, referred to as ICC(1,1), reflect the variability in ratings between PwA relative to the variability within PwA, without parceling out the contribution of rater-specific biases; that is, all within-PwA variability is considered to be error. Two-way models consider the contribution of rater-specific biases. If one-way and two-way models yield similar ICC results, it suggests that rater bias effects are small or absent; if these values differ, then one-way models should be rejected (Liljequist et al., 2019), as they will underestimate reliability. In random two-way models (used here), both subjects and raters are

assumed to be randomly sampled from their respective populations. In addition, for two-way models, two different outcomes have been defined: absolute agreement and consistency (McGraw & Wong, 1996). While absolute agreement reflects the degree to which raters assign the same value to a given target, consistency reflects the relative ranks of values that raters assign to different targets (Hallgren, 2012; Liljequist et al., 2019). For example, one rater may be biased to use the lower end of a rating scale, whereas another might tend to provide ratings at the upper end of the scale. Such raters might, despite having poor absolute agreement, still have good consistency, if they tend to rate PwA in the same order on the scale. The coefficient for absolute agreement, ICC(A,1), accounts for such systematic rater biases, whereas the coefficient for consistency, ICC(C,1), reflects an estimate of the ICC that would be obtained if systematic rater biases could be eliminated.

Analysis 2: Validity of perceptual ratings

To determine which objective measures most strongly influenced the respondents' perceptions, *z*-score ratings were correlated with the objective measures (Analysis 2a). In addition, *z*-score residuals were generated by regressing each of the seven speech-language dimensions (RATE, PAUSING, EFFORT, MELODY, PHRASE LENGTH, GRAMMAR, and LEXICAL RETRIEVAL) on the overall FLUENCY rating. This allowed us to factor out some of the shared variance between the ratings (i.e., halo effects; Thorndike, 1920). We also examined how respondents' perceptual ratings corresponded to expected patterns for different types of aphasia, comparing *z*-score ratings and residual ratings across aphasia types (Analysis 2b).

Analysis 3: Perceptions of the fluency construct

For the final analysis, post-rating responses about fluency assessment were analyzed. First, we examined potential variables affecting *which* dimensions were used to judge fluency, *how many* were typically used, and *how important* they were judged to be. These variables included professional characteristics of the respondents, specifically their years of experience, proportion of caseload with aphasia, number of PwA seen, and education level (Analysis 3a). Each characteristic was dichotomized to facilitate analysis (see the Results section) and to maximize power by keeping subgroups as large as possible but similar in size. Chi-square analyses were used to examine the proportions of respondents who endorsed each fluency dimension, and *t* tests were used to examine the mean importance given to each dimension. Finally, the open-ended responses were examined to identify main themes regarding the fluency concept (Analysis 3b). Each author reviewed the open-ended responses to identify themes that emerged from the data (i.e., they were not specified beforehand). After coming to a consensus on the number and nature of the themes, each author categorized the comments into one or more of the thematic categories.

Results

Respondents

Ninety-two people completed the survey: 28 completed the initial 20-sample version and 64 completed the 10-sample version. One of these (who responded to the mailed invitation) was not an SLP, and one reported having never interacted professionally with a person with aphasia. Responses from both participants were removed from the data set, leaving 90 respondents. An additional 22 individuals (nine for the 20-sample version and 13 for the 10-sample version) started the survey but did not complete it. However, because the PwA were randomly selected and ordered for each rater, we were able to include the rating data from these partial surveys. In all, 1,309 sets of ratings were collected, 1,175 from completed surveys and an additional 134 from partial surveys.

Demographic characteristics of the respondents are shown in Table 2. In brief, they were fairly well distributed across age groups from 20 to 70, and the highest degree for most of them (86%) was a Master of Arts or Master of Science. The most common work settings were private practice (40%) and rehabilitation units (29%). Respondents' experience, in years of practice, skewed negatively, with well over half (61%) having at least 10 years of experience and 38% having over 20 years of experience. A plurality of the respondents (39%) reported having worked with over a hundred PwA. However, relatively few worked primarily with PwA in their current setting—half reported that 20% or less of their typical caseload consisted of PwA.

Because of the random selection process, the number of respondents who rated each sample ranged from 0 to 16. To ensure that each PwA was rated by at least three respondents, we excluded four PwA: one with Broca's aphasia (two respondents), one with conduction aphasia (two respondents), and two with anomic aphasia (zero and one respondent). Thus, the final data set included 1,304 sets of ratings of 181 PwA, with an average of 7.2 ratings per PwA (range: 3–16) and an average of 11.6 ratings (range: 1–20) per respondent. The final set of PwA analyzed, along with their aphasia subtype classifications and severity measures, is provided in Supplemental Material S3.

Analysis 1: Interrater Reliability

ICCs for each of the eight perceptual fluency scales are presented in Table 3 and interpreted relative to Cichetti's (1994) guidelines: An ICC of < .40 indicates poor clinical significance, an ICC of .40–.59 is fair, an ICC of .60–.74 is good, and an ICC of .75 or higher is excellent. We took the confidence intervals into account in determining these levels. Values are provided for all the models, although we considered the two-way single-rater models as our benchmarks because these are able to account for individual rater biases. Judging from the two-way models, ratings of overall

Table 2. Characteristics of survey respondents.

Demographics	No. (%) of completed surveys	No. (%) of incomplete surveys	National data (ASHA, 2021a, 2021b)[a]	
Age				
20–30 years	16 (18%)	4 (18%)	**< 35**	**29%**
31–40 years	**22 (24%)**	**6 (27%)**		
41–50 years	**21 (23%)**	**6 (27%)**	**35–44**	**28%**
51–60 years	16 (18%)	3 (14%)	45–54	22%
61–70 years	13 (14%)	2 (9%)	55–64	13%
71–80 years	1 (1%)	1 (5%)	65+	8%
NA	1 (1%)	0 (0%)		
Education				
MA/MS	**77 (86%)**	**19 (86%)**		**98%**
PhD/Clinical doctorate	13 (14%)	3 (14%)		2%
Work setting				
Acute care	9 (10%)	6 (27%)		12%
Rehabilitation	26 (29%)	6 (27%)		
Long-term care	2 (2%)	0 (0%)		10%
Private practice	**36 (40%)**	6 (27%)		2%
Outpatient/home health[b]	12 (13%)	**7 (32%)**		16%
University	13 (14%)	6 (27%)		3%
Education[b]	10 (11%)	1 (5%)		**51%**
Other	2 (2%)	0 (0%)		7%
Length of practice				
< 1 year	1 (1%)	0 (0%)	**NA**	
1–5 years	13 (14%)	6 (27%)		
5–10 years	21 (23%)	2 (9%)		
10–20 years	21 (23%)	6 (27%)		
> 20 years	**34 (38%)**	**8 (36%)**		
Proportion of caseload with aphasia				
1%–20%	**36 (40%)**	**9 (41%)**	**NA**	
21%–40%	19 (21%)	3 (14%)		
41%–60%	11 (12%)	3 (14%)		
61%–80%	5 (6%)	3 (14%)		
81%–100%	10 (11%)	2 (9%)		
None currently	9 (10%)	2 (9%)		
No. PwA seen				
1–9	13 (14%)	3 (14%)	**NA**	
10–20	10 (11%)	2 (9%)		
21–50	21 (23%)	3 (14%)		
51–100	11 (12%)	6 (27%)		
> 100	**35 (39%)**	**8 (36%)**		

Note. Dominant responses for each group and each question are shown in bold font. NA = not available; PwA = people with aphasia.

[a]Estimated from ASHA (2021a, 2021b) *Profile of ASHA members and affiliates, year-end 2020 and profile of ASHA members and affiliates with PhDs, year-end 2020*. [b]The "Outpatient/Home Health" and "Education" categories were not provided on the survey but were frequent write-in responses in the "Other" category, so they have been included here separately.

FLUENCY yielded fair to good interrater reliabilities for both absolute agreement and consistency, as did ratings of MELODY and GRAMMATICALITY. Ratings of SPEECH RATE, PAUSING, and PHRASE LENGTH yielded good interrater reliabilities, whereas the reliabilities for EFFORT and LEXICAL RETRIEVAL were only fair. The small differences between ICC(A,1) and ICC(C,1) suggest little systematic rater bias.

Relative to the two-way tests, ICC values for the corresponding one-way tests were considerably lower, ranging from poor to fair for most of the dimensions. This was expected because one-way models attribute any rater variance to error variance. The relatively large differences between one-way and two-way models suggest that, although it may not be systematic as noted above, there does exist considerable variance across raters. In our design, this may be related in part to the random assignment of raters to PwA. In contrast to the single-rater models, corresponding ICC values for average-rater models (shown at the bottom of Table 3) are all much higher, in the range of excellent for all dimensions. This supports the conclusion that a significant amount of the variability across PwA can be attributed to the different raters and that this variability can be considerably reduced by averaging over multiple raters.

Analysis 2: Validity of the Ratings

2a. Relationship of Perceptual Ratings to Objective Measures

All rating dimensions were moderately to strongly related to each other, with correlations ranging from .395

Table 3. Interrater reliability characterized by intraclass correlation coefficients (ICCs) for each of the eight perceptual rating scales.

	Single-rater ICCs		
Rating scale	**ICC(1,1) [CI]**	**ICC(A,1) [CI]**	**ICC(C,1) [CI]**
FLUENCY	.454 [.39, .52][a–b]	.597 [.54, .66][b–c]	.603 [.54, .66][b–c]
SPEECH RATE	.548 [.49, .61][b–c]	.665 [.61, .72][c]	.669 [.62, .72][c]
PAUSING	.531 [.47, .59][b]	.665 [.61, .72][c]	.669 [.62, .72][c]
EFFORT	.372 [.31, .44][a–b]	.485 [.42, .55][b]	.492 [.43, .56][b]
MELODY	.405 [.34, .47][a–b]	.551 [.49, .61][b–c]	.559 [.50, .62][b–c]
PHRASE LENGTH	.540 [.48, .60][b]	.659 [.61, .71][c]	.666 [.61, .72][c]
GRAMMATICALITY	.425 [.36, .49][a–b]	.548 [.49, .61][b–c]	.556 [.50, .62][b–c]
LEXICAL RETRIEVAL	.375 [.31, .44][a–b]	.484 [.42, .55][b]	.489 [.43, .55][b]
	Average-rater ICCs		
Rating Scale	**ICC(1,*k*) [CI]**	**ICC(A,*k*) [CI]**	**ICC(C,*k*) [CI]**
FLUENCY	.857 [.82, .89][d]	.914 [.89, .93][d]	.912 [.90, .93][d]
SPEECH RATE	.900 [.87, .92][d]	.934 [.92, .95][d]	.935 [.92, .95][d]
PAUSING	.891 [.86, .91][d]	.934 [.92, .95][d]	.935 [.92, .95][d]
EFFORT	.808 [.76, .85][d]	.870 [.84, .90][d]	.873 [.84, .90][d]
MELODY	.829 [.79, .86][d]	.897 [.87, .92][d]	.900 [.88, .92][d]
PHRASE LENGTH	.894 [.87, .92][d]	.932 [.92, .95][d]	.934 [.92, .95][d]
GRAMMATICALITY	.836 [.80, .87][d]	.893 [.87, .92][d]	.896 [.87, .92][d]
LEXICAL RETRIEVAL	.811 [.77, .85][d]	.870 [.84, .90][d]	.872 [.84, .90][d]

Note. ICC(1,1) is a random one-way model using single raters as the unit of measurement. ICC(A,1) and ICC(C,1) are random single-rater two-way models using absolute agreement and consistency measures, respectively. Corresponding ICC models notated with *k* used average ratings as the unit of measurement (*k* is unspecified because the number of raters was variable across people with aphasia).
[a]Poor reliability (ICC < .40). [b]Fair reliability (.40 < ICC < .59). [c]Good reliability (.60 < ICC < .74). [d]Excellent reliability (ICC > .75).

between EFFORT and LEXICAL RETRIEVAL to .712 between SPEECH RATE and PAUSING (all *p*s < .0001).[3] Intercorrelations among the eight ratings are shown in Supplemental Material S4.

Table 4 shows correlations between *z* scores of the 16 continuous objective measures and the rating dimensions. Top rows show mean *z*-score ratings averaged across all respondents for a given PwA (n = 181 for each dimension). All significant correlations (p < .05) are shown. Middle rows show correlations above a small effect size (r > .10, p < .001; Cohen, 1988) for individual ratings (n = 1,304 for each dimension). Mean and individual ratings showed similar patterns, but with consistently stronger correlations for mean ratings (as would be expected, because interrater variability is eliminated by averaging). Individual ratings of FLUENCY were most strongly influenced by objective measures of speech rate (r = .559) and utterance length (r = .496) and also showed positive associations with measures of grammatical complexity (r = .410) and lexical diversity (r = .356) and a negative relationship with phonological errors (r = −.311). Figure 1 graphically illustrates the impact of each of these four objective measures on FLUENCY ratings. Despite the robust correlations, it is clear that there is a great deal of variability in the associations of the objective measures and FLUENCY ratings and that the relationships are driven by the more extreme values (e.g., low moving-average TTR [MATTR] scores, frequent phonological errors, and high rates of speech).

Because of the strong intercorrelations among perceptual ratings, all the dimensions showed similar patterns. However, some perceptual ratings showed particularly strong relationships to their corresponding objective measures. For example, ratings of SPEECH RATE and PAUSING were most strongly related to measured speech rate (*r*s = .652 and .658, respectively), and ratings of GRAMMATICALITY were strongly related to measures of grammatical complexity (r = .413) and grammatical errors (r = −.347). Other relationships that showed at least a medium effect size were between measured lexical diversity (MATTR) and ratings of PHRASE LENGTH (r = .390), GRAMMATICALITY (r = .383), and LEXICAL RETRIEVAL (r = .353) and between proportion of phonological errors and rated EFFORT (r = −.318). Lower ratings on all dimensions (all *p*s < .001) were given to the 63 PwA with concomitant apraxia of speech and the 21 PwA with dysarthria (all *p*s < .001 except lexical retrieval, p = .059).

Because the perceptual ratings showed a considerable amount of shared variance, we also calculated rating residuals by regressing the ratings of specific speech-language dimensions on the overall rating of FLUENCY. Factoring out the overall FLUENCY rating in this way helped identify measures contributing to each speech and language rating beyond their shared variance with overall

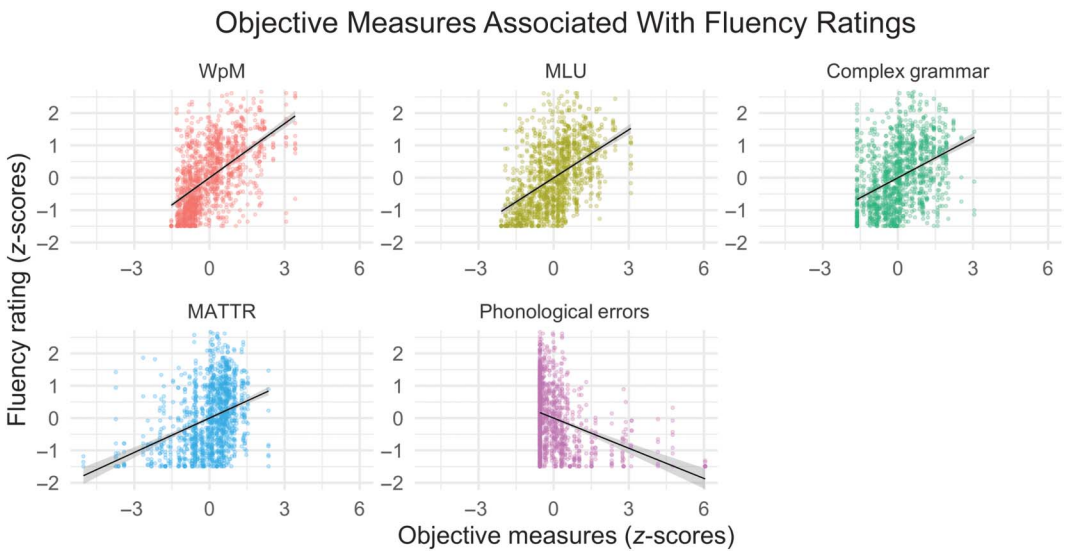
[3]Recall that, for ease of interpretation, all scales (even EFFORT and PAUSING) were structured such that low scores corresponded to lower fluency and high scores to greater fluency.

Table 4. Correlations between *z*-scores of the 16 continuous objective measures (columns) and mean *z*-score ratings (top), individual z score ratings (middle), and z scores residualized on fluency ratings (bottom).

Perceptual rating scales	WpM	MLU	Retrace	Con Fun	Gram Com	Vb Inflect	Prop Dens	MATTR	Gram Err	Morph Err	Neo Err	Phon Err	Sem Err	Circum	ES	Pitch Var
Mean *z*-score ratings																
FLUENCY	***0.76***	***0.69***	**0.30**	–0.29	***0.57***		0.22	**0.47**	–0.27	–0.16	–0.25	**–0.41**		**0.35**	0.26	
SPCH RATE	***0.83***	***0.58***	0.28	–0.26	**0.45**		0.16	**0.35**	–0.16		–0.27	**–0.38**		0.27	0.29	
PAUSING	***0.84***	***0.59***	0.26	–0.17	**0.48**		0.23	**0.35**	–0.15		–0.16	**–0.31**		0.25	0.28	
EFFORT	***0.68***	***0.56***	**0.32**	–0.20	**0.47**		0.20	**0.38**	–0.16		**–0.32**	**–0.45**		0.27	0.22	
MELODY	***0.73***	***0.51***	0.23	–0.21	**0.43**		0.16	**0.31**			**–0.31**	**–0.36**	–0.15	0.23	0.23	
PHRASE	***0.80***	***0.74***	**0.32**	**–0.32**	***0.57***		**0.31**	**0.49**	**–0.31**	–0.16	–0.26	**–0.35**		0.29	0.22	–0.15
GRAMM	***0.62***	***0.77***	**0.37**	**–0.38**	***0.59***	0.16	**0.32**	***0.55***	**–0.43**	–0.15	**–0.38**	**–0.34**	–0.18	0.29	0.16	–0.24
LEXICAL	***0.62***	***0.71***	0.29	–0.24	***0.54***		**0.31**	***0.52***	–0.24			**–0.32**	–0.24	0.26		
Individual *z*-score ratings																
FLUENCY	***0.56***	***0.50***	0.22	–0.21	**0.41**		0.16	**0.36**	–0.22	–0.14	–0.18	**–0.31**		0.25	0.19	
SPCH RATE	***0.65***	**0.46**	0.20	–0.20	**0.36**		0.12	0.28	–0.16	–0.13	–0.20	–0.29		0.23	0.24	
PAUSING	***0.66***	**0.46**	0.18	–0.13	**0.37**		0.17	0.28	–0.13	–0.12	–0.11	–0.25		0.19	0.23	
EFFORT	**0.46**	**0.39**	0.21	–0.13	**0.32**		0.11	0.26	–0.14	–0.10	–0.20	**–0.32**		0.19	0.14	
MELODY	***0.51***	**0.37**	0.15	–0.16	**0.31**			0.24	–0.13	–0.11	–0.20	–0.26		0.18	0.17	
PHRASE	***0.63***	***0.58***	0.24	–0.25	**0.45**		0.23	**0.39**	–0.27	–0.16	–0.20	–0.29	–0.10	0.22	0.18	–0.11
GRAMM	**0.42**	***0.53***	0.25	–0.27	**0.41**	0.15	0.20	**0.38**	**–0.35**	–0.16	–0.27	–0.26	–0.13	0.21		–0.14
LEXICAL	**0.43**	**0.49**	0.20	–0.16	**0.38**		0.20	**0.35**	–0.19	–0.11	–0.20	–0.23	–0.18	0.18		
z scores residualized on fluency ratings																
SPCH RATE	**0.35**	0.15									–0.10				0.15	0.12
PAUSING	**0.39**	0.18			0.14										0.15	
EFFORT	0.16	0.12	0.10		0.10						–0.12	–0.15				
MELODY	0.23										–0.11					0.15
PHRASE	**0.32**	**0.31**	0.11	–0.15	0.23		0.16	0.20	–0.15		–0.10					
GRAMM	0.13	0.28	0.14	–0.18	0.21	0.13	0.11	0.20	–0.25		–0.18					–0.10
LEXICAL	0.17	0.28	0.10		0.20		0.13	0.20			–0.12		–0.15			

Note. Only significant correlations are shown for mean ratings ($r \geq .15$, $p < .05$). For individual ratings, only those above a small effect size ($r \geq .10$) are shown. Medium-sized correlations ($r \geq .30$) are bolded; large-sized correlations ($r \geq .50$) are in bold italics. Please see Table 1 for an explanation of the variables. WpM = words per minute; MLU = mean length of utterance; Con Fun = content/function ratio; Gram Com = grammatical complexity; Vb Inflect = verb inflection; Prop Dens = propositional density; MATTR = moving-average type–token ratio; Gram Err = grammatical errors; Morph Err = morphological errors; Neo Err = neologistic errors; Phon Err = phonological errors; Sem Err = semantic errors; Circum = circumlocution; ES = empty speech; Pitch Var = pitch variation; SPCH RATE = speech rate; GRAMM = grammaticality.

Figure 1. Relationships between objective measures (on *x*-axes) associated with fluency ratings (on *y*-axes), showing all correlations with at least a medium effect size ($r > .30$). WPM = words per minute; MLU = mean length of utterance; MATTR = moving-average type–token ratio.



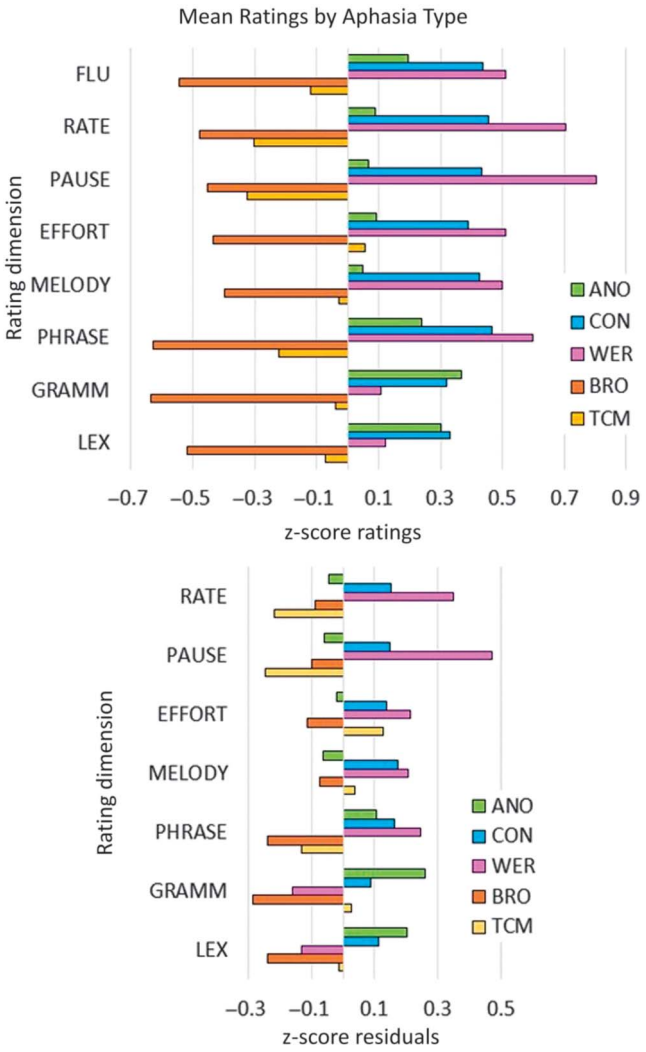
fluency. The bottom rows of Table 4 show correlations between *z* scores of the objective measures and the rating residuals, which provide a slightly more nuanced picture. Objective measures of speech rate and utterance length were still the most influential predictors overall: WpM most strongly predicted SPEECH RATE, PAUSING, EFFORT, MELODY, and PHRASE LENGTH residuals, whereas MLU most strongly predicted GRAMMATICALITY and LEXICAL RETRIEVAL residuals. Aside from WpM and MLU, the absence of phonological errors was the next strongest predictor of EFFORT residuals, and pitch variability was the next strongest predictor of MELODY residuals. GRAMMATICALITY residuals were affected by the absence of grammatical errors, and LEXICAL RETRIEVAL residuals were affected by lexical diversity (MATTR). In general, then, WpM and MLU captured some of the variance in all the perceptual ratings, but individual dimensions also reflected appropriate underlying measures of spontaneous speech.

2b. Relationship of Perceptual Ratings to Aphasia Subtypes

In the second validity analysis, perceptual ratings were compared across different aphasia types to determine whether the respondents' perceptions captured expected differences between aphasia syndromes. As in Analysis 2a, both rating *z* scores and rating residuals were regressed on overall FLUENCY. Figure 2 shows bar graphs of the average perceptual rating dimensions by aphasia type, with standardized ratings on the top and rating residuals on the bottom. Broad expected patterns were shown, in that speakers with Broca's and TCM aphasia received below-average ratings on almost all dimensions, and speakers with Wernicke's, anomic, and conduction aphasia received above-average ratings. For most of the dimensions, the contrast was greatest between Broca's and Wernicke's aphasia. More specifically, TCM aphasia received particularly low ratings for SPEECH RATE and PAUSING, while Broca's aphasia received the lowest ratings on PHRASE LENGTH and GRAMMATICALITY. Individuals with Wernicke's aphasia were rated highest on PAUSING and SPEECH RATE but lower on GRAMMATICALITY and LEXICAL RETRIEVAL. Those with anomic aphasia received intermediate ratings on most dimensions but relatively high ratings on GRAMMATICALITY and (somewhat unexpectedly) LEXICAL RETRIEVAL. These relatively high ratings might be attributed to the less severe nature of anomic aphasia; however, their lower ratings on SPEECH RATE, PAUSING, EFFORT, and MELODY do not seem to support this hypothesis.

The rating residuals illustrate discrepancies in the speech-language dimensions beyond what would be expected from the overall FLUENCY rating. For example, although speakers with Broca's aphasia received the lowest mean ratings on SPEECH RATE and PAUSING, those with TCM aphasia showed lower residuals, indicating that they were perceived to be worse on these dimensions than would be predicted from their overall FLUENCY ratings. Speakers with Wernicke's aphasia had positive residuals on ratings of SPEECH RATE and PAUSING but negative residuals on GRAMMATICALITY and LEXICAL RETRIEVAL. This reflects the relative ease with which speech is produced in this syndrome but lower grammaticality and word retrieval abilities than would be expected based on their perceived FLUENCY. By contrast, those with anomic aphasia showed

Figure 2. Perceptual ratings averaged by aphasia subtype. Top graph shows mean *z* score rating; bottom graph shows mean residuals of speech-language dimensions regressed on overall fluency ratings. Higher ratings indicate greater fluency on all dimensions. ANO = anomic aphasia (n = 62); CON = conduction aphasia (n = 31); WER = Wernicke's aphasia (n = 12); BRO = Broca's aphasia (n = 69); TCM = transcortical motor aphasia (n = 7). FLU = fluency; GRAMM = grammaticality; LEX = lexical retrieval.



positive residuals for GRAMMATICALITY and LEXICAL RETRIEVAL, suggesting that these abilities are better than expected from their FLUENCY ratings, whereas SPEECH RATE, PAUSING, MELODY, and EFFORT are roughly commensurate with overall FLUENCY. Thus, it does not appear that reductions in fluency in this syndrome are well accounted for by their perceived word retrieval deficits. Rating residuals in conduction aphasia were all positive, reflecting the relative fluency of this syndrome overall, but perhaps also that what gives rise to fluency disruption in these speakers (often phonological encoding difficulty) was not well represented in the rating dimensions.

Respondents also had the option of responding UR for any of the fluency dimensions. We examined where these responses occurred to identify what made perceptual ratings of different aspects of spontaneous speech more difficult. For this analysis, we retained the clinicians' diagnoses of global aphasia (rather than combining them with Broca's aphasia), as we suspected that severity would be an important contributor to rating difficulty. Of the 10,432 ratings (1,304 × 8 scales), 137 (1.3%) had UR responses. Forty-five PwA had at least one UR response, with an average of three (range: 1–24) UR responses each within this subset. As suspected, the majority of these occurred in rating global aphasia (n = 45, 18.8% of all global aphasia ratings) or Broca's aphasia (n = 62, 1.8% of all Broca ratings). Figure 3a shows the proportion of PwA of each subtype who had at least one UR response. Subtypes with the most frequent UR responses were more nonfluent, with frequency dependent on severity: global aphasia (3/4 = 75%), Broca's aphasia (24/68 = 35%), and TCM aphasia (2/7 = 29%). Supporting this, the correlation between the proportion of UR responses and WAB AQ was −.382 (p = .001).

The 45 PwA with at least one UR response had significantly lower perceptual ratings on all dimensions than the remaining 136 PwA (all ps < .001) and significantly lower scores on half of the continuous objective measures as well, with the largest differences on WpM and MLU (both ps < .001). Other significant differences were on retracing (p = .011), grammatical complexity (p < .001), propositional density (p = .004), MATTR (p = .018), neologistic errors (p = .022), and circumlocution (p < .001). Speakers with at least one UR response were also twice as likely to have apraxia of speech as those with no URs (56% vs. 28%), although the presence of dysarthria did not differ between the groups (11% vs. 12%).

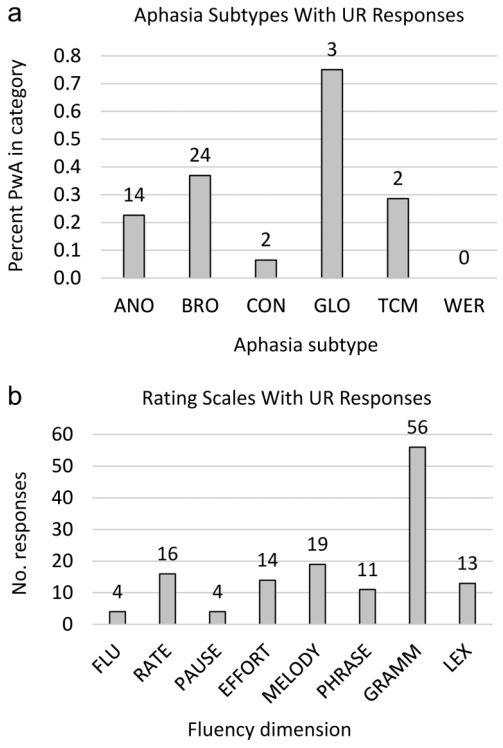
Figure 3b shows the number of UR responses (out of a total of 1,304 responses) on each perceptual rating dimension. Judgments of GRAMMATICALITY were by far the most likely to generate UR responses (4.3%), whereas overall FLUENCY and PAUSING (0.3% each) were least likely to receive UR responses. These findings suggest that certain dimensions require more connected speech than others, and judgments of grammaticality are particularly difficult when output is sparse.

Analysis 3: Conceptions of the Fluency Construct

3a: Methods of Fluency Measurement Used Clinically

The final analysis examined the post-rating responses of the 90 clinicians who completed the survey. In response to Q8 (*In the clinic, how would you usually measure or assess fluency in aphasia?*), respondents had the option of selecting any or all of nine options: five spontaneous speech dimensions

Figure 3. (a) Frequency of "unable to rate" (UR) responses for each aphasia subtype. Data labels show the raw number of individuals; *y*-axis shows the proportion of individuals in each group because numbers of each subtype vary widely. ANO = anomic aphasia; BRO = Broca's aphasia; CON = conduction aphasia; GLO = global aphasia; TCM = transcortical motor aphasia; WER = Wernicke's aphasia. (b) Frequency of UR responses for each perceptual rating scale. The total number of ratings for each scale (n = 1304) was the same. FLU = fluency; GRAMM = grammaticality; LEX = lexical retrieval.



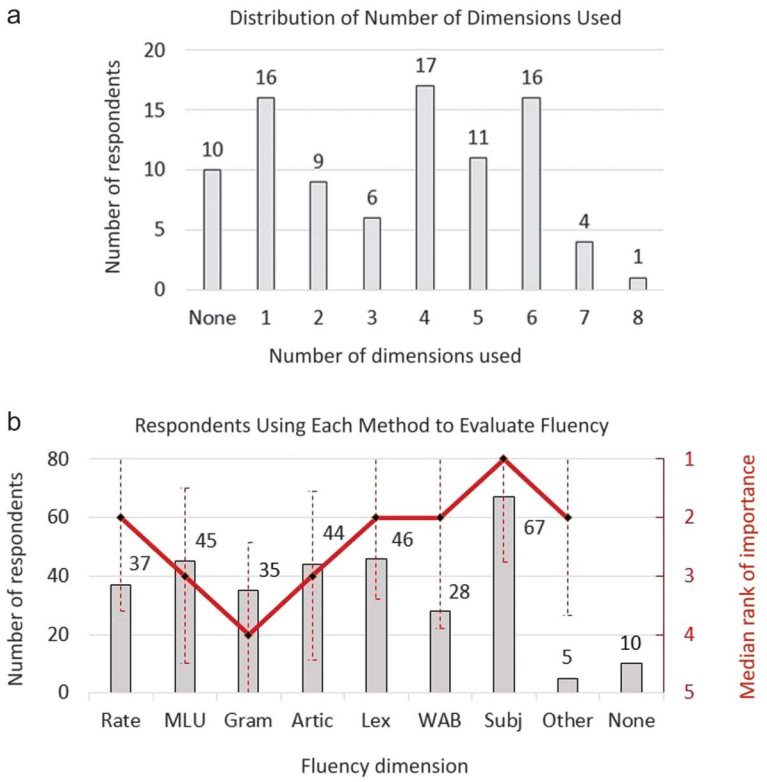
(speech rate, phrase length, grammatical competence, articulatory effort, and word retrieval), the WAB-R Fluency scale, subjective judgment, some other method specified by the respondent, and "none" (*I don't measure or assess fluency*). Figure 4a shows the number of dimensions reported by respondents. Most respondents reported using multiple methods of measuring fluency, with the mode being four methods. Of the 16 respondents who reported using only one method, 14 (88%) selected *making a subjective judgment based on one or more of the dimensions* and one selected the WAB-R scale. As both of these methods involve consideration of multiple dimensions, 88% (79/90) of all respondents reported relying on more than one dimension. Only one person relied on a single specific dimension, which was grammatical competence. Ten respondents reported that they did not measure or assess fluency; of these, four did not currently work with PwA, and five others reported aphasia caseloads of less than 20%.

Figure 4b shows the number of respondents who reported using each method to evaluate fluency in the clinic (left axis) and the average rank given to each method to indicate its importance (with 1 being *most important*). By far, the most common method was making a subjective judgment based on several dimensions (67 respondents). About half of respondents also reported measuring lexical retrieval, calculating phrase length, and assessing articulatory effort, whereas just over a third said that they typically calculate speech rate or measure some aspect of grammatical competence to assess fluency. Just under a third reported using the WAB-R spontaneous speech scores. A few respondents selected the "other" option and reported relying on measures of jargon, circumlocution, and empty speech; melodic quality; correct content units; repetition; and number of stuttering events (from a school-based clinician with minimal aphasia experience). The red line in Figure 4b reflects the median ranked importance of each dimension. The most important (i.e., the dimension most often ranked first) was subjective judgment. Assessing lexical retrieval, measuring speech rate, using WAB-R Fluency scores, and using other methods were most often ranked second; calculating MLU and assessing articulatory effort were usually ranked third; and evaluating grammatical competence was most often ranked fourth.

We examined potential sources of variability contributing to the choice of fluency assessment methods, how many were typically used, and how important they were considered to be by assessing the contributions of professional characteristics of the respondents. Results of these analyses are shown in Supplemental Material S5. In short, none of the variables assessed were shown to be strong predictors of different practices in fluency assessment. Specifically, comparing respondents with 0–10 years of experience (n = 35) to those with more than 10 years (n = 55) showed no significant difference in the distribution of respondents using the different dimensions (p = .730), in the number of dimensions typically used (p = .467), or in importance assigned to each dimension (all ps > .32). Similarly, no differences in the types of dimensions (p = .679), number used (p = .080), or rated importance (all ps > .20) were found between respondents with more than 20% PwA on their caseload (n = 45) and those with caseloads of 0%–20% PwA (n = 45). No differences in use (p = .960), number (p = .187), or importance (all ps > .11) were found between respondents who had seen 1–50 PwA (n = 44) and those who had seen more than 50 (n = 46). The 77 respondents with a master's degree also did not differ from the 13 with a PhD on the types (p = .326) or numbers of dimensions used (p = .497). PhD-level respondents did assign higher importance to the speech rate dimension than master's-level respondents (p < .001), but importance ratings did not depend on education level for any of the other dimensions (all remaining ps > .13).

The majority of respondents endorsed the idea that it is important to develop a more reliable way of measuring fluency in aphasia, with 92% of respondents giving ratings over 50 on the 100-point scale and 53% giving ratings over 80. The mean rating was 78.3. However,

Figure 4. (a) Number of respondents indicating how many dimensions they use to measure or assess fluency. (b) Number of respondents (bars and left *y*-axis) reporting that they used each fluency dimension and median importance rankings (line and right *y*-axis) for each dimension. Error bars indicate standard deviations of the ranked importance. Rate = speech rate; MLU = mean utterance length; Gram = grammaticality; Artic = articulatory facility; Lex = lexical retrieval; WAB = Western Aphasia Battery Fluency scale; Subj = subjective evaluation.



responses varied from 8 to 100. Comparing those who thought fluency assessment was less important (ratings ≤ 80, $n = 42$) to those who thought it more important (ratings > 80, $n = 48$) did not reveal any definitive reasons for this discrepancy. No significant differences were found between these subgroups in age ($p = .823$), years of experience ($p = .418$), proportion of caseload with aphasia ($p = .969$), or number of PwA seen ($p = .306$). Respondents judging reliable fluency measurement as more important had marginally higher levels of education ($p = .059$). This finding raised the possibility that the setting in which respondents worked might be the operative factor because those with PhDs mostly worked in university clinics. Indeed, raters who judged fluency measurement to be more important were more likely to work in university settings, whereas those judging it as less important were more likely to work in in-patient settings (acute care, rehab, and long-term care [LTC]) settings ($\chi^2 = 6.9$, $p = .032$).[4]

[4]For this analysis, inpatient (acute care, rehab, and LTC) settings were combined, and outpatient (private practice, home health, and outpatient office visits settings) were combined, and both of these categories were compared with university settings. Both subgroups were equally likely to work in outpatient settings.

On the basis of this finding, one more set of post hoc analyses was conducted comparing responses by work setting, following the hypothesis that inpatient settings (acute care, rehab, and LTC) would have greater time constraints than outpatient settings (private practice, home health, and outpatient clinics). Individuals working only in university settings ($n = 9$), in both inpatient and outpatient settings ($n = 5$), or not currently working ($n = 1$) were excluded. No difference was found in the distribution of dimensions used ($p = .463$) or the average number of dimensions used by each respondent ($p = .662$). In the rated importance of the different dimensions, a significant difference was found only for the WAB-R scale, with respondents in inpatient settings rating the scale higher ($n = 25$, $M = 1.9$) than those in outpatient settings ($n = 50$, $M = 3.9$, $p = .046$).

3b. Open-Ended Responses

The open-ended question (Q11: *Please add any suggestions, feedback, or other comments in the box below*) received responses from 49 (54%) of the respondents. Sixty-three discrete responses were identified and classified into three broad categories, as follows. Independent agreement on the categories was 87%; discrepancies were

resolved by discussion. Many included (a) *an expression of thanks or appreciation* for the importance of the research study (44%). About 19% commented on (b) *the survey format or the respondent's experience in taking the survey*, with suggestions such as including samples at the extreme ends of the fluency continuum or comment boxes to provide rationales for perceptual ratings. One respondent noted a tendency to rate speakers more harshly throughout the experiment, which may also indicate a need for training to calibrate clinicians on the scale. A post hoc analysis checked to see if this issue was widespread by correlating individual ratings with the order of presentation of the PwA samples. Correlations for each of the rating dimensions were smaller than .10 (the typical minimum benchmark indicating a small but meaningful effect; Cohen, 1988), indicating that order of presentation did not have a systematic effect on the ratings.

Over a third of the comments (37%) had to do with (c) *the measurement of fluency*. These were of most interest in this study, so verbatim responses (edited for length) are provided in the Appendix. Within this category, five subthemes were identified: (a) Several respondents commented on the complexity of fluency measurement, that is, the number of dimensions that contribute to impressions of fluency. (b) Related to this, a few of the respondents singled out word retrieval as an important component of fluency. (c) Some pointed out the extent to which conceptions of fluency vary by individual or by task. (d) A few respondents made the case that fluency measurement should be defined more broadly than verbal expression, taking into consideration aspects of nonverbal communication and the extent to which fluency disruptions in PwA affect activity and participation. (e) The final category identified more specific issues (e.g., time limitations) and suggestions regarding the measurement of fluency in clinical settings.

Discussion

This study sought to round out our understanding of what contributes to fluency perceptions by collecting from SLPs perceptual ratings of fluency based on audio samples from a range of individuals with aphasia and by analyzing the reliability and validity of the ratings, as well as respondents' ideas of how fluency is and should be measured clinically.

Reliability of Fluency Ratings

According to the two-way ICC models, all the speech-language dimensions showed acceptable levels of reliability, although reliability was lower (dipping into the "fair" range) for ratings of EFFORT and LEXICAL RETRIEVAL. This can be attributed to the ill-defined nature of the effort construct, which is subjective by nature, and to the fact that lexical retrieval difficulties may be difficult to identify in connected speech (Gordon & Kindred, 2011; Kavé & Nussbaum, 2012). The most reliable scales were SPEECH RATE, PAUSING, and PHRASE LENGTH, reinforcing their importance for fluency measurement.

Comparison of agreement and consistency models indicated that there was little systematic bias in how the raters used the scale. However, comparison of the one-way and two-way models suggested that there was a significant amount of variance attributable to raters, which is most likely related to some degree to the fact that respondents rated different subsets of the PwA. Such variance, although apparently not due to rater bias, should also not be considered error and should be taken into account. In this respect, the two-way models provide more appropriate estimates of interrater reliability. The impact of rater variance was also illustrated in the differences between single-rater and average-rater models. Averaging scores across raters considerably improved reliability estimates (as it did the magnitude of the correlations between perceptual ratings and objective measures). This suggests that fluency ratings can be quite reliable when the averages of several raters are used (a finding also reported by Casilio et al., 2019), and this would be a firm recommendation for using such measures in research. However, the average-rater reliabilities should not be generalized to clinical practice, where fluency is almost always judged by a single clinician. Thus, the need for a more reliable measure of fluency remains.

Relationship of Fluency Ratings to Objective Measures

Respondents' judgments about fluency were most strongly influenced by speakers' rate of speech, utterance length, and grammatical complexity (Analysis 2). This is consistent with long-standing conceptions that speech rate (Gordon & Clough, 2020; Halai et al., 2017; Howes, 1964; Nozari & Faroqi-Shah, 2017; Vermeulen et al., 1989; Wang et al., 2013) and utterance length (Goodglass et al., 1964; Gordon & Clough, 2020; Halai et al., 2017; Helm-Estabrooks, 1992; Vermeulen et al., 1989) serve as valid proxy measurements for fluency. Kerschensteiner et al. (1972) demonstrated that utterance length and pausing (which is closely related to speech rate) were most useful in discriminating between fluent and nonfluent aphasia. A factor analysis conducted by Vermeulen et al. (1989) showed that speech rate and MLU had the strongest loadings on their first factor, which represented fluency. More recently, Park et al. (2011) found that fluent/nonfluent classifications were best predicted by a combination of speech rate (syllables per minute) and speech productivity (proportion of time spent talking, that is, the

inverse of pause time). However, this study did not include any measures of utterance length or syntactic formulation.

Unfortunately, the identification of speech rate and utterance length as important to fluency does not take us very far in advancing our understanding of fluency impairments. Feyereisen et al. (1991) referred to these measures as "shallow measures," because they are unable to point to the "defective mechanism" that results in dysfluency. Survey respondents did, however, show sensitivity to more specific aspects of spontaneous speech, including an important impact of measured grammatical complexity on all the perceptual rating dimensions, of phonological errors on perceived EFFORT, of pitch variability on perceived MELODY, of grammatical errors and grammatical complexity on perceived GRAMMATICALITY, and of lexical diversity on perceived LEXICAL RETRIEVAL. These associations between perceptual ratings and objective measures help to validate the clinical ratings and identify some of the more specific aspects of production that underlie these perceptions.

The strong influence of grammatical complexity on fluency is also consistent with prior findings. In a factor analysis of spontaneous speech, Wagenaar et al. (1975) identified a fluency factor that included strong loadings of utterance length, utterance complexity, and speech tempo. Among the variables examined by Nozari and Faroqi-Shah (2017) using a path modeling approach, only their composite measure of syntactic production had a reliable direct effect on fluency (measured by the WAB-R Fluency scale and speech rate). In our own prior work, grammatical complexity was the strongest predictor of each of three fluency proxy measures—speech rate, MLU, and retracing—and among the strongest predictors of the WAB-R Fluency scale scores (Gordon & Clough, 2020), as well as binary fluency classifications based on the WAB-R scale (Clough & Gordon, 2020). Notably, grammatical measures did not contribute significantly to binary fluency classifications based on clinical impression (Clough & Gordon, 2020), a finding discussed further below.

Relationship of Fluency Ratings to Aphasia Subtypes

To further validate the clinicians' perceptual ratings, we examined mean values on each dimension by aphasia subtype (Analysis 2). For the most part, ratings reflected expected syndrome patterns, with maximum contrast between Broca's aphasia and Wernicke's aphasia, particularly on measures of SPEECH RATE, PAUSING, and PHRASE LENGTH. The patterns of rating residuals, which factored out the effect of overall FLUENCY, generated insights about specific dimensions that were perceived to differ among the syndromes. For example, although Broca's and Wernicke's aphasia remained maximally distinct in PHRASE LENGTH using the residual measures, it was the individuals with TCM aphasia who contrasted most with Wernicke's aphasia on SPEECH RATE and PAUSING, illustrating that these dimensions were perceived to be particularly disruptive to spontaneous speech production in TCM aphasia. Individuals with Wernicke's aphasia received above-average ratings on GRAMMATICALITY and LEXICAL RETRIEVAL, but residuals of both dimensions were below average, indicating that they were judged to be more impaired than would be expected from the speakers' level of rated FLUENCY. These observations are consistent with widely recognized deficits in Wernicke's aphasia: Despite the ability to produce long utterances, the structure of phrases is often distorted by paragrammatism (Bastiaanse et al., 1996; Goodglass et al., 2001a; Gordon & Slater, 2008; Matchin et al., 2020) and the content by paraphasic substitutions (Edwards, 2005; Goodglass et al., 2001a). By contrast, these same dimensions were better than predicted by FLUENCY ratings for anomic aphasia. The perception of relatively good lexical retrieval seems surprising for this group but may relate to ambiguity in the source of disfluency, particularly in speakers who can circumlocute around their word-finding difficulties in connected speech (Gordon & Kindred, 2011; Kavé et al., 2009).

Analysis by aphasia subtype also revealed that perceptual ratings of spontaneous speech are particularly difficult when output is sparse. Most notably, three quarters of the speakers with global aphasia received responses of UR on at least one speech-language dimension, as did about a third of those with Broca's and TCM aphasia. The paradox that fluency is more difficult to measure in individuals with disrupted fluency has been previously noted by Feyereisen et al. (1991). This problem arises partly because there is less available evidence to use for clinical assessment and partly because the available output is likely to be affected by multiple underlying deficits. Relatedly, UR responses were also found to be most frequent in judgments of grammaticality, which require a sufficient number of phrasal combinations. Goodglass et al. (2001a) recommend that the mostly single-word utterances of individuals with global and severe Broca's aphasia be characterized as "pseudo-agrammatism" because there is insufficient evidence upon which to judge grammatical competence.

Clinical Methods of Measuring Fluency

Most respondents reported assessing fluency with multiple measures. Although a third to half of respondents reported using some combination of lexical retrieval, MLU, articulatory effort, speech rate, and grammaticality measures, almost three quarters used subjective evaluation

that takes into account multiple dimensions. This method was also rated highest in importance, on average. This emphasis is likely related, at least in part, to the lack of availability of a more objective multidimensional tool, because the overwhelming majority of respondents endorsed the need for such a tool. Interestingly, the WAB-R Fluency scale, developed for this purpose, was endorsed by the fewest respondents (31%), suggesting an awareness of the scale's shortcomings. Although used less frequently, the WAB-R scale was nonetheless rated relatively high in importance by those who did use it. This result turned out to be driven by clinicians in inpatient settings, who rated the WAB-R scale as significantly more important than did those in outpatient settings.

The speech-language dimension measured least frequently and ranked lowest in importance was grammatical competence. This is surprising, given the importance of grammatical complexity in both predicting speech rate and utterance length (Gordon & Clough, 2020) and discriminating between fluent and nonfluent categories of aphasia based on the WAB-R scale (Clough & Gordon, 2020). In addition, in an earlier study in which clinicians were asked to identify *the most salient factor influencing the judgment of expressive language as "fluent" or "nonfluent,"* grammatical complexity was the most frequently cited aspect of spontaneous speech (Gordon, 1998). The difference between this finding and this study might be related to the nature of the question asked. Gordon (1998) asked what dimensions were most salient for classifying fluency; in this study, respondents were asked what they actually did in the clinic and how important they judged this method to be. The difference may lie in the extent to which clinicians consider the measurement of grammaticality to be a *feasible* method in practice. Notably, the classification of fluent *versus* nonfluent aphasia by *clinical impression* in the Clough and Gordon (2020) study (unlike classifications based on the WAB-R scale) did not include grammatical complexity as a significant predictor, lending support to the current findings in suggesting that clinicians did not find this dimensions to be as informative as other dimensions.

Barriers to Fluency Measurement

The variation in dimensions used and the importance ascribed to them bore no strong relationship to the respondents' experience with aphasia, whether measured by years of experience, number of PwA seen professionally, percentage of caseload with aphasia, or clinical setting. The only significant differences observed were (a) respondents with a doctoral degree (70% of whom worked in university settings) were more likely to rely on speech rate than respondents with a master's degree (who worked in a variety of settings); (b) respondents working in inpatient settings considered the WAB-R scale to be a more important measure of fluency than those in outpatient settings; and (c) those who considered the need for a better fluency measure to be greater were more likely to work in university settings, whereas those rating the need as less important were more likely to work in inpatient settings. These findings are all likely reflections of the time available for assessment in different settings. Inpatient settings are typically more constrained, with the result that clinicians place more value on quicker methods such as the WAB-R scale or subjective evaluation. University clinics, on the other hand, are typically guided less by efficiency and more by their teaching mission, which might explain the greater importance placed on calculating speech rate, a relatively time-consuming method of assessing fluency. In the open-ended responses, one respondent noted that, although they did not calculate speech rate in the clinic, they would rank it high in importance, suggesting that lack of use does not necessarily imply a perceived lack of importance.

Findings from previous surveys reinforce the idea that factors beyond clinicians' preferences or beliefs affect intervention practices. Of 10 general approaches to therapy, Australian clinicians (Rose et al., 2014) reported that discourse-based treatment was one of the least often used and that this was related to limitations in knowledge of and confidence with the approach. A survey by Bryant et al. (2017) focusing specifically on discourse analysis identified similar barriers. Although over 50% of the respondents agreed or strongly agreed with the statement that "Detailed linguistic analysis of discourse is important for the assessment of language in aphasia," only 30% endorsed the statement "I feel confident using discourse analysis to assess language in aphasia." Only 60% used discourse analysis at least some of the time; among these, 64% reported generating written transcripts and only 39% recorded samples. The most commonly reported dimensions of discourse analyzed were word-finding difficulty (~95%) and sentence structure (~80%); only 50% mentioned rate of speech. However, the specific *measures* most frequently reported (word counts, MLU, correct information units, and paraphasias) tended to focus on the word level, and none examined syntactic structure, consistent with the infrequent reliance on grammaticality reported in this study. The most frequently cited factors limiting use and depth of discourse analysis were lack of time and other resources and lack of relevant training or expertise. Similar findings have been found when surveying clinicians working with individuals with traumatic brain injury (Frith et al., 2014; Maddy et al., 2015). As in this study, practice differences in these studies were not accounted for by clinical experience (Bryant et al., 2017; Frith et al., 2014).

In the survey of Bryant et al. (2017), clinicians not only acknowledged the value of discourse analysis in

aphasia assessment, including recording and transcription, but also expressed the need for greater efficiency of discourse analysis in clinical context. The current findings echo this: The majority of respondents strongly endorsed the need for a better method of evaluating fluency, but their responses also revealed a reluctance to use time-intensive measures such as speech rate or grammaticality measures in clinical practice. Given the findings from previous surveys, it is likely that a lack of confidence in measuring grammaticality, as well as time limitations, contribute to its lack of use.

Limitations of This Study

We acknowledge several limitations in this study. First, the sample was relatively small, given the number of SLPs working with aphasia. Although the sample of respondents was sufficiently variable to allow some exploration of clinician variables, we may not have had sufficient power to identify significant differences. Our ability to identify rater-specific sources of variance may also have been limited by the design of the study, that is, the fact that not all raters rated each PwA. Although we used an analysis method specifically designed to accommodate missing data, the unbalanced design reduced the potential of the analysis to identify systematic rater biases. The generalizability of our conclusions may also be limited due to the single task used. Fluency judgments were based solely on audio samples of a storytelling task, and it is known that characteristics of spontaneous speech vary with elicitation context (e.g., Fergadiotis et al., 2011; Stark & Fukuyama, 2021). We expect that the task or context would be more likely to affect the *degree of fluency* than the *predictors of fluency* that are important for a given speaker, although we acknowledge that this is an untested assumption. An additional implication of relying on audio samples is the loss of visual information (e.g., eye contact, facial expression, and gestures) that is typically present in clinical interactions. Finally, as with most structured surveys, the options offered to respondents may have biased or limited their choices. For example, in asking participants about their use of fluency measures, we offered only one standardized test option—the WAB-R scale. Participants had the option to write in other options (e.g., the BDAE rating scales) but focusing on the WAB-R scale may have overemphasized its use.

Next Steps in Fluency Measurement

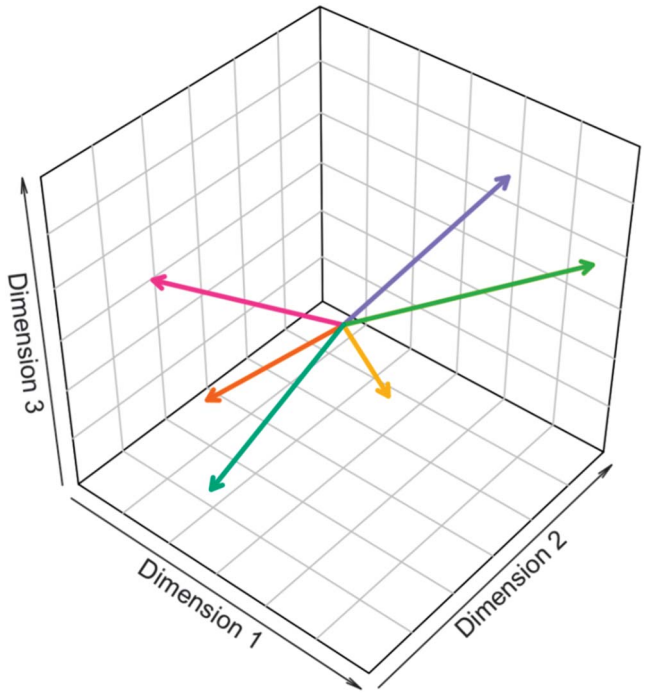
It is clear from this and past work that what is needed is a fluency measure that incorporates multiple dimensions but is clinically feasible to use, that provides greater objectivity than current methods, and that helps identify the deficit or deficits underlying dysfluency. We are currently developing such a measure based on findings from our earlier studies of speech-language dimensions contributing to dichotomous judgment (Clough & Gordon, 2020) and continuous measures (Gordon & Clough, 2020) of fluency, a factor analysis of spontaneous speech (Gordon, 2020), and the results of this study.

Several of the open-ended responses reinforce the direction that we are taking. First, fluency is complex, and its measurement must therefore allow for the consideration of multiple dimensions. One astute respondent noted that, for this reason, "fluent" and "nonfluent" may not be opposites; because speakers vary along multiple dimensions, they might be considered fluent on some aspects of spontaneous speech and nonfluent on others. Thus, a more nonfluent individual may not be the diametrical opposite of a more fluent individual. Figure 5 illustrates this concept by displaying dimensions of fluency as oblique vectors in multidimensional space rather than a single two-dimensional line.

Related to this, there is no one-to-one correspondence between overt behaviors and underlying deficits. A given impairment, such as lexical retrieval difficulty, may manifest in various ways (e.g., pausing, paraphasias, and sentence fragments). Similarly, a given behavior, such as pausing, may arise for different reasons (e.g., word-finding problems and syntactic formulation difficulties). As is evident from the preceding point, word retrieval difficulties have significant implications for fluency, a point mentioned by several of the respondents. This in itself calls into the question the validity of dichotomous classifications of fluency, because the most common type of "fluent aphasia" is anomic aphasia. Finally, disruptions in fluency depend on characteristics of the task and the individual; correspondingly, the dimensions of fluency that matter may vary by aphasia type, individual, and task. Intraindividual variability can result in what one respondent described as a given PwA being perceived as both fluent and nonfluent depending on the task and the dimensions deemed to be salient, which further calls into question the use of a dichotomous classification system.

Finally, a few of the respondents encouraged us to think beyond linguistic aspects of verbal production to view fluency with a wider lens. One suggestion was to include nonverbal communication. In particular, there is increasing interest in the types and functions of gestures PwA (and other neurogenic communication disorders) produce and how those gestures support communication and complement verbal output (e.g., Clough & Duff, 2020). In studies of gesture production, PwA produce higher rates of gestures per word (Carlomagno & Cristilli, 2006; de Beer et al., 2019; Feyereisen, 1983; Sekine et al., 2013; but, see Pritchard et al., 2015) and a larger variety of gesture types (Sekine & Rose, 2013) than neurotypical comparison participants. Moreover, PwA can use gesture to facilitate communication when speech fails. They are

Figure 5. Fluency represented as vectors in multidimensional space. Each vector represents an individual with aphasia and their performance on three hypothetical speech-language dimensions. The tendency for dimensions to correlate is represented in the directionality of the vectors, which cluster toward the high or low ends of each dimension, with variation in the extent to which individual dimensions are affected.



more likely than individuals without brain damage to produce essential gestures that convey information that is not present in the speech signal (Pritchard et al., 2015; van Nispen et al., 2017). The use of audio samples in this study prevented evaluation of nonverbal communication (e.g., gesture and eye gaze) by respondents; however, such nonverbal signals can contribute to the meaning of a communicated message and facilitate the flow of ideas between interlocutors. Research on the role of fluency in predicting gesture use has been equivocal, sometimes showing that more or more meaningful gestures are produced in fluent aphasia (Cicone et al., 1979), sometimes in nonfluent aphasia (Kong et al., 2017; Sekine et al., 2013), and sometimes showing no difference (Feyereisen, 1983). It is an open question how gestures might contribute to listener perceptions of fluency in aphasia.

Another comment was to take into account the role of fluency in cooperating with a listener in more interactive types of tasks, such as conversation. One respondent suggested that the assessment of fluency should consider its impact on the domains of activity and participation, how the facility of verbal production helps "connect the PwA in society." Although these concepts go beyond traditional definitions of *verbal* fluency in aphasiology, they are certainly relevant to the pragmatic functions of *communicative* fluency. Fluent language production signals to a listener that a speaker is still attempting to communicate a message. Failing linguistic fluency, PwA may make use of nonverbal fillers ("uh, um"), sound effects, or interactive gestures, that is, gestures that coordinate dialogue by, for example, passing a turn or holding the floor (Bavelas et al., 1992). Indeed, PwA produce more interactive gestures than neurotypical comparison participants in both spontaneous conversation and narrative retellings (de Beer et al., 2019), suggesting a greater reliance on nonverbal means to facilitate turn taking. How successfully a person with aphasia can participate in communicative tasks—whether verbally or nonverbally—is critical to their ability to participate in society.

We are taking these comments to heart in planning our next steps. To be clinically feasible, a fluency measurement tool must be fairly quick to administer. This and past studies (e.g., Casilio et al., 2019) suggest that ratings can be used to efficiently capture relevant dimensions of spontaneous speech. However, rating scales can be unreliable across clinicians (e.g., Gordon, 1998; Trupe, 1984), which may have implications for the accuracy of aphasia classification and the ability to identify appropriate and specific therapy targets. Thus, a clear protocol for implementing ratings is needed to ensure their reliability. Reliability will be further strengthened with the complementary use of objective measures, as long as the calculations are straightforward and consistently implemented. In addition, guidance is clearly needed regarding the measurement of grammaticality. To maximize internal validity, a fluency measurement tool must include measures to identify whether fluency is disrupted by lexical retrieval problems, grammatical formulation problems, or more peripheral aspects of speech production (prosody and motor speech), so that therapy can be appropriately directed. External validity, as pointed out by our survey respondents, will be enhanced by considering the communicative impact of fluency reductions in various spontaneous speech contexts and at International Classification of Functioning, Disability and Health levels of activity and participation.

Acknowledgments

This work was generously supported by a New Century Scholar grant from the American Speech-Language-Hearing Foundation. The authors would also like to acknowledge the developers and contributors to AphasiaBank, with special thanks to Davida Fromm for sharing her expertise. We are also grateful for the help provided by several industrious research assistants, notably Chaewon Park, Olivia Sourwine, and Jenna Kelly. We are also grateful to the support of the Iowa Social Science Research Center

(particularly Cassidy Branch) for their help in developing, administering, and managing the survey.

References

American Speech-Language-Hearing Association. (2021a). *Profile of ASHA members and affiliates, year-end 2020*. https://www.asha.org

American Speech-Language-Hearing Association. (2021b). *Profile of ASHA members and affiliates with PhDs, year-end 2020*. https://www.asha.org

Avidemux 2.6. (2017). [Computer software]. https://avidemux.org/

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports, 19*(1), 3–11. https://doi.org/10.2466/pr0.1966.19.1.3

Bastiaanse, R., Edwards, S., & Kiss, K. (1996). Fluent aphasia in three languages: Aspects of spontaneous speech. *Aphasiology, 10*(6), 561–575. https://doi.org/10.1080/02687039608248437

Bavelas, J. B., Chovil, N., Lawrie, D. A., & Wade, A. (1992). Interactive gestures. *Discourse Processes, 15*(4), 469–489. https://doi.org/10.1080/01638539209544823

Brueckl, M., & Heuer, F. (2021). *irrNA: coefficients of interrater reliability—Generalized for randomly incomplete datasets*. https://CRAN.R-project.org/package=irrNA

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

Carlomagno, S., & Cristilli, C. (2006). Semantic attributes of iconic gestures in fluent and non-fluent aphasic adults. *Brain and Language, 99*(1–2), 104–105. https://doi.org/10.1016/j.bandl.2006.06.061

Casilio, M., Rising, K., Beeson, P. M., Bunton, K., & Wilson, S. M. (2019). Auditory-perceptual rating of connected speech in aphasia. *American Journal of Speech-Language Pathology, 28*(2), 550–568.

Cichetti, 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

Cicone, M., Wapner, W., Foldi, N., Zurif, E. B., & Gardner, H. (1979). The relation between gesture and language in aphasic communication. *Brain and Language, 8*(3), 324–349. https://doi.org/10.1016/0093-934X(79)90060-9

Clough, S., & Duff, M. C. (2020). The role of gesture in communication and cognition: Implications for understanding and treating neurogenic communication disorders. *Frontiers in Human Neuroscience, 14,* 323. https://doi.org/10.3389/fnhum.2020.00323

Clough, S., & Gordon, J. K. (2020). Fluent or nonfluent? Part A. Underlying contributors to categorical classifications of fluency in aphasia. *Aphasiology, 34*(5), 515–539. https://doi.org/10.1080/02687038.2020.1727709

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

de Beer, C., de Ruiter, J. P., Hielscher-Fastabend, M., & Hogrefe, K. (2019). The production of gesture and speech by people with aphasia: Influence of communicative constraints. *Journal of Speech, Language, and Hearing Research, 62*(12), 4417–4432. https://doi.org/10.1044/2019_JSLHR-L-19-0020

Edwards, S. (2005). *Fluent aphasia*. Cambridge University Press. https://doi.org/10.1017/CBO9780511486548

Eysenbach, G. (2004). Improving the quality of web surveys: The Checklist for Reporting Results of Internet E-Surveys (CHERRIES). *Journal of Medical Internet Research, 6*(3), e34. https://doi.org/10.2196/jmir.6.3.e34

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

Feyereisen, P. (1983). Manual activity during speaking in aphasic subjects. *International Journal of Psychology, 18*(1–4), 545–556. https://doi.org/10.1080/00207598308247500

Feyereisen, P., Pillon, A., & De Partz, M.-P. (1991). On the measures of fluency in the assessment of spontaneous speech production by aphasic subjects. *Aphasiology, 5*(1), 1–21. https://doi.org/10.1080/02687039108248516

Frith, M., Togher, L., Ferguson, A., Levick, W., & Docking, K. (2014). Assessment practices of speech-language pathologists for cognitive communication disorders following traumatic brain injury in adults: An international survey. *Brain Injury, 28*(13–14), 1657–1666. https://doi.org/10.3109/02699052.2014.947619

GoldWave 6.31. (2017). *GoldWave* (Version 6.27). https://www.goldwave.com/goldwave.php

Goodglass, H., Kaplan, E., & Barresi, B. (2001a). *The assessment of aphasia and related disorders* (3rd ed.). Lippincott Williams & Wilkins.

Goodglass, H., Kaplan, E., & Barresi, B. (2001b). *Boston Diagnostic Aphasia Examination–Third Edition*. Lippincott Williams & Wilkins.

Goodglass, H., Quadfasel, F. A., & Timberlake, W. H. (1964). Phrase length and the type and severity of aphasia. *Cortex, 1*(2), 133–153. https://doi.org/10.1016/S0010-9452(64)80018-6

Gordon, J. K. (1998). The fluency dimension in aphasia. *Aphasiology, 12*(7–8), 673–688. https://doi.org/10.1080/02687039808249565

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

Gordon, J. K., & Clough, S. (2020). How fluent? Part B. Underlying contributors to continuous measures of fluency in aphasia. *Aphasiology, 34*(5), 643–663. https://doi.org/10.1080/02687038.2020.1712586

Gordon, J. K., & Kindred, N. K. (2011). Word retrieval in ageing: An exploration of the task constraint hypothesis. *Aphasiology, 25*(6–7), 774–788. https://doi.org/10.1080/02687038.2010.539699

Gordon, J. K., & Slater, M. (2008). *Understanding paragrammatism: A comparative case study*. Paper presented at the Clinical Aphasiology Conference, Jackson Hole, WY, United States.

Halai, A. D., Woollams, A. M., & Lambon Ralph, M. A. (2017). Using principal component analysis to capture individual differences within a unified neuropsychological model of chronic post-stroke aphasia: Revealing the unique neural correlates of speech fluency, phonology and semantics. *Cortex, 86,* 275–289. https://doi.org/10.1016/j.cortex.2016.04.016

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23–34. https://doi.org/10.20982/tqmp.08.1.p023

Helm-Estabrooks, N. (1992). *Aphasia Diagnostic Profiles*. Pro-Ed.

Holland, A. L., Fromm, D., & Swindell, C. S. (1986). The labeling problem in aphasia: An illustrative case. *Journal of Speech and Hearing Disorders, 51*(2), 176–180. https://doi.org/10.1044/jshd.5102.176

Howes, D. (1964). Application of the word-frequency concept to aphasia. In A. V. S. DeReuck & M. O'Connor (Eds.), *Disorders of language* (pp. 47–78). J. A. Churchill. https://doi.org/10.1002/9780470715321.ch4

Kavé, G., & Nussbaum, S. (2012). Characteristics of noun retrieval in picture descriptions across the adult lifespan. *Aphasiology, 26*(10), 1238–1249. https://doi.org/10.1080/02687038.2012.681767

Kavé, G., Samuel-Enoch, K., & Adiv, S. (2009). The association between age and the frequency of nouns selected for production. *Psychology and Aging, 24*(1), 17–27. https://doi.org/10.1037/a0014579

Kerschensteiner, M., Poeck, K., & Brunner, E. (1972). The fluency–non fluency dimension in the classification of aphasic speech. *Cortex, 8*(2), 233–247. https://doi.org/10.1016/S0010-9452(72)80021-2

Kertesz, A. (2006). Western Aphasia Battery–Revised. Pearson.

Kong, A. P.-H., Law, S. P., & Chak, G. W. C. (2017). A comparison of coverbal gesture use in oral discourse among speakers with fluent and nonfluent aphasia. *Journal of Speech, Language, and Hearing Research, 60*(7), 2031–2046. https://doi.org/10.1044/2017_JSLHR-L-16-0093

Liljequist, D., Elfving, B., & Roaldsen, K. S. (2019). Intraclass correlation—A discussion and demonstration of basic features. *PLOS ONE, 14*(7), Article e0219854. https://doi.org/10.1371/journal.pone.0219854

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk: Transcription format and programs* (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

Maddy, K. M., Howell, D. M., & Capilouto, G. J. (2015). Current practices regarding discourse analysis and treatment following non-aphasic brain injury: A qualitative study. *Journal of Interactional Research in Communication Disorders, 6*(2), 211–236. https://doi.org/10.1558/jircd.v7i1.25519

Matchin, W., Basilakos, A., Stark, B. C., den Ouden, D.-B., Fridrikkson, J., & Hickok, G. (2020). Agrammatism and paragrammatism: A cortical double dissociation revealed by lesion–symptom mapping. *Neurobiology of Language, 1*(2), 208–225. https://doi.org/10.1162/nol_a_00010

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

Nguyen, A. A., & Fabrigar, L. R. (2018). Visual Analog Scales. In B. B. Frey (Ed.), *The SAGE encyclopedia of educational research, measurement, and evaluation* (pp. 1797–1800). SAGE.

Nozari, N., & Faroqi-Shah, Y. (2017). Investigating the origin of nonfluency in aphasia: A path modeling approach to neuropsychology. *Cortex, 95,* 119–135. https://doi.org/10.1016/j.cortex.2017.08.003

Park, H., Rogalski, Y., Rodriguez, A. D., Zlatar, Z., Benjamin, M., Harnish, S., Bennett, J., Rosenbek, J. C., Crosson, B., & Reilly, J. (2011). Perceptual cues used by listeners to discriminate fluent from nonfluent narrative discourse. *Aphasiology, 25*(9), 998–1015. https://doi.org/10.1080/02687038.2011.570770

Poeck, K. (1989). Fluency. In C. Code (Ed.), *The characteristics of aphasia* (pp. 23–32). Taylor & Francis.

Pritchard, M., Dipper, L., Morgan, G., & Cocks, N. (2015). Language and iconic gesture use in procedural discourse by speakers with aphasia. *Aphasiology, 29*(7), 826–844. https://doi.org/10.1080/02687038.2014.993912

Rose, M., Ferguson, A., Power, E., Togher, L., & Worrall, L. E. (2014). Aphasia rehabilitation in Australia: Current practices, challenges and future directions. *International Journal of Speech-Language Pathology, 16*(2), 169–180. https://doi.org/10.3109/17549507.2013.794474

Sekine, K., & Rose, M. L. (2013). The relationship of aphasia type and gesture production in people with aphasia. *American Journal of Speech-Language Pathology, 22*(4), 662–672. https://doi.org/10.1044/1058-0360(2013/12-0030)

Sekine, K., Rose, M. L., Foster, A. M., Attard, M. C., & Lanyon, L. E. (2013). Gesture production patterns in aphasic discourse: In-depth description and preliminary predictions. *Aphasiology, 27*(9), 1031–1049. https://doi.org/10.1080/02687038.2013.803017

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

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

Swindell, C. S., Holland, A., & Fromm, D. (1984). *Classification of aphasia: WAB type versus clinical impression*. Paper presented at the Clinical Aphasiology Conference, Baltimore, MD, United States.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*(1), 25–29. https://doi.org/10.1037/h0071663

Trupe, E. H. (1984). *Reliability of rating spontaneous speech in the Western Aphasia Battery: Implications for classification*. Paper presented at the Clinical Aphasiology Conference, Baltimore, MD, United States.

van Nispen, K., van de Sandt-Koenderman, M., Sekine, K., Krahmer, E., & Rose, M. L. (2017). Part of the message comes in gesture: How people with aphasia convey information in different gesture types as compared with information in their speech. *Aphasiology, 31*(9), 1078–1103. https://doi.org/10.1080/02687038.2017.1301368

Vermeulen, J., Bastiaanse, R., & Van Wageningen, B. (1989). Spontaneous speech in aphasia: A correlational study. *Brain and Language, 36*(2), 252–274. https://doi.org/10.1016/0093-934X(89)90064-3

Wagenaar, E., Snow, C., & Prins, R. (1975). Spontaneous speech of aphasic patients: A psycholinguistic analysis. *Brain and Language, 2*(3), 281–303. https://doi.org/10.1016/S0093-934X(75)80071-X

Wang, J., Marchina, S., Norton, A. C., Wan, C. Y., & Schlaug, G. (2013). Predicting speech fluency and naming abilities in aphasic patients. *Frontiers in Human Neuroscience, 7,* 1–13. https://doi.org/10.3389/fnhum.2013.00831

Wertz, R. T., Deal, J. L., & Robinson, A. J. (1984). *Classifying the aphasias: A comparison of the Boston Diagnostic Aphasia Examination and the Western Aphasia Battery*. Paper presented at the Clinical Aphasiology Conference.

Appendix (p. 1 of 2)

Open-Ended Responses Commenting on the General Topic of Fluency Measurement

Subtheme 1: Fluency Measurement Is Complex

I am always second-guessing my fluency assessment of any given patient, because there are so many dimensions that one can look at to decide whether someone is fluent or nonfluent. It is a very nebulous concept, yet somehow all of our diagnoses are based upon that one fundamental distinction.

I have seen a range of patients from nonverbal to verbal with difficulties in all parameters. There are too many variables that could affect the fluency. I think each variable would need to be assessed and the fluency for each variable.

There is a huge gray area in measuring fluency versus intelligibility versus word retrieval/language challenges. Having a more objective way to distinguish among these areas of speech would be a great benefit.

Complexities to speech patterns necessitate better assessment methods.

What is the ultimate goal in finding a more reliable assessment? Is it determining underlying cause of the nonfluency, such as it is a matter of motor performance or language/word retrieval performance? Are we measuring fluency as a matter of sound repetitions or word repetitions in connected speech? How are we defining fluency in aphasia terms?

I found myself questioning whether the articulatory effort, pauses, and word-finding were influencing my fluency judgments. I attempted to take all measurements into consideration when judging fluency.

Subtheme 2: Importance of Word Retrieval

Struggling with word retrieval impacts all other aspects of measuring fluency. If the patient is grasping for words, there will be pauses, struggle with grammar and articulation.

I usually thought of aphasia as a word-finding difficulty that caused the fluency issue. So is there now research showing a brain injury or stroke can have effects on fluency and not so much considered word finding?

I have gone through stages of approaching fluency from a word-finding perspective, to not using the term at all, and now that I am teaching I have decided to kowtow to my impression of the traditional approach (aka a Wernicke's type of fluency).

Subtheme 3: Fluency Measurement Is Variable

I think the definition of fluency can vary, which can impact a person's response.

It is definitely a subjective measure that is more reliable with greater experience.

If I hear a person who seems to hesitate and self-correct (or attempt to) a lot, who struggles, who pauses, but who still emits functor words and some complex syntactical structures—I am thinking here of the conduction aphasic person—I will classify that person as fluent, even though a narrative connected speech sample would seem "not very fluent" to the non-speech pathologist.

Doing this survey made me think about which dimensions I listen for when deciding if someone is fluent or nonfluent, and I found that it varied across patients. I also found that the dimensions I feel are important for a judgment of "nonfluent" are not exactly the same as the dimensions I feel are important for a judgment of "fluent." That is, they do not map on directly. So, when I hear an effortful, slow speaker with extensive word-finding issues, I will judge their fluency more on articulation or pausing, but if I hear a speaker who has a normal rate of speech, I will pay more attention to grammaticality and word retrieval. This is intuitive, I suppose given our training, but I think it is important to note that "fluent" and "nonfluent" may not be exact opposites.

More often than not, I will think a patient is both fluent and nonfluent simultaneously, depending on the task at hand. For example, someone with word finding difficulty, who is otherwise a very fluent speaker—are they fluent or nonfluent? Alternatively, someone with apraxia of speech, who otherwise has very minimal aphasia—are they fluent or nonfluent? Someone who speaks fluently when they do have islands of continuous speech, but who is very hesitant and puts in a lot of effort pre- and post-islands of speech—are they fluent or nonfluent? The list goes on.

Subtheme 4: Fluency Should Be Considered in a Broader Context

I always come back to the individual's lived experience with their language impairment and how it impacts them during the conversations that matter most to them. That information and assessments of aspects of their language system (how they process verbal or graphemes input, retrieve sounds and words, structure output at word level/grammatical structures) are what help me structure therapeutic intervention, education, and introduction of compensatory strategies).

The items assessed and the importance of each of them is dependent upon the client and the most outstanding difficulties that they exhibit. The areas that are the most debilitating for the client are the areas of most importance during the assessment.

Fluency in aphasia is a very broad term and, in clinical settings/everyday life, should also include nonverbal communication and being able to cowork with the listener (fluency in communicating a message).

I also look at secondary characteristics (visual).

In my opinion, fluency in aphasia should be viewed … in terms of overall communications skills keeping in mind the needs, demands, and wants of persons with aphasia. Furthermore, it should look at how the PwA's activity and participation are influenced and what impact that has on the identity of persons with aphasia. How the verbal components connect PwA in society should be one of the important indicators.

Appendix (p. 2 of 2)

Open-Ended Responses Commenting on the General Topic of Fluency Measurement

Subtheme 5: Limitations and Solutions
During my practice with aphasia patients, I often used informal checklists and assessments to gather information. Often, the patient was not available long enough to get a sufficient assessment. Many times the family and/or doctors wanted immediate feedback. So time was critical especially in the hospital setting. In a home health or nursing setting, a more reliable assessment would be great.
I do note speech rate and take that into account as an important aspect for fluency assessment, but I generally do not calculate it formally in the clinic (e.g., #words/minute). . . . I also note the presence of apraxia of speech which of course impacts articulatory effort and slows speech/makes it more effortful. However, I have worked with individuals whose language profiles match an anomic classification (a fluent type of aphasia) with concomitant AOS that results in their speech/language output looking more non fluent.
Fluency measures are limited for stuttering severity as well.
I think fluency is challenging and the WAB-R judgment is not a good reflection of fluency, yet so many use that test.