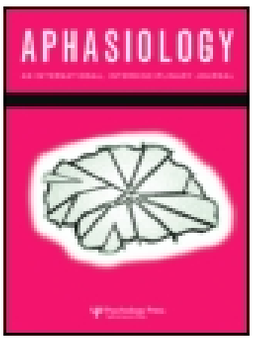

Aphasiology

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

Fluent or nonfluent? Part A. Underlying contributors to categorical classifications of fluency in aphasia

Sharice Clough & Jean K. Gordon

To cite this article: Sharice Clough & Jean K. Gordon (2020): Fluent or nonfluent? Part A. Underlying contributors to categorical classifications of fluency in aphasia, Aphasiology, DOI: 10.1080/02687038.2020.1727709

To link to this article: https://doi.org/10.1080/02687038.2020.1727709

APHASIOLOGY
https://doi.org/10.1080/02687038.2020.1727709

Routledge
Taylor & Francis Group

Fluent or nonfluent? Part A. Underlying contributors to categorical classifications of fluency in aphasia

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

[a]Department of Hearing and Speech Sciences, Vanderbilt University, Nashville, Tennessee, USA;
[b]Department of Communication Sciences & Disorders, University of Iowa, Iowa City, Iowa, USA

ABSTRACT

Background: The concept of fluency is widely used to dichotomously classify aphasia syndromes in both research and clinical practice. Despite its ubiquity, reliability of fluency measurement is reduced due to its multi-dimensional nature and the variety of methods used to measure it.

Aims: The primary aim of the study was to determine what factors contribute to judgements of fluency in aphasia, identifying methodological and linguistic sources of disagreement.

Methods & Procedures: We compared fluency classifications generated according to fluency scores on the revised *Western Aphasia Battery* (WAB-R) to clinical impressions of fluency for 254 English-speaking people with aphasia (PwA) from the AphasiaBank database. To determine what contributed to fluency classifications, we examined syndrome diagnoses and measured the predictive strength of 18 spontaneous speech variables extracted from retellings of the Cinderella story. The variables were selected to represent three dimensions predicted to underlie fluency: grammatical competence, lexical retrieval, and the facility of speech production.

Outcomes & Results: WAB-R fluency classifications agreed with 83% of clinician classifications, although agreement was much greater for fluent than nonfluent classifications. The majority of mismatches were diagnosed with anomic or conduction aphasia by the WAB-R but Broca's aphasia by clinicians. Modifying the WAB-R scale improved the extent to which WAB-R fluency categories matched clinical impressions. Fluency classifications were predicted by a combination of variables, including aspects of grammaticality, lexical retrieval and speech production. However, fluency classification by WAB-R was largely predicted by severity, whereas the presence or absence of apraxia of speech was the largest predictor of fluency classifications by clinicians.

Conclusions: Fluency judgements according to WAB-R scoring and those according to clinical impression showed some common influences, but also some differences that contributed to mismatches in fluency categorization. We propose that, rather than using dichotomous fluency categories, which can mask sources of disagreement, fluency should be explicitly identified relative to the underlying deficits (word-finding, grammatical formulation, speech production, or a combination) contributing to each individual

ARTICLE HISTORY
Received 24 June 2019
Accepted 29 January 2020

KEYWORDS
Aphasia; assessment; diagnosis; fluency; reliability

CONTACT Jean K. Gordon jean-k-gordon@uiowa.edu Department of Communication Sciences & Disorders, University of Iowa, Iowa City, USA

PwA's fluency profile. Identifying what contributes to fluency disruptions is likely to generate more reliable diagnoses and provide more concrete guidance regarding therapy, avenues we are pursuing in ongoing research.

Introduction

One of the most common ways of describing aphasia is by the fluency of language output. Defined broadly, fluency refers to the ease and speed with which a task can be completed. In language production, fluency arises from the ability to smoothly coordinate linguistic subtasks – the timely retrieval of words to be integrated into an emerging syntactic framework, and the seamless programming of the formulated message for articulation. Nonfluent aphasia is characterised by increased effort, impaired prosody, articulatory errors, reduced grammaticality, and a predominance of content words, and is associated with anterior lesion sites (Feyereisen, Pillon, & De Partz, 1991; H. Goodglass & Kaplan, 1983; Kertesz, 2006). In contrast, fluent aphasia, associated with more posterior lesions, is characterised by uninterrupted runs of speech with a variety of syntactic structures, in which rate or phrase length is normal or even "hyper-normal", but output is typically erroneous (Edwards, 2005; Feyereisen et al., 1991; Goodglass & Kaplan, 1983; Kertesz, 2006). In classical taxonomic approaches, non-fluent aphasia syndromes include global, Broca's, and transcortical motor aphasia, whereas fluent aphasias include Wernicke's, transcortical sensory, conduction, and anomic aphasia.

This traditional taxonomic approach to aphasia has often been criticized over the years (e.g., Poeck, 1989; Tremblay & Dick, 2016) but continues to dominate the field of aphasiology. Similarly, treating fluency as a dichotomy has generated significant criticism, and for good reason. Within each broad fluency category, patterns of spontaneous speech are highly variable, giving rise to questions about the validity of the dichotomy. Poeck (1989) noted that "for classification purposes, the fluency – nonfluency dimension is too broad to be useful" (p. 30). Nevertheless, the traditional taxonomy, including the fluency dichotomy, underlies three of the most commonly used aphasia tests – the *Western Aphasia Battery-Revised* (WAB-R, Kertesz, 2006), the *Boston Diagnostic Aphasia Exam* (BDAE, Goodglass, Kaplan, & Barresi, 2001), and the *Aphasia Diagnostic Profiles* (ADP, Helm-Estabrooks, 1992). Despite the variability within fluency categories, it has been argued (e.g., Goodglass, Quadfasel, & Timberlake, 1964) that fluency in aphasia is categorical, not continuous, because fluent and nonfluent aphasia are such qualitatively different syndromes, and arise from different lesion site areas. Even when syndromes evolve during recovery, this evolution does not typically cross the "fluency barrier" (Code & Rowley, 1981). In addition to its use in clinical contexts, the fluency dichotomy has been widely used in research to study broad differences in aphasia, both historically (e.g., Albert & Sandson, 1986; Marshall & Tompkins, 1982; Whitehouse, Caramazza, & Zurif, 1978) and more recently (e.g., Alyahya, Halai, Conroy, & Lambon Ralph, 2018; Hazamy & Obermeyer, 2019; Kong & Wong, 2018; Lee, Kocherginsky, & Cherney, 2018). Like many dichotomies (indeed, many taxonomies), the fluent-nonfluent distinction entails significant simplification, but has been found to be

a "useful shorthand" to describe broad differences in oral expressive behavior in aphasia (Edwards, 2005).

The fluency reliability problem

Despite its widespread use by both clinicians and aphasia researchers, and despite flurries of research into the topic over the years, the measurement of fluency continues to be unreliable. This is likely due to several factors. First, the concept of fluency is applied in two ways in aphasia diagnosis – as a dichotomous classification of aphasia syndromes (as described above), and as a descriptor of continuously variable speech output, and these two uses are not always concordant with each other. Second, because fluency is multiply determined by several underlying factors, there is still no agreed-upon measure for evaluating it, and different clinicians and researchers use different methods and prioritise different criteria. As a first step toward improving the reliability of fluency judgements, the goal of the current study is to identify the most salient features underlying dichotomous classifications of aphasia fluency. To do so, we make use of AphasiaBank (MacWhinney, Fromm, Forbes, & Holland, 2011), a large database of individuals with aphasia (https://aphasia.talkbank.org), and we compare two sources of classification available in AphasiaBank – classifications according to the revised WAB-R (Kertesz, 2006) and classifications according to clinical impressions. In a companion paper (Gordon & Clough, 2020), we extend our analysis by examining predictors of continuous measures of fluency.

Underlying components of fluency

Fluency may be disrupted by one or a combination of deficits to underlying processes supporting speech/language production. Several studies have contributed to the attempt to identify these underlying processes. Casilio and colleagues conducted a factor analysis of 27 spontaneous speech variables, each rated on a 5-point scale (Casilio, Rising, Beeson, Bunton, & Wilson, 2019). Four main factors accounted for 79.5% of the variance in speech profiles of the PwA. These included paraphasia, logopenia (or paucity of speech), agrammatism, and motor speech. Vermeulen and colleagues identified 5 factors among 17 spontaneous speech measures: syntactic ability, phonological paraphasia, neologistic paraphasia, articulatory impairment, and vocabulary (Vermeulen, Bastiaanse, & Van Wageningen, 1989). In each study, the identified factors capture the three primary dimensions of spontaneous speech production that also underlie fluency: *grammatical competence, lexical retrieval*, and *motor speech production*.

It is well-known that nonfluent aphasia is associated with grammatical deficits, because agrammatism is a common feature of Broca's aphasia, the prototypical nonfluent type of aphasia (e.g., Goodglass et al., 2001; Kertesz, 2006). Agrammatic aphasia is characterised by impoverished sentence production, including difficulty producing verbs, simplified verb structures, and omission of function words and morphemes (Berndt & Caramazza, 1980; Rochon, Saffran, Berndt, & Schwartz, 2000; Thompson & Bastiaanse, 2012). People with agrammatic aphasia have difficulty producing complex sentence structures, often producing simple canonical-order sentences (Roelien Bastiaanse, Rispens, Ruigendijk, Rabadan, & Thompson, 2002; Thompson, Lange, Schneider, & Shapiro, 1997). Quantitative analyses have identified that people with nonfluent aphasia demonstrate

lower verb:noun ratios (Gordon, 2006), and that those who demonstrate agrammatism produce fewer grammatical morphemes, obligatory determiners, and verb inflections (Rochon et al., 2000; Saffran, Berndt, & Schwartz, 1989). However, not all people with nonfluent aphasia are agrammatic. Agrammatism is generally associated with Broca's aphasia – those with global aphasia usually have output that is too sparse to accurately judge grammatical competence, and the nonfluency of those with transcortical motor aphasia (TCM) is typically due to difficulties initiating speech (Goodglass et al., 2001). Even among those diagnosed with Broca's aphasia, agrammatism is not always a prominent feature. For example, Rochon and colleagues (2000) reported that, of the 29 individuals with Broca's aphasia in their study, 20 (69%) were identified as having agrammatism.

Lexical retrieval difficulties are common to both nonfluent and fluent aphasia types, although they may show qualitative differences. For example, people with agrammatism have more difficulty retrieving function than content words (Saffran et al., 1989) and verbs than nouns (Gordon, 2006; Zingeser & Berndt, 1990), while those with anomia show the opposite pattern (Zingeser & Berndt, 1990). In fluent aphasias, word-finding difficulty often manifests as empty, circumlocutory, or paraphasic output that reduces the specificity and efficiency of language production but not necessarily its fluency. Lexical diversity is often reduced for these individuals, especially for verbs (Bastiaanse, Edwards, & Kiss, 1996; Edwards & Bastiaanse, 1998). In contrast, people with nonfluent aphasia often demonstrate greater lexical diversity than people with fluent aphasia, relative to the amount of content produced (Gordon, 2008). In both fluent and nonfluent aphasias, word retrieval difficulty may result in slowed rate and frequent pauses and repairs that disrupt the flow of output (Andreetta & Marini, 2015; Edwards, 2005).

Fluency is also affected by more peripheral aspects of expression, especially the presence of motor speech disorders. Articulatory and prosodic disturbances are prominent features of dysarthria and apraxia of speech (AoS), which frequently co-occur with nonfluent aphasia (Duffy, 2013; McNeil, Robin, & Schmidt, 2009). Deficits in motor speech may give rise to dysfluent signs such as slowed speech rate, phonological errors, revisions, and perceptually effortful speech. But such errors are not restricted to the nonfluent aphasias – successive phonemic approximations common to conduction aphasia may result in frequent restarts that can be difficult to differentiate from the sound distortions and substitutions accompanying AoS (Haley, Jacks, & Cunningham, 2013; Laganaro, 2012).

Fluency assessment

Fluency can be characterised by objective measurement or subjective rating scales. The ADP (Helm-Estabrooks, 1992) takes the former approach, calculating fluency on the basis of the longest average phrase length across three discourse tasks. In an early study, Goodglass and colleagues also suggested a measure based on phrase length: they calculated a ratio of the number of long (5+ words) to short (1–2 word) utterances, which generated a bimodal distribution in their sample of 49 PwA (Goodglass et al., 1964). Rate of speech is commonly used, usually measured in words per minute (e.g., Howes, 1964), sometimes in syllables per minute (e.g., Park et al., 2011). One criticism of such measures is that reducing fluency to a single dimension may not capture relevant aspects of fluency disruption.

In clinical assessment, the most common approach to solving the fluency problem is to rate multiple dimensions of fluency to achieve some degree of convergence from the evidence. The BDAE (Goodglass et al., 2001) provides rating scales for 6 potential dimensions of fluency: melodic line, phrase length, articulation, grammaticality, paraphasia, and word-finding ability. These ratings, in combination with repetition and auditory comprehension scores from the battery, provide profiles which are used by the clinician to judge the best-fitting syndrome for each patient. One drawback of this approach is that each rating scale requires a subjective judgement, leaving plenty of room for disagreement.

The WAB-R sidesteps the problem of multiple rating scales by condensing judgements of spontaneous speech to just two rating scales, one reflecting informational content (not examined in the current study) and the other fluency, grammatical competence, and paraphasias. The latter scale is commonly referred to as the "fluency scale", including by Kertesz himself (2006), although it encompasses much more than fluency. On this scale, the clinician matches the spontaneous speech of a PwA to one of 11 qualitative descriptions, numbered 0 ("no words or short, meaningless utterances") to 10 ("sentences of normal length and complexity without definite slowing, halting, or paraphasias"). Ratings from 0 to 4 are considered nonfluent and from 5 to 10 are considered fluent. Like the BDAE, these descriptions include dimensions of grammaticality, word-finding, prosody and articulatory effort. Unlike the BDAE, multiple dimensions are compiled into each level, which requires the administrator to select the one description that best fits the PwA's language production.

Although the reduced number of dimensions and the *a priori* designation of levels on the scale as either "fluent" or "nonfluent" were intended to improve reliability, the requirement to force patients into predefined profiles means that the fit may be less than optimal. Trupe (1984) examined the inter-rater reliability of 5 clinicians' ratings on the WAB fluency scale for 20 transcribed samples. Agreement was described as "poor". Although no index of agreement was calculated, the raw data in the article indicate that fluency ratings for 11 of the 20 samples showed "acceptable" discrepancies of 0 or 1 on the fluency scale, while the other 9 showed discrepancies as large as 2 to 4 points on the scale. Clarifying the scoring criteria reduced these discrepancies; however, this improved reliability was not maintained with new judges.

Ratings on the WAB-R fluency scale are combined with auditory comprehension, repetition, and word retrieval scores from the battery to identify the best-fitting syndrome for each PwA. The two spontaneous speech scales together constitute 20% of the WAB-R AQ (Kertesz, 2006). Thus, small differences in ratings can greatly impact overall severity, as well as alter aphasia syndrome classifications. Disagreements on fluency may contribute to discrepancies in aphasia type classification. Comparing BDAE and WAB outcomes, Wertz, Deal, and Robinson (1984) found that the two tests generated the same syndrome type for only 27% (n = 12) of a sample of 45 PwA. Agreement was highest for global aphasia, and lowest for transcortical motor, conduction, and anomic aphasia. Comparing WAB syndromes to their own clinical impressions, Swindell, Holland, and Fromm (1984) found a rate of 54% agreement for 69 PwA, with the lowest agreement for Broca's and anomic aphasia. A more recent analysis of 82 acute ischemic stroke patients found that 71% of participants classified with anomic aphasia by WAB-R were classified with Broca's aphasia by clinical impression (John et al., 2017). It is worth noting that the primary difference between anomic and Broca's aphasia subtypes, as characterized by the WAB-R, is in the fluency scale;

the two syndromes overlap considerably in the other dimensions of comprehension, repetition, and naming. Thus, a one-point difference on the fluency scale (from 4 to 5) would be enough to change an aphasia subtype from Broca's aphasia to anomic aphasia. Such discrepancies may result from differences in the perception of effort or concomitant motor speech deficits. For example, apraxia of speech is confusable with phonological paraphasias, particularly when a given speaker produces both (Laganaro, 2012).

Despite claims of poor reliability, the WAB-R scale remains one of the most popular standardised methods of measuring fluency. The attraction is in its relatively "automatic" syndrome classification, including the division into fluent and non-fluent syndromes. A survey of 108 clinicians and researchers from the Unites States, Canada, United Kingdom, New Zealand, Australia, and Chile examined the types of assessments commonly used across clinical and research settings (Kiran et al., 2018). While clinicians used a large variety of tests, the WAB-R was the most frequently used across settings. Of clinicians working in inpatient, outpatient, and home-care settings, almost 70% reported using the WAB-R for clinical purposes. In intensive aphasia programs, between 80-90% of respondents reported using the WAB-R for clinical purposes. In aphasia research settings, the WAB-R dominated, with 100% of respondents (n = 11) reporting that they used the WAB-R for research purposes. In other settings as well, the WAB-R was one of the most frequently used tests for research purposes.

Perceptions of fluency

Perceptual judgements form the basis of subjective rating scales such as those in the WAB-R and the BDAE. Similarly, the clinical gold standard for differential diagnosis of motor speech disorders is auditory-perceptual judgement (Duffy, 2013; Kent, 2009). Perceptual scales have been used to quantify the presence and severity of neurogenic apraxia of speech (e.g., Strand, Duffy, Clark, & Josephs, 2014) and dysarthria (e.g., Bunton, Kent, Duffy, Rosenbek, & Kent, 2007). Auditory-perceptual rating also serves as the standard for characterizing and diagnosing voice disorders (e.g., Kempster, Gerratt, Abbott, Barkmeier-Kraemer, & Hillman, 2009). Goodglass and colleagues (1964) endorsed this approach specifically for differentiating fluent and nonfluent types of aphasia. More recently, Casilio and colleagues (2019) developed a tool for rating aphasic expression, and reported that most participants' ratings reliably reflected objective measures from AphasiaBank, providing support for the use of subjective ratings.

Given the lack of consensus on objective alternatives, a clinician's ear continues to be one of the most widely used diagnostic tools for diagnosing certain aspects of communication disorders. Nevertheless, perceptual judgements of fluency are often unreliable when compared across listeners. Gordon (1998) found that, despite the fact that a sample of 24 clinicians reported frequently using fluency to characterise aphasia, there was little agreement among them on the fluency diagnoses of 10 PwA. Only half of the audio-recorded speech samples were rated as fluent or nonfluent by a consensus of at least two-thirds of the clinicians. Holland and colleagues (1986) found little agreement even among experts. Twenty-two experienced speech-language pathologists and neuropsychologists judged the applicability of 21 possible descriptors to characterise the spontaneous speech of one PwA. All agreed that that the sample was characterised by dysprosody, difficulty initiating speech, and slow and effortful speech, although ratings of severity varied. The

majority of raters also identified the presence of apraxia and mild word retrieval difficulties; however, identification of the presence of other characteristics such as agrammatism and dysarthria was equivocal – only about half of the raters detected them.

Differences in perceptual judgement can be traced to the variety of criteria that might be used. Studies focusing on listener perceptions of fluency show that clinicians differ in the relative importance they place on different spontaneous speech characteristics when classifying fluency. Gordon (1998) found that almost half of clinicians (42%) identified grammaticality as the most salient characteristic of fluency, while slightly fewer identified articulatory effort (37%) and still fewer word-finding difficulties (21%). Park and colleagues (Park et al., 2011) asked expert clinicians to classify fluency in a large and varied group of speakers, including non-brain-damaged adults, individuals with dementia, and PwA. The feature that carried the most weight in judgements of fluency was "speech productivity", that is, the amount of time spent speaking compared to pause time. Speech rate (syllables per minute) and perceived audible struggle also significantly predicted fluency classifications, but lexical specificity (measured by type-token ratio) and the use of fillers did not.

The current study

Clinicians vary in their judgements of fluency in aphasia, and this is due, in large part, to differences in the methods by which fluency is measured and the criteria used to distinguish fluent from nonfluent categories. Thus, the goal of the current study is to build on past research (notably Gordon, 1998; Holland et al., 1986; Park et al., 2011; Trupe, 1984) to investigate how methods of fluency classification affect the outcomes, and what speech and language characteristics contribute to perceptions of fluency. To do so, we take advantage of the large and well-described database in AphasiaBank. Specifically, in that database, fluency categories were determined using two different methods: scores on the WAB-R spontaneous speech fluency scale, and "clinical impression". Our aim was to identify, not only contributors to fluency judgements, but contributors to *disagreements* in fluency, whether methodological or related to characteristics of the aphasia. The significant contributions of the study lie in the large dataset and the analysis of a wide and theoretically motivated range of speech and language variables.

Methods

This study is part of a larger project funded by the American Speech-Language-Hearing Foundation. It was approved by the Institutional Review Board (IRB) of the University of Iowa.

Data collection

Potential participants in the study included all unique English-speaking individuals in the AphasiaBank database as of May 2018 (n = 305). AphasiaBank consists of demographic information and standardized assessment results for PwA collected from multiple sites across the country according to a standardized and well documented protocol (see MacWhinney et al., 2011 for details). The AphasiaBank protocol includes administration of the WAB-R (Kertesz, 2006), as well as additional tests of word retrieval, grammatical

competence, and discourse production. For the current study, we required transcripts of the Cinderella story retelling task; WAB-R results, including the Aphasia Quotient (reflecting overall severity) and Spontaneous Speech fluency scale scores; and clinical impressions regarding fluency (fluent *vs* nonfluent). Thus, the sample of 305 PwA was narrowed down to include only participants with completed transcriptions of the Cinderella Story, WAB-R scores, and clinician fluency classifications, resulting in a final sample of 254 participants.

Clinical impressions

In AphasiaBank, PwA are typically assigned a classification of "fluent" or "nonfluent" by the clinicians who assess them. In addition, these classifications can be generated according to scores on the WAB-R spontaneous speech fluency scale (i.e., 0–4 = nonfluent; 5–10 = fluent). Aphasia syndromes were also designated by both clinical impression and WAB-R scores. In the majority of cases (n = 169, 67% of the current sample), the AphasiaBank investigators (M. Forbes and A. Holland, two expert clinical aphasiologists) completed the AphasiaBank protocol, including administration of the WAB-R and the Cinderella story and completion of the demographic form including their clinical impressions of fluency and aphasia subtype. In these cases, clinical impressions were typically arrived at by consensus between them, sometimes with input from the contributing clinician/researcher (Fromm, D. & Forbes, M. Personal communication, Sept., 2019). In the other third of cases, contributing clinicians or clinical researchers administered the WAB-R, completed the AphasiaBank protocol, and contributed their clinical impressions. Instructions for forming clinical impressions specified that they need not be based on any standardized method or prescribed sources of information, and contributors were not asked to justify their impressions.

Thus, the two sets of diagnoses included in AphasiaBank reflect two different clinical methods, even if contributed by the same examiners: 1) diagnoses guided by the standardized testing and interpretation of the WAB-R and 2) clinical impressions informed by a varying combination of sources of information. Importantly, both methods are commonly used and representative of clinical practice. For most of the PwA in our sample, these two methods were carried out by the AphasiaBank investigators; for the other PwA, the examiners varied. Thus, we included post-hoc tests (described below) to assess whether these examiner differences affected the results.

Spontaneous speech samples

Samples of spontaneous speech from the task of retelling the Cinderella story were analyzed to identify characteristics of spontaneous speech. Of the various tasks involving spontaneous speech in AphasiaBank (e.g., stroke story interviews, picture descriptions), the story retelling task was determined to strike the best balance for our purposes between standardizing the topic and requiring generation of lexical content and syntactic structure. According to the AphasiaBank protocol, examiners elicited the Cinderella story using a wordless picture book, sometimes prompting the PwA with generic prompts such as "What happened next?" or, if necessary, with trouble-shooting questions (e.g., "Did Cinderella go to the ball?"). Computerized Language Analysis (CLAN, MacWhinney, 2000)

was used to extract the Cinderella story from the transcripts, and to calculate the majority of the connected speech measures.

Connected speech measures

To identify contributors to fluency classifications, we examined measures of the two major linguistic domains of spoken language production (grammatical competence and lexical retrieval), as well as measures of speech production, all of which might contribute to fluency. Variables were selected to focus as clearly as possible on one of these three dimensions, although a few of them might reflect multiple underlying components (see Table 1 and

Table 1. Predictor variables reflecting underlying components of fluency.

Underlying component	Dimension	Predictor variable	CLAN code/analysis [#]
Grammatical competence	Grammatical accuracy	Proportion of ungrammatical utterances **(Gram Errs)**	[+gram] coded at the utterance level
	Grammatical complexity	Proportion of complex grammatical relations (**Gram Comp**)	Compilation of 10 grammatical relations (coded on the GRA tier) that mark embeddings
	Morphological accuracy	Proportion of morphological errors (**Morph Errs**)	compilation of 21 possible morphological errors coded [*m]
	Morphological complexity	Proportion of verbs inflected (**Inflect Vbs**)	Verb inflections (coded on the MOR tier) extracted using the EVAL command
Lexical retrieval ability	Lexical accuracy	Proportion of semantic errors (**Sem Errs**)	Compilation of 4 semantic error codes [*s]: related and unrelated semantic errors, perseverated semantic errors and unknown semantic errors
	Lexical specificity	Proportion of empty utterances (**Empty**)	[+es] coded at the utterance level
	Lexical efficiency	Proportion of circumlocutory utterances (**Circum**)	[+cir] coded at the utterance level
		Propositional density (**Prop Dens)**	Counts noun relations (e.g. verbs, adjectives, prepositions) but not nouns; available through EVAL
	Lexical diversity	Moving Average Type-Token Ratio (**MATTR**)	Calculated with a moving window of 10 words; obtained through a FREQ command
Facility of speech production	Phonological encoding	Proportion of phonological errors (**Phon Errs**)	Compilation of 4 phonological error codes [*p]
		Proportion of neologistic errors (**Neo Errs**)	Compilation of 3 neologistic error codes [*n]
	Motor speech	Apraxia of Speech: Y/N (**AoS**)	Coded on the demographics sheet
		Dysarthria: Y/N (**Dys**)	Coded on the demographics sheet
	Melodic line	Pitch variability: SD of F_0 (**Pitch Var**)	Standard deviation of F_0 of a 60-second sample; calculated using Praat
Combined measures	Grammatical & lexical	Content:function word ratio (**Con:Fun**)	Ratio of content to function words; available through EVAL
	Grammatical, lexical & speech	Proportion pause time (**Pauses**)	% of total sample duration consisting of pauses longer than 150 msec; calculated using Goldwave
		Proportion of utterances retraced (**Retrace**)	[//] in CLAN indicates revision (usually syntactic) of an utterance; available through EVAL
Global measure		Total Utterances (**Total Utts**)	Excludes nonword and nonverbal turns; available through EVAL command

[#]All measures were extracted from CLAN, except where noted.

further explanation below). Our final set included the Aphasia Quotient (AQ) from the WAB-R, as an index of severity, and 18 linguistic variables (Table 1), which were selected to represent the three dimensions but without a high degree of multicollinearity among the variables. All inter-correlations among selected predictor variables were < |.500| (see Appendix 1).

Grammatical competence

To capture grammaticality, we used sentence-level and morphological measures of accuracy and complexity, based on previous findings that impairments in these domains are reflective of nonfluent output (e.g., Rochon et al., 2000; Saffran et al., 1989; Thompson et al., 1997). *Grammatical accuracy* was calculated as the proportion of total utterances classified as ungrammatical. *Morphological accuracy* was calculated as the number of morphological errors divided by total words. *Grammatical complexity* was measured as the proportion of grammatical relations coded in CLAN that were classified as complex, i.e. containing syntactic embeddings. This procedure was found to have an accuracy of 95%, comparable to that of human coders (MacWhinney, 2000). *Morphological complexity* was calculated as a proportion of verbs that were inflected (we focused on verbs because of their vulnerability in agrammatism, e.g., Bastiaanse et al., 1996).

Lexical retrieval

Measures of lexical retrieval were selected based on previous findings that fluent and nonfluent groups differ on the basis of semantic sufficiency (Gordon, 2008) and on clinical observation that word retrieval deficits frequently give rise to dysfluency. Proportions of semantic (real-word) and neologistic (non-word) errors, calculated as a proportion of total words produced, were used as indices of *lexical accuracy*. In AphasiaBank coding, the category of "semantic" errors includes real-word errors, whether semantically related or unrelated, that do not meet criteria for phonological relatedness, while neologistic errors are non-words with a known or unknown target. Details on error coding are in the CHAT manual available on the AphasiaBank website (MacWhinney, 2000).

The proportions of utterances produced that were characterised in CLAN as empty or circumlocutory reflected *lexical specificity* and *lexical efficiency*, respectively. An additional measure of lexical efficiency was propositional (or idea) density, which is calculated by dividing the number of words that form propositions (verbs, adjectives, adverbs, and prepositions) by the total number of words (after Brown, Snodgrass, Kemper, Herman, & Covington, 2008; Snowden et al., 1996). A common measure of *lexical diversity* is type-token ratio (TTR), a ratio of the number of different words relative to the number of total words produced. We used an automated method that calculates a moving-window average TTR (MATTR, Covington & McFall, 2008), which reduces the impact of sample size. MATTR has been incorporated into the CLAN tools.

Facility of speech production

To quantify more peripheral aspects of speech-language production, we calculated proportions of phonological and neologistic errors to reflect difficulties in *phonological encoding*. In AphasiaBank, phonological errors may be real words or nonwords, but must meet a minimum overlap of phonological structure with the target (see CHAT manual for specific criteria). Presence or absence of AoS and dysarthria, as judged by the clinician, were included as categorical variables reflecting *motor speech*. In order to quantify

melodic line across the speech sample, we used the standard deviation of the speaking fundamental frequency as a measure of pitch variability (following Baken & Orlikoff, 2000). To calculate pitch variability, we first used GoldWave Digital Audio Editing Software (GoldWave, 2017) to remove any examiner speech, including segments with voicing overlap between examiner and participant, so that only the PwA's voiced signal remained in the sample. Using Praat (Boersma & Weenink, 2017), we selected 60-second samples of each audio file from the middle of the sample. Voicing was detected within the standard range of 75–500 Hz. The analysis window was then narrowed before calculating the standard deviation, so that the lower margin was just below the speaker's minimum F_0 and the upper margin just above the speaker's maximum F_0 to reduce the influence of any background noise.

Combined measures

Some of the measures potentially represent disruption at multiple levels. Content-to-function word ratio, retracing, and pausing can result from both grammatical and lexical deficits; retracing and pausing might also arise due to speech production difficulties. To calculate the proportion of pauses, we used GoldWave to eliminate background noise, then removed all pauses longer than 150 milliseconds, a conservative measure for differentiating silent pauses from articulatorily conditioned pauses (Pakhomov & Kotlyar, 2011). Proportion pause time was calculated by the sum of all identified pause durations divided by the unedited sample length in seconds. Because the steps involved depended on identifying background noise, introducing subjective judgement, reliability was calculated from two independent editors. Editor 2 was first trained on 15 practice files by Editor 1. Inter-rater reliability was then calculated on 50 samples (20%). Correlation between editors was 0.989 for total pause time.

Finally, we included the number of total utterances as a covariate to account for sample length, since this can affect the influence of some of the other linguistic variables.

Analysis

Before addressing our primary goal of determining speech-language contributors to fluency diagnoses, we examined the agreement between fluency classifications (fluent *vs* nonfluent) arising from the WAB-R and those from clinical impressions. To do this, we calculated the proportions of matching and mismatching fluency categories. Mismatches were further analyzed by aphasia syndrome diagnoses, to illustrate sources of confusion underlying fluency classification. To identify specific speech-language characteristics contributing to fluency classification, we used logistic regression, examining the contributions made by aphasia severity (WAB-R AQ) and the 18 linguistic variables to fluency categories as assigned either by WAB-R scores or by clinical impressions. Continuous spontaneous speech measures were first converted to *z*-scores to put them on the same scale. For each regression model, we used a backwards elimination process: we began by entering all 19 predictor variables, then removed them one by one, beginning with the highest probability values (least likelihood of being significant). To ensure that the results were robust, the final models retained only predictors with *p*-values less than .01.

Results

Our final sample of PwA consisted of 254 PwA, of whom 45% were female, 85% were white and 11% were African-American. The average age of the group was 61 years, with a range of 26 to 91 years (n = 1 unknown). Education level ranged from 8 to 25 years, with an average of 15 years (n = 8 unknown). The primary language spoken for 99% of the PwA was English (n = 3 unknown), and the vast majority (88%) were monolingual (n = 3 unknown). Of those with known handedness (n = 4 unknown), 92% were right-handed. The etiology of aphasia was stroke for most (97%) of the sample (n = 1 unknown). Of these, 41% were ischemic, 15% hemorrhagic, and the remainder (n = 94) were unknown. Of those with documented lesion laterality (n = 28 unknown), 96% were unilateral left-hemisphere lesions and 3% were unilateral right-hemisphere lesions. The average time post-onset was 5.3 years, with a range of .1 to 27 years. The average time spent in speech-language therapy was 3.2 years, ranging from none to 16 years. The data were contributed from a total of 21 different sites. The examiners' mean (and median) years of experience was 15 years (range: 0.5 to 40 years).

Figure 1(a) shows the distribution of aphasia subtypes according to WAB-R syndrome classifications. The group consisted of 34% anomic aphasia, 26% Broca's aphasia, 18% conduction aphasia, 7% Wernicke's aphasia, 4% transcortical aphasias, and 11% with AQs above 93.8, designated by WAB-R criteria as having no residual aphasia (but see Fromm et al., 2017). This distribution is similar to past epidemiological studies of chronic aphasia, in that Broca's aphasia and anomic aphasia were the most frequent types (e.g., Obler, Albert, Goodglass, & Benson, 1978; Pedersen, Vinter, & Olsen, 2004). A caveat is the paucity of individuals with global aphasia in this sample, which occurs because they are less likely to be able to complete the AphasiaBank protocol. Nevertheless, AQs ranged widely, from 10.8 to 99.6 (mean = 72.3), and WAB-R fluency scale scores ranged from 0 to 10 (mean = 6.3).

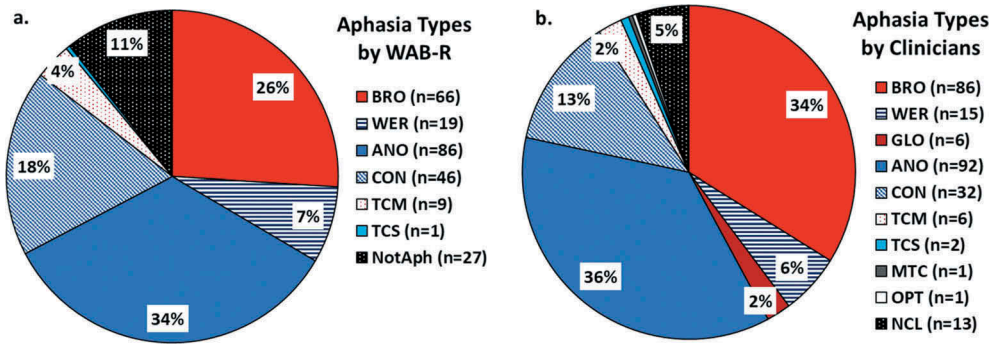

Figure 1. Proportions of aphasia syndromes by (a) WAB-R scores, and (b) clinical impression. Abbreviations: BRO, Broca's aphasia; WER, Wernicke's aphasia; GLO, Global aphasia; ANO, Anomic aphasia; CON, Conduction aphasia; TCM, Transcortical motor aphasia; TCS, Transcortical sensory aphasia; MTC, Mixed transcortical aphasia; OPT, Optic aphasia; NCL, not classified; NotAph, not aphasic.

Agreement of WAB-R and clinician classifications

Figure 1 shows the breakdown of aphasia types by WAB-R (1a) and by clinical impression (1b). Clinical impression resulted in more diagnoses of Broca's aphasia and fewer of conduction aphasia compared to WAB-R diagnoses, and included a few diagnoses not represented in the WAB-R categories in this sample (global, mixed transcortical/isolation, and optic aphasia). A few of the PwA were not assigned an aphasia type by clinical impression (n = 13, 5.1%), which might indicate that their aphasia was considered to be resolved, or to be unclassifiable. Unlike clinical impressions, the WAB-R criteria explicitly designated several participants as not aphasic, defined as scoring equal to or greater than 93.8 on the WAB-R AQ (Kertesz, 2006, p. 91).

Of the 254 PwA, clinical impressions classified 139 as fluent and 115 as nonfluent, while the WAB-R fluency scale yielded a classification of 180 fluent and 74 nonfluent PwA. There were 43 PwA whose WAB-R fluency classifications did not match the clinical impressions, generating an overall agreement of 83%. Agreement was excellent (99%) for those that were classified by clinical impression as having fluent aphasia, while agreement was much poorer (64%) for those clinically classified as having nonfluent aphasia. (Notably, the likelihood of agreement was not related to years of clinical experience of the examiner [p = .941] or the relationship of the examiner to the PwA [researcher *vs* clinician, p = .709]). Proportions of fluency mismatches are shown by aphasia syndrome in Table 2. Of the 43 *mismatched* PwA, all but one (n = 42) were classified as fluent by the WAB-R but nonfluent by clinicians. Specifically, the WAB-R categorized 51% of mismatches with anomic aphasia (n = 22), 30% with conduction aphasia (n = 13) and 12% with Wernicke's aphasia (n = 5), but only 5% with Broca's aphasia (n = 2). One was categorized as not aphasic (2%). Clinical impressions designated most of these as having Broca's aphasia (n = 30, 70%), and a few

Table 2. Confusion matrices of all fluency mismatches, comparing clinical syndrome diagnoses to WAB-R-generated syndrome diagnoses.

	Clinician Diagnoses							
WAB-R Diagnoses	GLO	BRO	TCM	WER	ANO	CON	Not Class	Total by WAB-R
a) Original WAB-R fluency scale, in which 0–4 = non-fluent, 5–10 = fluent.								
GLO	0	0	0	0	0	0	0	**0**
BRO	0	1	0	0	1	0	0	**2**
TCM	0	0	0	0	0	0	0	**0**
WER	0	5	0	0	0	0	0	**5**
ANO	0	12	3	0	4	0	3	**22**
CON	0	12	0	0	0	1	0	**13**
Not Aph	0	0	0	0	1	0	0	**1**
Total by Clinicians	**0**	**30**	**3**	**0**	**6**	**1**	**3**	**43**
b) Adjusted WAB-R fluency scale, in which 0–5 = non-fluent, 6–10 = fluent								
GLO	0	0	0	0	0	0	0	**0**
BRO	0	0	0	0	1	0	0	**1**
TCM	0	0	0	0	0	0	0	**0**
WER	1	0	0	1	0	0	0	**2**
ANO	0	8	3	0	7	0	1	**19**
CON	0	6	0	1	0	4	0	**11**
Not Aph	0	0	0	0	1	0	0	**1**
Total by Clinicians	**1**	**14**	**3**	**2**	**9**	**4**	**1**	**34**

Abbreviations: BRO = Broca's aphasia; WER = Wernicke's aphasia; GLO = Global aphasia; ANO = Anomic aphasia; CON = Conduction aphasia; TCM = Transcortical motor aphasia; Not Class = not classified; Not Aph = not aphasic.

with (nonfluent) anomic aphasia (n = 6, 14%) or TCM aphasia (n = 3, 7%). One was labelled as conduction (2%), and 3 were unclassified (7%). (Note that, even when syndrome diagnoses matched, fluency classifications sometimes mismatched. For example, clinical impressions might designate someone with anomic aphasia as nonfluent, whereas WAB-R classifications would always designate anomic aphasia as fluent.)

A large proportion of the 42 mismatches designated fluent by WAB-R were scored as a "5" on the fluency scale (n = 18, 43%), which is described as follows: "Often telegraphic but more fluent speech with some grammatical organization; marked word-finding difficulty. Paraphasias may be prominent; few, but more than two propositional sentences." (Kertesz, 2006, p. 33). Clearly, this is a transitional level between fluent and nonfluent categories. Kertesz notes that those with improving Broca's aphasia often receive this score, but the WAB-R scoring protocol puts it in the fluent range. However, it appears that most clinicians consider such a behavioral profile to warrant a classification of nonfluent aphasia.

Given our finding that many of the mismatches were participants classified as "5" on the WAB-R spontaneous speech scale, we adjusted the fluency cut-off of the WAB-R scale so that ratings of 0–5 were classified as nonfluent and 6–10 as fluent, and re-assessed reliability of the scale relative to clinical impressions. The adjusted scale resulted in fewer mismatches overall (43 to 34), meaning that agreement improved from 83% to 87%).[1] Predictably, this change resulted in a slight reduction of agreement for PwA classified by clinical impression as fluent (down from 99% to 93%), but considerably improved agreement for PwA classified by clinical impression as nonfluent (up from 64% to 79% with the adjusted scale). The syndrome labels by clinical impression for the remaining 34 mismatches are shown in Table 2.[2] With the adjusted scale, mismatches were mostly labelled as anomic aphasia (n = 19, 53%) or conduction aphasia (n = 11, 31%) according to WAB-R criteria, but as Broca's aphasia according to clinical impression (n = 14, 39%). Clinical impressions also designated several as having anomic (n = 9, 25%) or conduction (n = 4, 11%) aphasia. The adjusted scale resulted in a more reliable classification rate overall, so we used this in the subsequent analyses.

Predictors of fluency outcomes

In the following analyses, we identified spontaneous speech characteristics predicting fluency classifications. Means of fluent and nonfluent categories determined by both the adjusted WAB-R fluency scale and by clinical impression are shown in Table 3. As an initial exploration, we conducted *t*-tests on all the variables, using corrected p-values of .05/18 speech/language measures = .0028. Most of the variables show significant differences between fluency categories, whether diagnosed by the WAB-R or by clinical impression. Notable exceptions were the measures of morphology (verb inflection and morphological errors), measures of semantic accuracy and specificity (semantic errors and empty speech), pitch variability, and the presence of dysarthria. Unexpectedly, retraces were more common in fluent than nonfluent aphasia. Length of sample (total utterances) and circumlocution seem to be important for the WAB-R fluency categories but not clinical impressions. The bar graphs in Figure 2 graphically illustrate the differences between fluent and nonfluent aphasia, plotting mean *z*-scores of each of the measures to facilitate comparison across predictors.

To determine the variables contributing *most* to WAB-R and clinician fluency classifications, taking into account any relationships among the predictors, we conducted separate

Table 3. Descriptive statistics and *t*-test results comparing age and spontaneous speech characteristics for fluent and nonfluent categories by WAB-R (adjusted fluency scale) and clinician impression.

	WAB-R Fluency Classification			Clinician Fluency Classification		
	Fluent (n = 153)	Nonfluent (n = 101)	p-value	Fluent (n = 139)	Nonfluent (n = 115)	p-value
	M (SD)	M (SD)		M (SD)	M (SD)	
Age	62.6 (12.6)	57.8 (11.9)	***.002***	63.2 (12.2)	57.7 (12.3)	***<.001***
SpSp Fluency	8.0 (1.4)	3.8 (1.2)	***<.001***	7.9 (1.5)	4.5 (2.0)	***<.001***
WAB-R AQ	81.6 (13.9)	58.3 (14.9)	***<.001***	80.2 (15.4)	62.7 (16.9)	***<.001***
Gram Errs	.195 (.19)	.399 (.26)	***<.001***	.188 (.18)	.383 (.26)	***<.001***
Gram Comp	.065 (.03)	.032 (.03)	***<.001***	.066 (.03)	.035 (.03)	***<.001***
Morph Errs	.002 (.01)	.004 (.01)	.136	.003 (.01)	.004 (.01)	.362
Inflect Vbs	.392 (.09)	.355 (.14)	.026	.389 (.09)	.365 (.14)	.124
Sem Errs	.014 (.02)	.023 (.03)	.023	.014 (.02)	.023 (.03)	.007
Empty	.047 (.08)	.031 (.08)	.128	.052 (.09)	.026 (.07)	.008
Circum	.020 (.04)	.002 (.01)	***<.001***	.017 (.04)	.008 (.03)	.022
Prop Dens	.463 (.05)	.406 (.15)	***<.001***	.466 (.05)	.409 (.14)	***<.001***
MATTR	.905 (.03)	.843 (.08)	***<.001***	.908 (.03)	.847 (.08)	***<.001***
Phon Errs	.010 (.02)	.031 (.05)	***<.001***	.009 (.01)	.030 (.04)	***<.001***
Neo Errs	.008 (.02)	.027 (.05)	***<.001***	.008 (.02)	.025 (.05)	***<.001***
Pitch Var	17.4 (6.6)	20.5 (10.7)	.011	17.3 (6.6)	20.2 (10.2)	.010
Con:Fun	.850 (.50)	1.17 (.92)	***.002***	.799 (.15)	1.196 (1.01)	***<.001***
Pauses	.350 (.17)	.477 (.18)	***<.001***	.350 (.18)	.462 (.18)	***<.001***
Retrace	.275 (.13)	.183 (.15)	***<.001***	.279 (.13)	.190 (.15)	***<.001***
Total Utts	32.7 (22.2)	24.1 (19.3)	***.001***	32.1 (22.7)	26.0 (19.4)	.022
	N (%)	N (%)		N (%)	N (%)	
Apraxia #	20 (.13)	57 (.63)	***<.001***	22 (0.17)	70 (.67)	***<.001***
Dysarthria #	8 (.03)	15 (.17)	.048	10 (0.08)	16 (.16)	.055

Bold italics indicates significant at corrected *p*-values <.0028. # Percentages were calculated excluding those for whom AoS or Dys was not diagnosed.

logistic regression analyses for the two sets of fluency classifications. The outcomes of each final model are presented in Table 4, with WAB-R-determined diagnoses at the top and clinical impressions at the bottom. For each model, significant predictors are listed from the most to least predictive, as indicated by the odds ratio. The odds ratios should be interpreted as the change in the odds of a fluent designation, given an increase of one standard deviation (because the variables are expressed as *z*-scores) in the variable under consideration.

WAB-R criteria

PwA were more likely to be classified as fluent by WAB-R criteria if they had higher WAB-R AQs, produced more complex grammar, and used more diverse words but also more empty speech and semantic errors (all *ps* < .001). The odds of being labelled as "fluent" were more than 18 times more likely with each SD increase in the Aphasia Quotient (OR = 18.53). The odds of a fluent label were quadrupled by each SD increase in lexical diversity (OR = 4.43), and almost tripled for increases in grammatical complexity (OR = 2.97), empty speech (OR = 2.79) and semantic errors (OR = 2.61). The predictive power of the model was calculated using the McFadden Pseudo R^2. Together these measures accounted for almost 60% of the variance in WAB-R fluency diagnoses.

Clinical impressions

PwA were more likely to be classified as fluent by clinical impression if they did not have apraxia of speech, had higher WAB-R AQs, and produced a more diverse sample of words

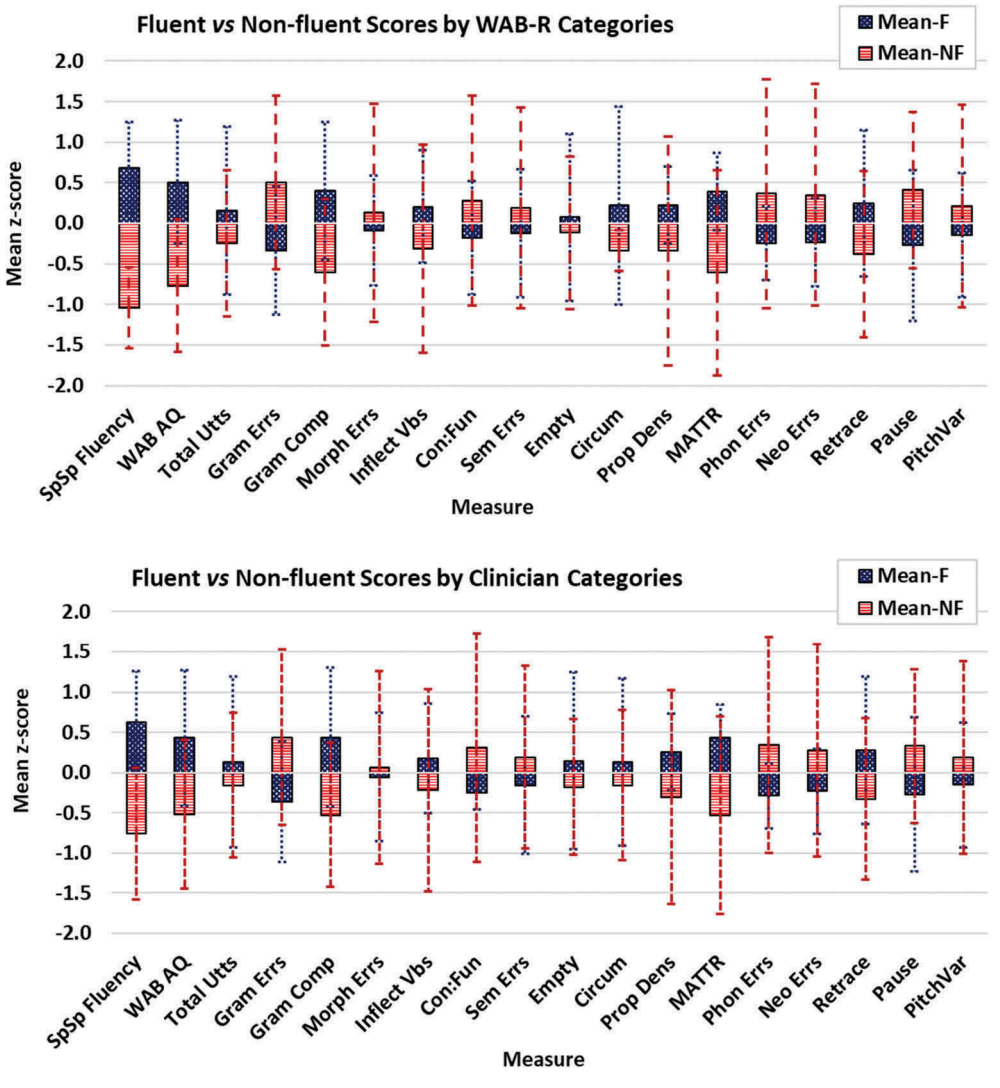

Figure 2. Bar graphs showing mean z-scores of spontaneous speech measures for fluent and nonfluent aphasia, by WAB-R classification (top) and clinician classification (bottom). Error bars indicate standard deviations.

but more empty speech (all *ps* < .001). The absence of apraxia of speech increased the odds of a fluent label over 9 times (OR = 9.31), and each SD increase in AQ more than quadrupled the odds of a fluent label (OR = 4.24). Greater lexical diversity more than tripled the odds (OR = 3.78), while empty speech almost doubled the odds (OR = 1.99) of being classified as fluent by clinical impression. Together these measures accounted for 53% of the variance in clinical impressions of fluency.

Comparing the two models illustrates several things. First, fluency diagnoses by both WAB-R and clinical impression were influenced by aphasia severity and by aspects of word retrieval (lexical diversity and lexical specificity in both, as well as lexical accuracy in the

Table 4. Results of reduced logistic regression models predicting binary WAB-R and clinician fluency classification. Odds ratios represent the likelihood of a fluent classification.

					Confidence intervals	
Significant predictors	Estimate	*z*-value	*p*-value	Odds ratio	2.5%	97.5%
Likelihood of fluent classification by WAB-R (R^2_{McF} = .598, n = 252)						
(Intercept)	0.71	3.05	.002	2.03	1.30	3.25
AQ (severity)	2.92	6.76	<.001	18.53	8.53	47.02
MATTR	1.49	4.20	<.001	4.43	2.30	9.31
Complex grammar	1.09	4.14	<.001	2.97	1.81	5.13
Empty speech	1.03	4.07	<.001	2.79	1.74	4.76
Semantic errors	0.96	3.85	<.001	2.61	1.63	4.38
Likelihood of fluent classification by clinicians (R^2_{McF} = .526, n = 234)						
(Intercept)	0.97	3.77	<.001	2.64	1.61	4.45
AoS-No	2.23	5.30	<.001	9.31	4.20	22.06
AQ (severity)	1.44	5.16	<.001	4.24	2.52	7.59
MATTR	1.33	3.86	<.001	3.78	1.99	7.73
Empty speech	0.91	3.87	<.001	2.49	1.61	4.09

WAB-R model). Second, clinical impressions – but not WAB-R classifications – were influenced by aspects of motor speech (as reflected in the presence of apraxia), whereas WAB-R classifications – but not clinical impressions – were influenced by aspects of grammatical competence (grammatical complexity). Third, in terms of the strength of the predictors, for WAB-R fluency classifications, aphasia severity was paramount, whereas the most important predictor of clinical impressions of fluency was the absence of apraxia of speech. These differences are discussed further below. It is also noteworthy that some of variables that did not distinguish between groups in the preliminary (*t*-test) analyses were shown to be predictive in the regression analyses, once other important variables were accounted for. For example, semantic errors and empty speech increased the odds of a fluent classification according to the WAB-R, when aphasia severity was taken into account. This illustrates the importance of considering the combined effects of multiple sources of variance simultaneously.

Discussion

While fluency is often used dichotomously to categorise aphasia, the utility of this classification depends on the consistency with which fluency categories can be assigned by different judges and different methods of assessment. Our results indicated that, although overall agreement between clinical impressions of fluency and fluency classification based on the WAB-R was quite high (> 80%), there were certain types of aphasia for which agreement was relatively poor. Fluency mismatches were most frequent for individuals with Broca's, anomic and conduction aphasia (see Table 2), which replicates prior research. John and colleagues (2017) found that the most common source of disagreement in aphasia type diagnoses was for clinical diagnoses of Broca's aphasia, over 70% of which were diagnosed as anomic by the WAB-R. Swindell and colleagues (1984) noted that more than half of PwA designated as anomic by the WAB were disputed by clinicians. Among the 10 samples analyzed by Gordon (1998), 5 most closely matched the BDAE profiles of anomic or conduction aphasia; for only one of these did clinicians reached a 2/3 consensus on whether the PwA was better characterised as fluent or nonfluent. Wilson

and colleagues (Wilson, Eriksson, Schneck, & Lucanie, 2018) described similar sources of disagreement between clinician impressions and the WAB-R.

In the current study, clinical impressions classified almost a quarter (42/180 = 23%) of PwA scoring in the fluent range on the standardised WAB-R fluency scale as nonfluent. The relatively low agreement for these PwA was in large part attributable to ambiguity at the midpoint of the scale. The majority of mismatched PwA with ratings of 5 on the WAB-R fluency scale were diagnosed by clinicians with Broca's aphasia. Indeed, the WAB-R manual describes a rating of 5 as recovering Broca's aphasia (Kertesz, 2006) but assigns a classification of "fluent" aphasia .[3] Swindell and colleagues noted that "the WAB reserves the classification of Broca's Aphasia only for the more severe language impairments" (1984, p. 51), classifying milder nonfluent patterns as anomic instead. Although shifting the scale decreased agreement slightly for clinician-diagnosed fluent PwA, it greatly improved agreement for nonfluent PwA, and improved overall agreement of fluency diagnoses in the process. Therefore, we recommend considering WAB-R fluency ratings of 0–5 as an adjusted criterion for nonfluent categorization. This is an appropriate first step to more closely align this standardised measure of fluency with clinical perceptions.

One important reason for these diagnostic disagreements seems to be the many underlying sources of dysfluency, and a lack of consistency about which characteristics of dysfluent speech qualify a PwA to be labelled as "nonfluent". Based on the disagreements here, particularly between anomic and Broca's aphasia, word retrieval difficulties is one of the most contentious characteristics. Although typically thought to be a milder fluent syndrome, anomic aphasia is characterised by primary deficits in word retrieval which may manifest in many dysfluent behaviors, such as hesitations, fillers, revisions, and sentence fragments. However, these do not "qualify" as fluency deficits according to the WAB-R categories. In addition, it is often difficult to determine why behaviors such as hesitations and revisions arise. Thus, word retrieval difficulties may be attributed to agrammatism instead of anomia, generating an aphasia subtype of Broca's aphasia and a classification as nonfluent. Conduction aphasia is often characterised by successive phonemic approximations that can be difficult to distinguish from AoS (Haley et al., 2013) and can lead to similar dysfluencies, like phonemic errors, restarts and revisions, and the impression of effortful speech. Although traditional aphasia classification models such as that instantiated in the WAB-R do not consider these syndromes to be diagnostically nonfluent, clinicians often consider these patterns to be descriptively nonfluent. The divergence of nonfluency as a diagnostic category or a descriptive characteristic is illustrated further by the fact that clinicians themselves occasionally designated PwA as nonfluent even while diagnosing a traditionally fluent syndrome (n = 9). Thus, a second specific recommendation arising from these results is to explicitly consider the impact of word retrieval and phonological encoding difficulties on fluency.

Our regression analyses confirmed a role for the three primary dimensions of spontaneous speech in fluency disruptions in aphasia: grammatical competence, lexical retrieval, and the facility of speech production. For both WAB-R classifications and clinical impressions, fluent and nonfluent PwA were differentiated by measures of lexical diversity and specificity. WAB-R categories were also affected by grammatical complexity. Although there was considerable overlap in the variables contributing to fluency outcomes as measured by the WAB-R scores and by clinical impression (not surprising, given the overall high degree of agreement in fluency classifications), there were some differences.

Most notably, the WAB-R spontaneous speech scale was more strongly influenced by aphasia severity than by any of the specific speech-language measures reflecting underlying dimensions of fluency. Part of the reason for this is, of course, that the fluency scale contributes to the calculation of the WAB-R AQ, so the conflation with severity is an inherent component of the WAB-R fluency scale. But this makes identification of fluency more difficult. As Goodglass and colleagues note, "one of the central problems in classification is to distinguish variations based on severity from those based on type of aphasia" (1964, p. 138). Another notable difference was that clinical impressions were much more strongly influenced by the presence of apraxia of speech, something that is not well captured by the WAB-R scale. These observations lead to two further recommendations. First, if the WAB-R scale is used to measure fluency, it is important to account for its potential confounding with severity, either statistically (e.g., by including severity as a covariate) or by equating groups on severity. Second, the WAB-R scale does not seem to be a useful indicator of fluency for those individuals whose oral expression is significantly affected by apraxia.

Finally, we found that grammatical competence significantly predicted fluency categories by WAB-R, but did not have a significant impact on clinical impressions of fluency. This might indicate that grammatical features are less salient to clinical impression, unless attention is explicitly drawn to them, as it is in the WAB-R scale. Several levels of the scale refer to sentence completeness and grammatical organization. The explicit description of relevant dimensions is one advantage of multidimensional scales such as those in the WAB-R and BDAE over less structured clinical impressions, and is a feature we advocate including in any measure of fluency.

Although fluency is often viewed as a dichotomy (e.g., Goodglass et al., 1964; Helm-Estabrooks, 1992; Kong & Wong, 2018; Park et al., 2011), it is clear from these results that finding a workable cut-off point between fluent and nonfluent aphasia can be difficult. Adjusting the WAB-R scale improved its overall correspondence to clinical intuition, but there remained mismatches. Furthermore, we suspect that the level of agreement on fluency classification is over-estimated in our study, since some of the clinical impressions of fluency were based, at least in part, on the WAB-R. The administration protocol of AphasiaBank asks contributors to document the basis for their classification. No basis was identified for 32 (12.6%) of the PwA and the WAB-R was specifically mentioned for only 20 (9.0%) of the PwA, although it might have influenced perceptions nonetheless. For the vast majority of PwA, judgements were based on some combination of formal testing, clinical interaction, and observation.

One solution to the lack of a clear cut-off between fluent and nonfluent aphasia has been to measure fluency as a continuum. Although the WAB-R fluency scale functions to automatically assign PwA to a fluent or nonfluent subtype, the 11-point scale can also be used continuously. Using it this way may improve sensitivity in detecting the underlying deficits that disrupt fluency along the continuum. Researchers also use objective continuous measures as proxies for fluency, such as mean utterance length (e.g., Halai, Woollams, & Lambon Ralph, 2017) and speech rate (e.g., Wang, Marchina, Norton, Wan, & Schlaug, 2013). However, it is an open question to what extent these measures adequately cover the range of fluency behaviours. In a companion paper (Gordon & Clough, 2020), we investigate whether such continuous measures can capture relevant

aspects of fluency. In addition, given the multidimensional nature of fluency, it may be difficult to find measures on which clinicians and researchers can agree.

To address these issues, we propose an alternative solution to measuring fluency. When a clinician identifies a behavior that disrupts the flow of speech, it may be less helpful to ask whether the participant is diagnostically fluent or nonfluent and more appropriate to unpack the deficits contributing to the dysfluency of expressive language and to acknowledge the continuous variability of such underlying deficits. In her indictment of the WAB fluency scale, Trupe (1984) noted that "a multidimensional scale such as the WAB fluency scale … cannot reliably be used to score spontaneous speech," asserting that each dimension should be evaluated separately to be useful. Casilio et al. (2019) followed this approach, noting that PWA may be more accurately understood as points in a multidimensional symptom space. Their factor analysis revealed the same three dimensions we identified here as underlying fluency – word retrieval, grammatical formulation and speech production. Although our findings indicate that clinicians already attend to these three dimensions in judging fluency, they lack a tool that allows for an objective and consistent assessment of how they contribute to fluency.

Developing such a tool is a primary goal of our ongoing project. An assessment that standardises measures directly related to each of these underlying components would help clinicians identify the deficits contributing to each client's profile of fluency impairment. Doing so would not only improve diagnostic reliability, but would also allow more specific identification of targets for therapy and tracking of recovery over time. In addition, an assessment process that focuses on the speech-language deficits underlying disrupted fluency is in keeping with state-of-the-art approaches to rehabilitation treatment specification, in which clinicians are encouraged to identify the active ingredients of therapy and the mechanisms by which ingredients affect targeted components (e.g., Van Stan et al., 2019).

Limitations

Our results are limited in several ways. As a retrospective study, it was not possible to control, beyond the constraints of the AphasiaBank protocol, the conditions under which clinical impressions were made, or who made them. Nevertheless, such variability is an inherent component of clinical practice, and the unconstrained use of clinical impression was of explicit interest in the current study. Although we examined a large number of variables from different domains of spoken language production, it is possible that there are aspects of spoken language not measured here that may also contribute to differences in fluency classification. Among those we did include, the measures may not adequately capture the range of variance in each hypothesized dimension, a potential issue of content validity. In addition, low incidence occurrences, such as morphological errors, may not show up as important in a group analysis, but may be much more salient for some individual PwA. As in any study, our findings are also dependent on the particular set of participants analyzed. The large sample provided by AphasiaBank is a strength of the current study, and it is representative except for the relative lack of individuals with global aphasia. A sample that included more speakers with global aphasia might, for example, have shown

stronger influences of grammatical accuracy measures or pausing on fluency diagnoses. Finally, it is important to acknowledge that measuring fluency is just one aspect of aphasia diagnosis, and clinicians make use of many other sources of information in identifying syndromes, judging fluency, and making other clinical decisions. Nevertheless, fluency remains an integral component of aphasia diagnosis and, as such, the field needs a more reliable and standardized approach to its measurement.

Summary and conclusions

In this study, we examined the concordance of two common methods of classifying fluency in aphasia – using the WAB-R fluency scale and relying on clinical impressions – and investigated why disagreements arise. Our analyses revealed three shortcomings of the WAB-R method: 1) it is confounded with severity of aphasia; 2) it does not capture the influence of motor speech impairments on fluency; and 3) the cut-off between fluent and nonfluent does not accord well with clinical impression. A disadvantage of clinical impressions was the relative lack of consideration of grammaticality. We make several concrete recommendations to account for these issues. Many before us have pointed out problems with fluency classification, particularly as implemented in the WAB-R. However, its persistence suggests that it serves as a helpful cognitive heuristic in understanding how aphasia manifests in oral expressive behaviours. Our aim here was to move the discussion forward by revealing why fluency classifications may be unreliable. Based on these findings and those in our companion paper (Gordon & Clough, 2020), we are developing an approach to fluency measurement that takes into account its multidimensionality and the continuity of its underlying dimensions.

Notes

1. In response to a reviewer's query, we also calculated agreement for the subset of PwA (n = 169) whose AQ data and clinical impressions were provided by the AphasiaBank investigators. Agreement rates between clinical impressions of fluency and AQ fluency classification were 81% and 85%, respectively, for the original and adjusted WAB-R fluency scale. These rates did not differ from the agreement rates for the remainder of the set (n = 85), who were judged by external clinicians/researchers ($p = .269$ for the original WAB-R scale; $p = .421$ for the adjusted scale).
2. Note that we did not take the liberty of changing the WAB-R syndrome diagnoses after adjusting the fluency scale, although some syndrome types might change according to WAB-R criteria.
3. It should be acknowledged that these issues have been written about extensively by Kertesz and his colleagues, and the manual for the test acknowledges the arbitrary nature of the cut-off score between fluent and nonfluent syndromes.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was generously supported by a New Century Scholars Grant from the American Speech-Language-Hearing Foundation awarded to the two authors.

ORCID

Jean K. Gordon http://orcid.org/0000-0003-0073-8016

References

Albert, M. L., & Sandson, J. (1986). Perseveration in aphasia. *Cortex*, *22*, 103–115. doi:10.1016/S0010-9452(86)80035-1

Alyahya, R. S. W., Halai, A. D., Conroy, P., & Lambon Ralph, M. A. (2018). Noun and verb processing in aphasia: Behavioural profiles and neural correlates. *NeuroImage: Clinical*, *18*, 215–230. doi:10.1016/j.nicl.2018.01.023

Andreetta, S., & Marini, A. (2015). The effect of lexical deficits on narrative disturbances in fluent aphasia. *Aphasiology*, *29*, 705–723. doi:10.1080/02687038.2014.979394

Baken, R. J., & Orlikoff, R. F. (2000). *Clinical measurements of speech and voice* (2nd ed.). Albany, NY: Thomson Delmar Learning.

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

Bastiaanse, R., Rispens, J., Ruigendijk, E., Rabadan, O. J., & Thompson, C. K. (2002). Verbs: Some properties and their consequences for agrammatic Broca's aphasia. *Journal of Neurolinguistics*, *15*, 239–264. doi:10.1016/S0911-6044(01)00032-X

Berndt, R. S., & Caramazza, A. (1980). A redefinition of the syndrome of Broca's aphasia: Implications for a neuropsychological model of language. *Applied Psycholinguistics*, *1*, 225–278. doi:10.1017/S0142716400000552

Boersma, P., & Weenink, D. (2017). Praat: Doing phonetics by computer (Version 6.03) [Computer software]. Retrieved from http://www.praat.org/

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

Bunton, K., Kent, R. D., Duffy, J. R., Rosenbek, J. C., & Kent, J. F. (2007). Listener agreement for auditory-perceptual ratings of dysarthria. *Journal of Speech, Language, and Hearing Research*, *50*, 1481–1495. doi:10.1044/1092-4388(2007/102)

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. doi.10.1044/2018_AJSLP-18-0192.

Code, C., & Rowley, D. (1981). Age and aphasia type: The interaction of sex, time since onset and handedness. *Aphasiology*, *1*, 339–345. Retrieved from https://doi.org/10.1080/02687038708248854

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

Covington, M., & McFall, J. D. (2008). The moving-average Type-Token ratio. Paper presented at the Linguistics Society of America, Chicago, IL.

Duffy, J. (2013). *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management* (3rd ed. ed.). St. Louis, MO: Elsevier/Mosby.

Edwards, S. (2005). *Fluent aphasia*. Cambridge, UK: Cambridge University Press.

Edwards, S., & Bastiaanse, R. (1998). Diversity in the lexical and syntactic abilities of fluent aphasic speakers. *Aphasiology*, *12*, 99–117. doi:10.1080/02687039808250466

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*, 1–21. doi:10.1080/02687039108248516

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

GoldWave Inc. (2017). Goldwave (Version 6.26) [Computer software]. Retrieved from http://www.goldwave.com.

Goodglass, H., & Kaplan, E. (1983). *The Assessment of Aphasia and Related Disorders*. Philadelphia, PA: Lea & Febiger.

Goodglass, H., Kaplan, E., & Barresi, B. (2001). *Boston Diagnostic Aphasia Examination* (3rd ed.). Philadelphia, PA: Lippincott, Williams & Wilkins.

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

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

Gordon, J. K. (2006). A quantitative production analysis of picture description. *Aphasiology*, *20*, 188–204. doi:10.1080/02687030500472777

Gordon, J. K. (2008). Measuring the semantic specificity of picture description in aphasia. *Aphasiology*, *22*, 839–852. doi:10.1080/02687030701820063

Gordon, J. K., & Clough, S. (2020). How fluent? Part B. Underlying contributors to continuous measures of fluency in aphasia. *Aphasiology*. *this issue*.

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

Haley, K. L., Jacks, A., & Cunningham, K. T. (2013). Error variability and the differentiation between apraxia of speech and aphasia with phonemic paraphasia. *Journal of Speech, Language, and Hearing Research*, *56*, 891–905. doi:10.1044/1092-4388(2012/12-0161)

Hazamy, A. A., & Obermeyer, J. (2019). Evaluating informative content and global coherence in fluent and non-fluent aphasia. *International Journal of Communication Disorders*, *55*, 110–120. doi:10.1111/1460-6984.12507

Helm-Estabrooks, N. (1992). *Aphasia Diagnostic Profiles*. Austin, TX: 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*, 176-180.

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). London: Eng.: J.A. Churchill.

John, A. A., Javali, M., Mahale, R., Mehta, A., Acharya, P. T., & Srinivasa, R. (2017). Clinical impression and Western Aphasia Battery classification of aphasia in acute ischemic stroke: Is there a discrepancy?. *Journal of Neurosciences in Rural Practice*, *8*, 74–78. doi:10.4103/0976-3147.193531

Kempster, G. B., Gerratt, B. R., Abbott, K. V., Barkmeier-Kraemer, J., & Hillman, R. E. (2009). Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol. *American Journal of Speech-Language Pathology*, *18*, 124–132. doi:10.1044/1058-0360(2008/08-0017)

Kent, R. (2009). Perceptual sensorimotor speech examination for motor speech disorders. In M. R. McNeil (Ed.), *Clinical Management of Sensorimotor Speech Disorders* (pp. 19–29). New York, NY: Thieme Medical Publishers.

Kertesz, A. (2006). *Western Aphasia Battery-Revised*. San Antonio, TX: Pearson.

Kiran, S., Cherney, L. R., Kagan, A., Haley, K. L., Antonucci, S. M., Schwartz, M. S., … Simmons-Mackie, N. (2018). Aphasia assessments: A survey of clinical and research settings. *Aphasiology*, *32(S1)*, 47–49. doi:10.1080/02687038.2018.1487923

Kong, A. P.-H., & Wong, C. W.-Y. (2018). An integrative analysis of spontaneous storytelling discourse in aphasia: Relationship with listeners' rating and prediction of severity and fluency status of aphasia. *American Journal of Speech-Language Pathology*, *27*, 1491–1505. doi:10.1044/2018_AJSLP-18-0015

Laganaro, M. (2012). Patterns of impairments in AoS and mechanisms of interaction between phonological and phonetic encoding. *Journal of Speech, Language, & Hearing Research*, *55*, S1535–S1543. doi:10.1044/1092-4388(2012/11-0316)

Lee, J. B., Kocherginsky, M., & Cherney, L. R. (2018). Attention in individuals with aphasia: Performance on Conners' Continuous Performance Test-2nd Edition. *Neuropsychological Rehabilitation*. doi:10.1080/09602011.2018.1460852

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk: Transcription format and programs* (3rd ed.). Lawrence Erlbaum Associates Publishers.

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

Marshall, R. C., & Tompkins, C. A. (1982). Verbal self-correction behaviors of fluent and nonfluent aphasic subjects. *Brain & Language*, *15*, 292–306. doi:10.1016/0093-934X(82)90061-X

McNeil, M. R., Robin, D. A., & Schmidt, R. A. (2009). Apraxia of speech: Definition and differential diagnosis. In M. R. McNeil (Ed.), *Clinical management of sensorimotor speech disorders* (pp. 249–268). New York, NY: Thieme Medical Publishers.

Obler, L. K., Albert, M. L., Goodglass, H., & Benson, D. F. (1978). Aphasia type and aging. *Brain and Language*, *6*, 318–322. doi:10.1016/0093-934X(78)90065-2

Pakhomov, S., & Kotlyar, M. (2011). Prosodic correlates of individual physiological response to stress. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech, Aug. 2011, Florence, Italy, 2949-2952..

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

Pedersen, P. M., Vinter, K., & Olsen, T. S. (2004). Aphasia after stroke: Type severity and prognosis. *Cerebrovascular Disease*, *17*, 35–43. doi:10.1159/000073896

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

Rochon, E., Saffran, E. M., Berndt, R. S., & Schwartz, M. F. (2000). Quantitative analysis of aphasic sentence production: Further development and new data. *Brain & Language*, *72*, 193–218. doi:10.1006/brln.1999.2285

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

Snowden, D., Kemper, S., Mortimer, J. T., Greiner, L. H., Wekstein, D. R., & Marksbery, W. R. (1996). Linguistic ability in early life and cognitive function and Alzheimer's disease in late life. *Journal of the American Medical Association*, *275*, 528–532. doi:10.1001/jama.1996.03530310034029

Strand, E. A., Duffy, J. R., Clark, H. M., & Josephs, K. (2014). The apraxia of speech rating scale: A tool for diagnosis and description of apraxia of speech. *Journal of Communication Disorders*, *51*, 43–50. doi:10.1016/j.jcomdis.2014.06.008

Swindell, C. S., Holland, A., & Fromm, D. (1984). *Classification of aphasia: WAB type versus clinical impression*. Clinical Aphasiology Conference Proceedings, Seabrook Island, SC.

Thompson, C. K., & Bastiaanse, R. (2012). Introduction to agrammatism. In R. Bastiaanse & C. K. Thompson (Eds.), *Perspectives on Agrammatism* (pp. 1–16). London, UK: Psychology Press.

Thompson, C. K., Lange, K. L., Schneider, S. L., & Shapiro, L. P. (1997). Agrammatic and non-brain-damaged subjects' verb and verb argument structure production. *Aphasiology*, *11*, 473–490. doi:10.1080/02687039708248485

Tremblay, P., & Dick, A. S. (2016). Broca and Wernicke are dead, or moving past the classic model of language neurobiology. *Brain & Language*, *162*, 60–71. doi:10.1016/j.bandl.2016.08.004

Trupe, E. H. (1984). *Reliability of rating spontaneous speech in the Western Aphasia Battery: Implications for classification*. Clinical Aphasiology Conference Proceedings, Seabrook Island, SC.

Van Stan, J. H., Dijkers, M. P., Whyte, J., Hart, T., Turkstra, L. S., Zanca, J. M., & Chen, C. (2019). The rehabilitation treatment specification system: Implications for improvements in research design, reporting, replication, and synthesis. *Archives of Physical Medicine and Rehabilitation*, *100*, 146–155. doi:10.1016/j.apmr.2018.09.112

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

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*, 831. doi.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. Clinical Aphasiology Conference Proceedings, Seabrook Island, SC.

Whitehouse, P., Caramazza, A., & Zurif, E. B. (1978). Naming in aphasia: Interacting effects of form and function. *Brain & Language*, *6*, 63–74. doi:10.1016/0093-934X(78)90044-5

Wilson, S. M., Eriksson, D. K., Schneck, S. M., & Lucanie, J. M. (2018). A quick aphasia battery for efficient, reliable, and multidimensional assessment of language function. *PloS One*, *Feb. 9*, 1–29. Retrieved from https://doi.org/10.1371/journal.pone.0192773

Zingeser, L. B., & Berndt, R. S. (1990). Retrieval of nouns and verbs in agrammatism and anomia. *Brain & Language*, *39*, 14–32. doi:10.1016/0093-934X(90)90002-X

Appendix 1

Inter-correlations among predictor variables, and between predictors and spontaneous speech fluency scale scores

Measures	Gram Acc	Gram Comp	Morph Errs	Inflect Vbs	Sem Errs	Empty	Circum	Prop Dens	MATTR	Phon Errs	Neo Errs	AoS	Dys	Pitch Var	Con: Fun	Pauses	Retrace	Total Utts
SpSp Fluency	-0.341	0.507	-0.062	0.191	-0.252	-0.017	0.253	0.362	0.557	-0.324	-0.305	-0.346	-0.149	-0.169	-0.179	-0.308	0.308	0.185
WAB-R AQ	-0.253	0.388	0.039	0.218	-0.415	-0.250	0.189	0.286	0.494	-0.196	-0.263	-0.219	0.006	-0.197	-0.066	-0.166	0.294	0.153
Gram Acc		-0.335	0.244	-0.086	0.012	-0.031	-0.172	0.001	-0.251	0.260	0.197	0.192	0.080	0.095	0.474	0.125	-0.088	-0.049
Gram Comp			-0.162	0.061	-0.152	0.143	0.229	0.307	0.380	-0.243	-0.243	-0.342	-0.100	-0.181	-0.014	-0.350	0.350	0.016
Morph Errs				-0.069	-0.041	-0.015	-0.050	-0.048	-0.013	0.199	0.035	0.118	0.034	-0.042	0.081	0.068	0.030	0.076
Inflect Vbs					-0.109	0.038	0.081	0.068	0.178	-0.081	-0.019	-0.086	0.011	-0.210	-0.052	0.037	0.095	-0.236
Sem Errs						0.003	-0.057	-0.015	-0.169	0.154	0.250	0.064	-0.008	-0.013	-0.012	0.140	-0.115	-0.023
Empty							0.058	0.017	0.032	-0.009	0.019	-0.134	-0.115	-0.020	-0.176	-0.140	0.155	-0.047
Circum								0.098	0.195	-0.135	-0.141	-0.096	-0.099	-0.171	-0.095	-0.105	0.287	0.023
Prop Dens									0.408	-0.163	-0.170	-0.277	-0.112	-0.165	-0.088	-0.224	0.274	0.146
MATTR										-0.227	-0.265	-0.330	-0.087	-0.105	-0.146	-0.186	0.307	0.087
Phon Errs											0.414	0.247	0.258	0.118	0.101	0.225	-0.146	-0.029
Neo Errs												0.255	0.089	0.078	0.087	0.101	-0.080	-0.093
AoS													0.216	0.074	0.187	0.239	-0.135	-0.041
Dys														-0.038	0.047	0.063	-0.091	-0.107
Pitch Var															0.084	-0.139	-0.251	0.236
Con:Fun																0.092	-0.098	-0.153
Pause																	-0.276	-0.300
Retrace																		0.140

Note: Please see Table 1 for an explanation of each variable. Shading indicates the strength of the correlation according to Cohen's (1988) benchmarks: *r* > .10 but < .30 = small (light shading); *r* = /> .30 but < .50 = medium (darker shading), r = /> .50 = large (bright shading). Red shades indicate negative correlations; blue positive correlations.