ASHA American Speech-Language-Hearing Association

AJSLP

Research Article

The Flu-ID: A New Evidence-Based Method of Assessing Fluency in Aphasia

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

[a] Department of Communicative Disorders, The University of Rhode Island, Kingston [b] Multimodal Language Department, Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands

ARTICLE INFO

Article History:
Received November 13, 2023
Revision received May 9, 2024
Accepted July 24, 2024

Editor-in-Chief: Rita R. Patel
Editor: Brent Archer

https://doi.org/10.1044/2024_AJSLP-23-00424

ABSTRACT

Purpose: Assessing fluency in aphasia is diagnostically important for determining aphasia type and severity and therapeutically important for determining appropriate treatment targets. However, wide variability in the measures and criteria used to assess fluency, as revealed by a recent survey of clinicians (Gordon & Clough, 2022), results in poor reliability. Furthermore, poor specificity in many fluency measures makes it difficult to identify the underlying impairments. Here, we introduce the Flu-ID Aphasia, an evidence-based tool that provides a more informative method of assessing fluency by capturing the range of behaviors that can affect the flow of speech in aphasia.
Method: The development of the Flu-ID was based on prior evidence about factors underlying fluency (Clough & Gordon, 2020; Gordon & Clough, 2020) and clinical perceptions about the measurement of fluency (Gordon & Clough, 2022). Clinical utility is maximized by automated counting of fluency behaviors in an Excel template. Reliability is maximized by outlining thorough guidelines for transcription and coding. Eighteen narrative samples representing a range of fluency were coded independently by the authors to examine the Flu-ID's utility, reliability, and validity.
Results: Overall reliability was very good, with point-to-point agreement of 86% between coders. Ten of the 12 dimensions showed good to excellent reliability. Validity analyses indicated that Flu-ID scores were similar to clinician ratings on some dimensions, but differed on others. Possible reasons and implications of the discrepancies are discussed, along with opportunities for improvement.
Conclusions: The Flu-ID assesses fluency in aphasia using a consistent and comprehensive set of measures and semi-automated procedures to generate individual fluency profiles. The profiles generated in the current study illustrate how similar ratings of fluency can arise from different underlying impairments. Supplemental materials include an analysis template, extensive guidelines for transcription and coding, a completed sample, and a quick reference guide.
Supplemental Material: https://doi.org/10.23641/asha.27078199

The most widely used metric by which narrative skills in aphasia are judged is fluency. As a feature of connected speech, fluency is multidimensional in nature, reflecting the interaction of several language production skills, from conceptual planning to lexical retrieval and grammatical formulation to articulation. The popularity of the fluency construct stems from its ability to capture a myriad of language production deficits that occur in aphasia and, thus, to reflect functional verbal competence in natural speaking contexts. That said, not all language production skills fall under the umbrella of fluency. Oral expression may be fluent but rendered nonsensical or empty of meaning by inaccuracies of conceptual flow (as in cases of schizophrenia), lexical retrieval errors (as in jargon aphasia or anomic aphasia), and/or grammatical formulation difficulties (as in paragrammatism). It is important, therefore, to distinguish not only between underlying sources of disfluency but also between disorders of fluency and other expressive impairments. The assessment procedure described in the current article provides an evidence-based method to both identify factors contributing to disfluency in individuals with aphasia and to quantify their impact.

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

Downloaded from: https://pubs.asha.org University of Rhode Island on 10/10/2024, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

The production of fluent speech requires "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" (Gordon & Clough, 2020, p. 1521). The involvement of these particular subtasks was supported by our prior work examining predictors of several different metrics of fluency in story retelling samples from a large and varied sample of 254 people with aphasia (PwA) stored in AphasiaBank (MacWhinney et al., 2011). Clough and Gordon (2020) conducted logistic regressions and found that the likelihood of classification of aphasia type as "fluent" or "nonfluent" on the Western Aphasia Battery–Revised (WAB-R; Kertesz, 2006) Fluency scale depended on aphasia severity, grammatical complexity, and several measures of lexical retrieval (lexical diversity, empty speech, and semantic errors). In the same study, fluent/nonfluent classifications by clinical impression were dependent on the presence of apraxia of speech, aphasia severity, and lexical retrieval (lexical diversity and empty speech). A companion study (Gordon & Clough, 2020) used linear regression to predict the degree of fluency, as reflected in utterance length and speech rate, two common proxy measures of fluency (Cordella et al., 2024). Mean utterance length (MLU) was influenced most heavily by grammatical and lexical variables: grammatical complexity, propositional density, lexical diversity, and content–function word ratio. Speech rate was predicted by these same four variables, as well as two articulatory variables: pitch variation and the presence of apraxia of speech. Together, these analyses demonstrated the multidimensional nature of fluency while revealing that the particular subskills reflected may vary according to the fluency metric used.

These studies highlighted a long-recognized disadvantage of fluency measures—a lack of agreement about the linguistic subskills that are of primary importance in measuring fluency, a finding reinforced by a recent scoping review (Cordella et al., 2024). Forty years ago, Trupe (1984) attributed low agreement on the WAB Fluency scale to its multidimensionality, noting that several different ratings might be justified for a single speaker, depending on which variables were attended to. The Boston Diagnostic Aphasia Examination–Third Edition (BDAE-3; Goodglass et al., 2001a), in which oral expression is separately rated on several fluency-relevant dimensions—melodic line, phrase length, articulation, grammatical form, paraphasia, and word-finding—does not fare much better. Gordon (1998) asked 24 clinicians to rate 10 PwA on each BDAE dimension and identify the person with aphasia as fluent or nonfluent. Clinicians agreed on a fluency classification for only half of the PwA, using a fairly lax criterion of agreement by two thirds of the clinicians. Across clinicians, ratings for a given person with aphasia on a given 7-point scale were typically spread over a range of 5 or 6 points. Rating variability was highest for the lexical retrieval and articulation scales. Thus, even breaking fluency down into its components did not seem to substantially improve the reliability of ratings. A more objective method of measurement is needed to accomplish this.

In a more recent and extensive examination of clinical perceptions of fluency, Gordon and Clough (2022) collected ratings along a visual analogue scale (VAS) from 112 clinicians on eight dimensions (SPEECH RATE, PAUSING, EFFORT, MELODY, PHRASE LENGTH, GRAMMATICALITY, LEXICAL RETRIEVAL, as well as OVERALL FLUENCY)[1] and asked them about their clinical use of the fluency concept. For each respondent in that study, speech samples from 10 or 20 PwA were randomly selected from a subset of 185 PwA from AphasiaBank. Interrater reliability was good for ratings of SPEECH RATE, PAUSING, and PHRASE LENGTH, but only fair for ratings of EFFORT and LEXICAL RETRIEVAL, replicating findings from Gordon (1998). Ratings averaged over respondents showed significant relationships with corresponding objective measures, attesting to their validity. Measures of utterance length and speech rate, the two strongest predictors, predicted ratings in a manner consistent with the earlier linear regression models (Gordon & Clough, 2020). Specifically, objectively measured utterance length was most strongly associated with ratings of GRAMMATICALITY and LEXICAL RETRIEVAL, whereas objectively measured speech rate was more strongly associated with ratings of PAUSING, EFFORT, and MELODY, demonstrating its association with articulatory aspects of speech.

In addition to evidence that fluency perceptions—just like objective measures of fluency—are affected by multiple underlying dimensions of speech production, survey responses from Gordon and Clough (2022) attested to the perceived importance of fluency measurement for clinical practice. The vast majority of respondents (89%) reported using some kind of fluency measurement in their clinical practice; of those who did not, most worked minimally with PwA. All but one of the respondents who measured fluency reported using a variety of methods to do so, with the most common method being subjective judgment. The least common methods were the WAB-R Fluency scale, measurement of speech rate (despite its ability to capture the most dimensions of fluency; Gordon & Clough, 2020), and grammaticality (despite its demonstrated importance in predicting fluency; Clough & Gordon, 2020; Gordon & Clough, 2020). Grammaticality, in addition to being among the least frequent, was also judged to be the least important dimension. Over 90% of the

[1]Small caps are used to differentiate dimensions rated by clinicians (Gordon & Clough, 2022) from dimensions used in the current study.

Downloaded from: https://pubs.asha.org University of Rhode Island on 10/10/2024, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

respondents endorsed the need for a more reliable measure of fluency, with over half providing strong endorsement for this need. The measures that were implemented appeared to depend more on practical factors such as the time available for assessment than on the impact of underlying dimensions on fluency, which is consistent with prior work (e.g., Bryant et al., 2017; Rose et al., 2014). Our survey findings suggest that the construct of fluency is widely used in clinical contexts and is acknowledged to be complex (see also Cordella et al., 2024). However, there remains substantial variability in how fluency is perceived and what dimensions are considered most important to measure.

The method of measuring fluency matters not only to achieve reliability but also to validly represent the degree and nature of dysfluency for specific individuals with aphasia. This is because the various sources of dysfluency may dissociate across individuals. For example, a given PwA may be deemed significantly nonfluent when judged by the apparent effort of their speech production but relatively fluent when judged by the grammaticality of their utterances. Another may speak agrammatically but with intonational contours that create the impression of fluency. Evidence supporting this comes from a recent factor analysis of language production in aphasia (Gordon, 2020), in which the identified factors reflected different aspects of fluency. One reflected basic phrase-building skills, evident in measures of propositional density and verb inflection. Another reflected narrative-level fluency, evident in rate of speech and total words produced. A third reflected repair behaviors, and a fourth grammatical complexity. In addition, a latent profile analysis of the same data generated seven aphasia profiles, of which only four showed a clear fluency pattern: two profiles consisted of 90% or more fluent PwA; two consisted of 90% or more nonfluent PwA; the other three consisted of a mix of fluent and nonfluent types (71%/29%, 38%/62%, and 42%/58%). The idea that elements of fluency dissociate in individuals (also discussed by Hula et al., 2010, with reference to the WAB-R Fluency scale) was represented by Gordon and Clough (2020, Figure 5) as separate vectors in multidimensional space, as opposed to the more typical conception of fluency as a linear dimension with "fluent" at one end and "nonfluent" at the other.

Acknowledging that dimensions of fluency may fractionate at the individual level will also facilitate goal setting in aphasia therapy. To put this in terms of the Rehabilitation Treatment Specification System (Fridriksson et al., 2021; Zanca et al., 2019), treatment should be based on a theory of therapy that identifies targets (e.g., increased fluency), ingredients of the treatment approach (e.g., elements of Melodic Intonation Therapy: Albert et al., 1973; Curtis et al., 2020; or entrainment: Fridriksson et al., 2012; Kershenbaum et al., 2023), and mechanisms of action through which ingredients are hypothesized to achieve targets. Understanding the source or mechanism of breakdown in a given individual is essential to proposing methods by which the deficit may be remediated, and this allows the clinician to make appropriate decisions about treatment targets and ingredients.

To address these issues, the goal of the current study was to develop a method of assessing fluency that embraces the multidimensionality of the construct while improving the reliability of measurement and providing information about underlying contributors for individual PwA, which can then be used to direct treatment. Importantly, our goal is not to facilitate global binary diagnoses of "fluent" versus "nonfluent." Although these terms may serve a useful purpose as shorthand labels for prototypical complexes of behaviors (Goodglass et al., 2001b), a dichotomous classification can be misleading and overly simplistic at an individual level, as discussed above and in our prior work (Gordon & Clough, 2022). The method proposed here, called the Flu-ID ("floo-eye-dee" for its aim of identifying underlying contributors to fluency), is intended to balance an evidence-based approach with the practical demands of clinical time constraints. The accommodation of time constraints is achieved by automating several of the measures. This, along with detailed guidelines regarding use of the assessment and the interpretation of the findings, is also intended to enhance reliability of measurement. Enhanced validity is achieved by allowing the identification of specific variables affecting fluency in individual speakers.

Methods: Development of the Flu-ID

The Flu-ID is a method for identifying in individuals with aphasia the underlying factors affecting the fluency of their speech production and the degree to which fluency is affected. To assess fluency, narrative speech samples are transcribed and segmented into utterances using an Excel macro-enabled template (provided in Supplemental Material S1). Fluency-relevant behaviors are coded in the sample, allowing automated counts of their frequency of occurrence. From these counts, dimension scores are generated and displayed visually on a color-coded graph. Dimension scores are classified into three broad domains of language production—Grammatical Competence (GC), Lexical Availability (LA), and Articulatory Facility (AF). Domain scores are generated by averaging dimension scores to indicate the extent to which each of these affects the fluency of language production. As an alternative to the coding process, the assessment also allows a Quick Version, which involves subjectively rating each fluency behavior. Extensive evidence-based

Downloaded from: https://pubs.asha.org University of Rhode Island on 10/10/2024, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

guidelines and a coding reference sheet are provided for each stage of the process (see Supplemental Materials S2 and S3, respectively). Supplemental Material S4 provides a sample completed transcript on the Excel template.

To develop and assess the Flu-ID, we coded a set of 18 test samples from the AphasiaBank database (MacWhinney et al., 2011). Because we used previously collected data, no institutional approval was required. (See the AphasiaBank website, https://talkbank.org/share/irb/, for details about the consenting processes used to contribute data to AphasiaBank.)

Narrative Speech Samples

The development of the Flu-ID was based on samples from PwA retelling the story of Cinderella from AphasiaBank. It is intended for use with narrative samples, because narrative is less likely than conversation to yield sentence fragments. Because conversation is coconstructed (Carragher et al., 2023; Clark, 1996; Goodwin, 1979; Hengst, 2020), it seeks to minimize joint effort and increase efficiency by reducing the complexity of referring expressions (Clark & Wilkes-Gibb, 1986). This provides many opportunities for a speaker to use elliptical speech that takes advantage of shared knowledge and information in prior utterances (e.g., Speaker 1: *Where did you eat?* Speaker 2: *In the 7th floor caf*), and this might render an overestimate of fluency (Tavakoli, 2016). Monologic narrative, on the other hand, better reflects a speaker's ability to construct and connect utterances independently, just as many other language assessment tasks aim to remove cues from the conversation partner and the context. Although it is important to acknowledge that decontextualized linguistic tasks do not fully reflect the functional communication competence of PwA in interactive and situated communication settings (Doedens & Meteyard, 2022), they can provide detailed characterization of verbal language abilities and impairment profiles, as is the goal here. With this aim in mind, we further recommend using narrative tasks without visual support (e.g., story retelling vs. picture description), as the presence of a shared visual stimulus may mask clinically relevant differences (Fergadiotis & Wright, 2011; Fergadiotis et al., 2011). Narrative tasks have also been shown to elicit more grammatically complex and lexically varied samples than picture description (Bose et al., 2022; Schnur & Wang, 2023; Stark, 2019).

In the Flu-ID guidelines, methods for transcription are recommended to ensure consistency and to facilitate the analysis of fluency. We make use of prior methods of analysis where possible to further enhance consistency and validity (e.g., Berndt et al., 2000; Bernstein Ratner & Brundage, 2022; MacWhinney et al., 2011). For example, Bernstein Ratner and Brundage (2022) recommend parsing utterances on the basis of conversational units, as defined by pauses, intonation, and grammatical structure. We use these cues as well; however, in keeping with our current goals and our focus on narrative, we follow guidelines from the Quantitative Production Analysis (QPA; Berndt et al., 2000; Saffran et al., 1989), prioritizing syntactic and prosodic cues. A sample of 20–30 utterances is recommended to maximize reliability while maintaining clinical feasibility. Our previous work (Clough & Gordon, 2020; Gordon & Clough, 2020) indicates that PwA produce an average of 29 utterances in the Cinderella task (32 for fluent aphasia, 26 for nonfluent aphasia), equating to about 187 words (241 for fluent, 121 for nonfluent).

Dimensions Reflecting Fluency

Many of the coding conventions used in the Flu-ID are similar to those used in AphasiaBank (MacWhinney et al., 2011) but were selected specifically to focus on aphasia fluency behaviors (and adapted where necessary for use in Microsoft Excel). Twelve measures (or dimensions) were ultimately selected to represent factors previously identified as underlying fluency (Clough & Gordon, 2020; Gordon, 2020; Gordon & Clough, 2020, 2022). Additional measures (e.g., type–token ratio, content–function word ratio) were considered but ultimately omitted because they would require further manual coding that was determined to be not worth any potential gain in informativeness. Based on the prior studies referenced above, the dimensions were divided into three broad categories (or domains): GC (Grammatical Competence, reflecting grammatical complexity as well as accuracy), LA (Lexical Availability, encompassing efficiency of lexical retrieval and specificity of lexical items), and AF (Articulatory Facility, including accuracy of phonological formulation and effort of articulation). Some of the dimensions contribute to more than one domain in keeping with prior evidence as outlined below.

Grammatical Competence

Six dimensions were originally selected to represent GC: (a) proportion of utterances with embeddings, (b) MLU, (c) maximum utterance length (MaxLU), proportions of (d) agrammatic and (e) paragrammatic utterances, and (f) speech rate. The proportion of embedded clauses was a direct reflection of syntactic complexity, but this measure was ultimately excluded as there were too few embeddings to make it a meaningful measure. Utterance length measures were included as more indirect reflections of syntactic complexity. MLU has the advantage of being easy to calculate and was also the strongest contributor to ratings of GRAMMATICALITY in our survey study (Gordon

Downloaded from: https://pubs.asha.org University of Rhode Island on 10/10/2024, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

& Clough, 2022). However, MLU may underestimate the competence of a person with aphasia to occasionally produce more complex structures, so we also include MaxLU. Adapting procedures in the Aphasia Diagnostic Profiles (Helm-Estabrooks, 1992), MaxLU is calculated by averaging the three longest utterances in a sample.

Proportions of agrammatic and paragrammatic utterances were included to capture grammatical accuracy. Agrammatic utterances are characterized by the omission of one or more obligatory syntactic elements (Goodglass et al., 1993), whereas paragrammatic utterances involve the misuse or inappropriate use of grammatical elements (Butterworth & Howard, 1987; Matchin et al., 2020) through their substitution, addition, or misordering. Given the current lack of understanding regarding the nature of paragrammatism as a deficit (e.g., Matchin et al., 2020), we classify each utterance (rather than each individual PwA) as containing Agrammatism, Paragrammatism, or both. This approach is probably more reliable and reflects findings that a mixture of agrammatic and paragrammatic errors is common within a given individual (Butterworth & Howard, 1987; Gordon et al., 2022; Weisenberg & McBride, 1935). The differentiation of agrammatic and paragrammatic types of errors represents an additional degree of specificity over the catch-all code of [+gram] in AphasiaBank, promoting more accurate interpretation of the factors contributing to disruptions in narrative fluency.

The last grammatical measure was Speech Rate, measured in narrative words per minute (WpM). Speech rate may be affected by grammatical, lexical, and motor speech factors (Gordon & Clough, 2020), so we include it as a contributor to each of these broad domains. Narrative words refer to the core of words contributing to the semantic content and syntactic form, excluding, for example, side comments, repairs, neologisms, and perseverations (Saffran et al., 1989). Calculating rate using the number of narrative words provides an index of the quantity (though not necessarily the accuracy or relevance) of content that is conveyed. Measures of speech rate using sublexical units such as syllables are also relevant to fluency (e.g., Harmon et al., 2016), but are more likely to reflect lower levels of speech production (i.e., phonological formulation and articulation). In addition, counting syllables would require additional time-consuming coding.

Lexical Availability

Six dimensions were originally selected to represent LA: (a) Speech Rate; (b) MLU; and the proportions of utterances containing (c) fillers, (d) significant silent pauses (see below for definition), (e) repairs, and (f) empty speech. As noted above, Speech Rate was included because lexical measures were found by Gordon and Clough (2020) to be significant predictors of speech rate. Lexical measures were also shown to predict MLU, so MLU was included here. Fillers include both nonverbal fillers (e.g., *uh, um, hm*) and verbal fillers reflecting hesitations (e.g., *like, you know, well*), self-cueing (e.g., *J-K-L-M*, *mice*), asides (e.g., *or whatever her name is*), or hedges (e.g., *I guess*).

Significant pauses (Pauses) were defined as within-utterance pauses lasting approximately 1 s or more. From a practical perspective, 1 s has the advantage of being relatively easy to measure in a clinical setting by counting ("one, one thousand"). From a theoretical perspective, pauses within utterances are more likely to reflect microlinguistic planning (our focus here), such as word retrieval and grammatical formulation (Goldman-Eisler, 1958; Hartsuiker & Notebaert, 2010), whereas between-utterances pausing is more likely to reflect macrolinguistic planning, such as the recall of story concepts (Grande et al., 2012; Lee et al., 2019). Although reported pause times in spontaneous narrative speech vary widely, there seems to be widespread agreement that pauses of greater than 1 s are qualitatively distinct from pauses shorter than 1 s in a wide variety of contexts, such as quantifying production breakdowns in corpus data (e.g., Clark & Fox Tree, 2002), comparing neurologically healthy and impaired populations (e.g., Pistono et al., 2019; Sluis et al., 2020), and assessing fluency in second-language learners (e.g., De Jong, 2016; Shea & Leonard, 2019).

Repair behaviors (Repairs) include the following, as defined in AphasiaBank (MacWhinney et al., 2011): verbatim repetitions of words or phrases; retraces, in which the basic idea is repeated but the form of the sentence is modified; and reformulations, in which the message is changed. We consider these together, as they can be difficult to differentiate, and all of them may be traceable to word retrieval difficulties. At a population level, repairs are more likely to be associated with fluent than nonfluent aphasia (Casilio et al., 2019; Gordon, 2020; Gordon & Clough, 2020, 2022). However, this is somewhat misleading, as retracing shows positive associations with other indicators of fluency in nonfluent aphasia but negative associations in more fluent aphasia (Gordon, 2020). Thus, repairs are of clinical interest for at least two reasons: First, repairs have a significant impact on qualitative impressions of fluency, particularly for more fluent speakers (Feyereisen et al., 1991; Gordon & Clough, 2022); second, the ability to repair utterances may enhance fluency for more nonfluent speakers.

The sixth lexical dimension is the proportion of utterances containing empty speech or circumlocution (Empty Speech). Unlike most other manifestations of word retrieval difficulty, empty speech and circumlocution typically serve to facilitate, rather than disrupt, fluency (Clough & Gordon, 2020; Gordon & Clough, 2020). We include it here for its clinical utility in determining whether word retrieval difficulties are a prominent feature

Downloaded from: https://pubs.asha.org University of Rhode Island on 10/10/2024, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

of production for a given PwA, and whether fluency for that individual is maintained at the expense of content.

Articulatory Facility

Five dimensions were selected to represent the facility of articulation: (a) Speech Rate and (b) Pauses (as described above), the proportions of utterances containing (c) Phonological Errors and (d) Abstruse Neologisms, and (e) rated Effort. Although phonologically related errors are linguistic in nature, in practice, it is difficult to distinguish between errors of phonological origin and errors that originate from apraxia of speech (Haley & Jacks, 2023), so both phonological errors and abstruse neologisms are relevant here. Effort is coded at the utterance level to capture behaviors indicative of motor speech impairment, such as lengthened phonemes, phoneme distortions, syllable segmentation, and prosodic disruptions such as equal and excess stress or reduced stress contrast (Haley & Jacks, 2023; Kent & Rosenbek, 1983; McNeil et al., 2017). Like fluency itself, the impression of effort in production may be fostered by many factors, including the struggle to retrieve words and formulate grammatical utterances. However, effort is an important—if ill-defined—contributor to fluency disruptions in many speakers with aphasia. Our goal here was to define effort as originating from more peripheral aspects of production, including phonological formulation and articulatory implementation, in line with Goodglass et al.'s (2001a) conception of articulatory effort.

Calculation of Scores

Dimension Scores

In the Flu-ID, all dimensions except for Speech Rate, MLU, and MaxLU are calculated as a proportion of utterances containing each behavior. Each raw score is then converted to a dimension score so that all are on the same 5-point scale, and such that the more fluency is disrupted, the lower the dimension score (see Table 1 for an example). The ranges of proportions corresponding to each dimension score were chosen to represent the verbal descriptors of each score and thereby maximize face validity. For example, the extremes of the distribution (91%–100% and 0%–9%) include narrower ranges than the middle categories of the distribution, because they represent the verbal descriptors "always or almost always" and "never or almost never," respectively. The descriptor "sometimes" is represented by the widest range in the middle of the scale.

Table 1. Conversion of proportional measures to dimension scores, using the Agrammatism dimension as an example.

Frequency of fluency behavior	Proportion of utterances	Dimension score
Utterances are always or almost always agrammatic	91%–100%	1
Utterances are often or usually agrammatic	65%–90%	2
Utterances are sometimes agrammatic	35%–64%	3
Utterances are occasionally agrammatic	10%–34%	4
Utterances are never or almost never agrammatic	0%–9%	5

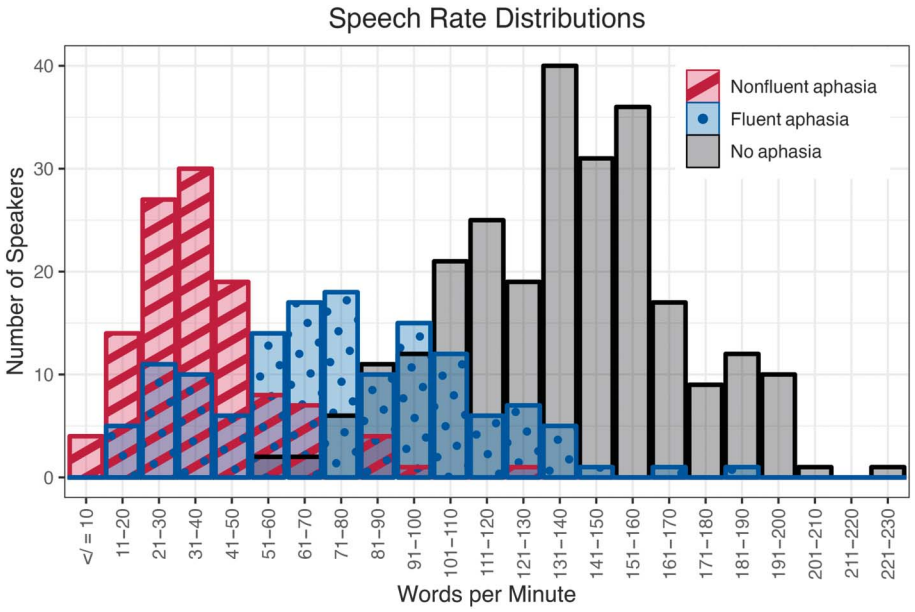
For Speech Rate, the conversion of raw scores to dimension scores was based on speech rate values calculated from 254 individuals with aphasia (115 nonfluent, 139 fluent as classified by the clinicians contributing the data to AphasiaBank; Gordon & Clough, 2020) as well as 255 individuals without aphasia in AphasiaBank, using the EVAL command in Computerized Language Analysis (CLAN; MacWhinney, 2000). The normative sample represents all unique individuals in the control database as of August 2022. By default, EVAL calculates speech rate excluding repetitions and revisions, which approximates the narrative speech rate measure used in the Flu-ID. Distributions of speech rates from these samples are shown in Figure 1. Individuals without aphasia had a median speech rate of 137 WpM, compared to median speech rates of 72 WpM for those with fluent aphasia and 34 WpM for those with nonfluent aphasia. The conversion based on these distributions is shown in Table 2.

As with Speech Rate, we compared the utterance length distributions for individuals with fluent and nonfluent aphasia subtypes (from Gordon & Clough, 2020), along with normative values from AphasiaBank (shown in Figure 2). Median utterance lengths were 4.0 words for nonfluent aphasia and 7.4 words for fluent aphasia, compared to 9.4 words for individuals without aphasia. The conversion of MLU and MaxLU measures to dimension scores was based on these distributions, as well as the types of syntactic structures typically corresponding to different utterance lengths, as illustrated in Table 3. An average of one to two words, reflecting a dearth of word combinations and a predominance of single nouns, corresponds to a score of 1.[2] A score of 2 reflects basic but mostly incomplete syntactic structures (combinations of two to four words, such as *article* + *noun*; *subject* + *verb*; *subject* + *verb* + *object*). A score of 3 reflects a restricted syntactic range (four to six words), corresponding to mostly complete sentences, but with a limited range of structures (e.g., *article* + *subject* + *verb* + *object*; *article* + *subject* + *auxiliary verb* + *main verb, article* + *subject* + *main verb* + *prepositional phrase*). Longer utterances, scored 4 and 5, reflect typical to extended ranges,

[2]The relative dearth of individuals with utterance lengths at this lowest level reflects the paucity of individuals with global aphasia in AphasiaBank.

Downloaded from: https://pubs.asha.org University of Rhode Island on 10/10/2024, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

Figure 1. Distribution of speech rates for speakers with and without aphasia in AphasiaBank.



respectively. Of course (as noted above), utterance length is a proxy measure for syntactic structure, but we observed that utterances of the lengths listed in Table 3 typically corresponded to such syntactic structures.

For the MaxLU measure (average of the three longest utterances), the ranges corresponding to the descriptions are extended and shifted upward relative to the ranges for MLU, such that fewer points are given for each category. The rationale for this is that a speaker's competence to produce syntactic structures (as reflected in the MaxLU dimension) is expected to be higher than their typical performance in producing syntactic structures (as reflected in MLU). This is certainly what we observed in the present sample of PwA, with a mean difference of about seven words between the two measures.

Domain Scores

For each of the three broad domains outlined above, domain scores are calculated by averaging the included dimension scores, with two exceptions. The frequency of Empty Speech is not included in the LA domain score, and the frequency of Paragrammatism is not included in the GC domain score. Although empty speech does reflect LA and paragrammatism reflects GC, their relationships to fluency are not the same as the other measures in their respective domains; that is, fluency tends to be facilitated rather than inhibited by empty or paragrammatic speech. Because the intent of providing domain scores is to allow clinicians to draw conclusions about which domains of language production contribute to fluency *disruption*, these dimensions are not included in the domain scores. However, they are retained as dimension scores to provide additional information about possible trade-offs of fluent speech production. An overall fluency domain score is also calculated by averaging the three domain scores. Because several of the variables are included in more than one domain (e.g., Speech Rate, MLU), this is essentially a weighted average that gives extra weight to these important dimensions.

Table 2. Conversion of speech rates measure to Speech Rate dimension scores.

Rate of speech relative to norms	Speech rate (words per minute)	Dimension score
Slow nonfluent range	0–20	1
Typical nonfluent range	21–40	2
Fast nonfluent/slow fluent range	41–60	3
Typical fluent/slow normal range	61–100	4
Typical normal to fast normal range	> 100	5

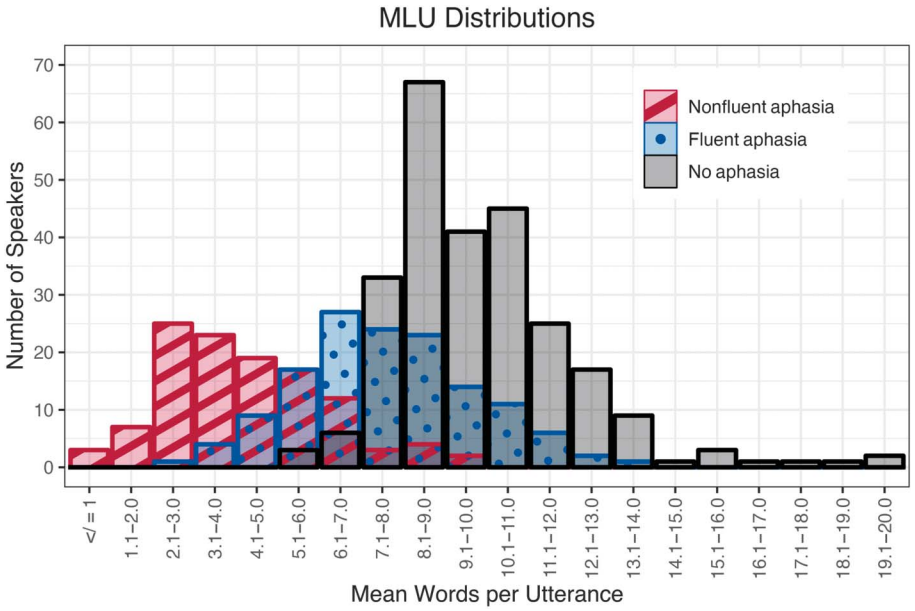
Testing the Flu-ID

Selecting the Sample of PwA

To assess the utility of the Flu-ID for a broad range of PwA, we made use of data collected from our survey of clinicians (Gordon & Clough, 2022), as described above. Respondents rated the fluency of each PwA on OVERALL FLUENCY, as well as seven subscales: SPEECH RATE, PAUSING, EFFORT, MELODY, PHRASE LENGTH, GRAMMATICALITY, and LEXICAL RETRIEVAL. Using these

Downloaded from: https://pubs.asha.org University of Rhode Island on 10/10/2024, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

Figure 2. Distribution of mean utterance lengths for speakers with and without aphasia in AphasiaBank. MLU = mean utterance length.



clinical ratings, we identified nine pairs of PwA with equivalent OVERALL FLUENCY ratings when averaged across coders (e.g., both PwA in the pair receiving an average overall fluency rating of 70 on the VAS), but who differed across subscale ratings, reflecting discrepant profiles of underlying impairments contributing to fluency. The average ratings of overall fluency and of the seven underlying fluency subscales from the previous study are illustrated in Figure 3 for each pair of PwA analyzed here (and in Supplemental Material S5 for all pairs of PwA shown together in the same graph). Demographic and aphasia-related information about the 18 PwA is also provided in Supplemental Material S6.

Our goal in selecting these individuals for analysis of the Flu-ID was to illustrate how overall fluency ratings can often mask clinically important differences and to help reveal the underlying causes of disfluency in each individual. For example, as shown in Figure 3, the OVERALL FLUENCY of two speakers (Pair 2, represented by medium blue lines) obtained very similar average OVERALL FLUENCY ratings of 57.5 and 57.4, but they diverged by more than 20 points on rated PAUSING (with PwA 2A rated higher) and by over 30 points on LEXICAL RETRIEVAL (with PwA 2B rater higher). Thus, it is clear that the factors contributing to perceived fluency are quite different in the two speakers, despite their equivalent OVERALL FLUENCY ratings.

The nine pairs of speakers were also selected to represent a range of fluency levels. As illustrated in Figure 3, their average OVERALL FLUENCY ratings (on the VAS of 0–100) ranged from 11.6 to 70.8 (Gordon & Clough, 2022). We did not consider type of aphasia when identifying pairs, but the resulting set of PwA consisted of a variety of taxonomic subtypes, according to the clinical diagnoses on AphasiaBank: seven individuals with anomic aphasia, five with Broca's aphasia, two with conduction aphasia, two with transcortical motor aphasia, and two with Wernicke's aphasia. This breakdown is similar to the

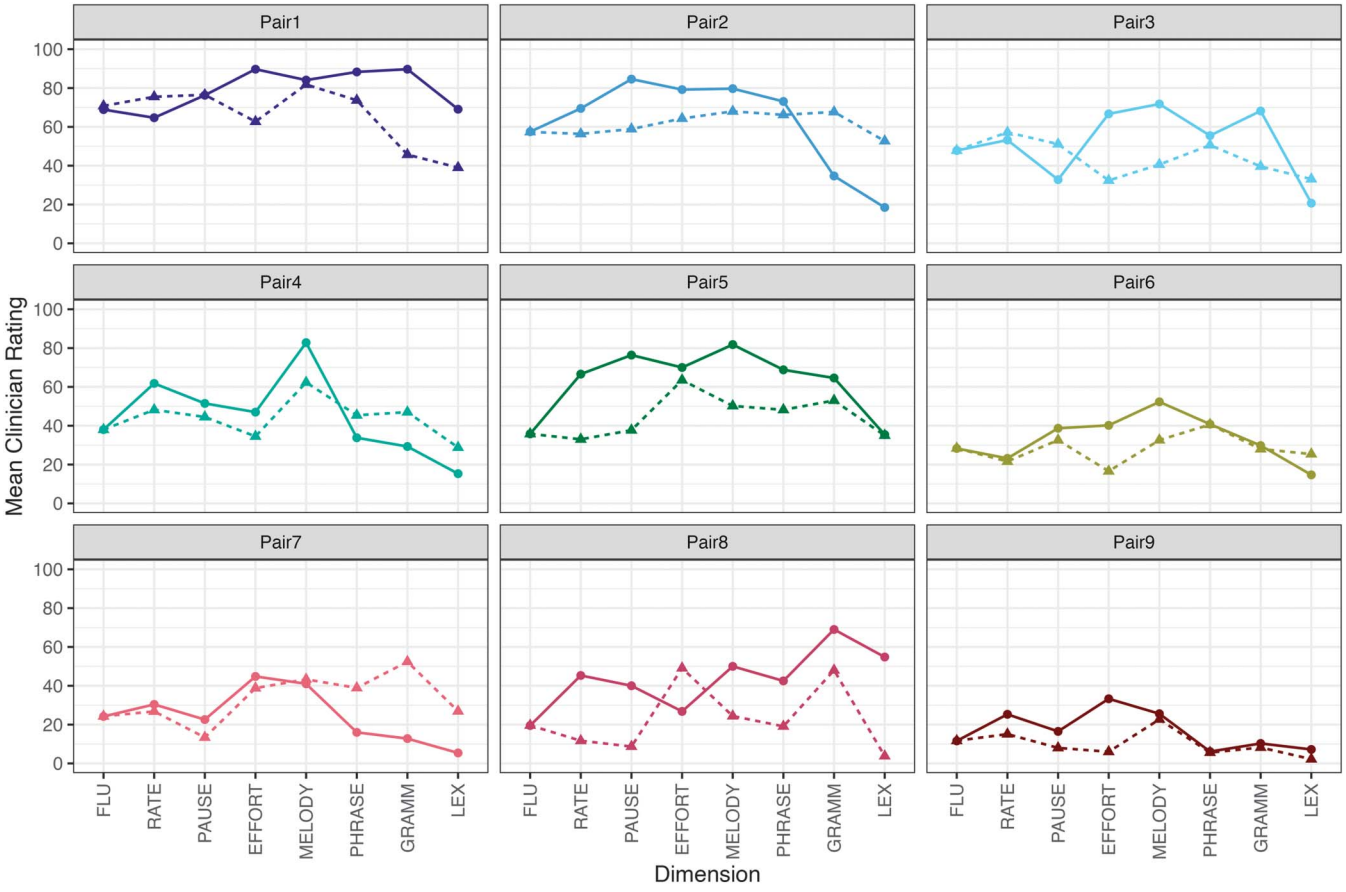
Table 3. Conversion of utterance length measures to dimension scores.

Typical syntactic structures	Utterance length (words) for MLU	Utterance length (words) for MaxLU	Dimension score
Minimal syntactic structure, if any	0–1.9	0–3.9	1
Basic syntactic structures	2–3.9	4–6.9	2
Restricted syntactic range	4–6.9	7–11.9	3
Typical syntactic range	7–10.9	12–17.9	4
Extended syntactic range	≥ 11	≥ 18	5

Note. MLU = mean utterance length; MaxLU = maximum utterance length.

Downloaded from: https://pubs.asha.org University of Rhode Island on 10/10/2024, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

Figure 3. Fluency ratings from Gordon and Clough (2022) for the nine pairs of speakers examined in the present study. All rating scales are oriented with higher scores indicating better or more fluent performance. FLU = overall fluency; RATE = rate of speech; PAUSE = degree of pausing; EFFORT = effortful speech; MELODY = melodic or intonational contour; PHRASE = phrase length; GRAMM = grammatical competence; LEX = lexical retrieval.



distribution of PwA in AphasiaBank, in that the sample includes a predominance of individuals with Broca's and anomic aphasia.

Assessing the Reliability of Coding

Each sample was coded independently by the two authors. We began with the transcribed samples from AphasiaBank, which were already segmented into utterances. AphasiaBank codes were then removed, and samples were timed by the first author (J.K.G.) using the time codes in AphasiaBank. Each coder then coded the fluency behaviors specified above (e.g., Pauses, Repairs, Phonological Errors, Effort) by watching the videotaped sample in AphasiaBank. The first 10 samples (Set A) were coded in three small batches of two to four samples each, with the coders checking interim reliability for each batch. At each checking stage, we discussed sources of disagreement and coding errors and made changes to the coding procedure as needed. For example, we began with a 10-point scoring scale but changed this to the 5-point scale described above, as we discovered that the 10-point scale reduced our reliability without a sufficient gain in information. It was also during this process that we decided to remove manual counting of content words and propositions as too time-consuming. The coding of syntactic embeddings was also eventually removed, as there was insufficient variability in this measure (i.e., embeddings were too infrequent) to provide meaningful information. Other changes involved elaborating on the guidelines used to code certain behaviors, particularly Effort and Empty Speech, which were frequent sources of disagreement.

Once the coding system was finalized, each coder separately recoded Set A in accordance with any changes made. At this point, our codes were not completely independent, as we had discussed each of the samples. However, during this initial discussion, we did not try to come to a consensus on any specific codes; rather, we aimed to keep the discussion at a general level to decide on coding principles. The recoding, then, might be

Downloaded from: https://pubs.asha.org University of Rhode Island on 10/10/2024, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

considered semi-independent. Next, each coder independently coded an additional eight PwA using the finalized protocol (Set B). To examine interrater reliability, we calculated point-to-point agreement for dimension scores and intraclass correlations (ICCs) for raw scores and dimension scores.

Assessing the Validity of Coding

To assess the construct validity of the Flu-ID, we examined the coherence of our three broad fluency domains by calculating intercorrelations among the dimensions contributing to each domain. Because the data were ordinal, we used Spearman rank-order correlations. To assess convergent validity, we compared the correspondence of our coding results (using averaged scores in the case of disagreements between coders) to the fluency profiles subjectively rated by clinicians in our prior study (Gordon & Clough, 2022), again using ranked dimension and domain scores.

Results: Reliability and Validity of the Flu-ID

Ranges of dimension scores and domain scores are illustrated in Figure 4. Relatively restricted ranges were noted for the dimensions of Paragrammatism, Pauses, and Abstruse Neologisms.

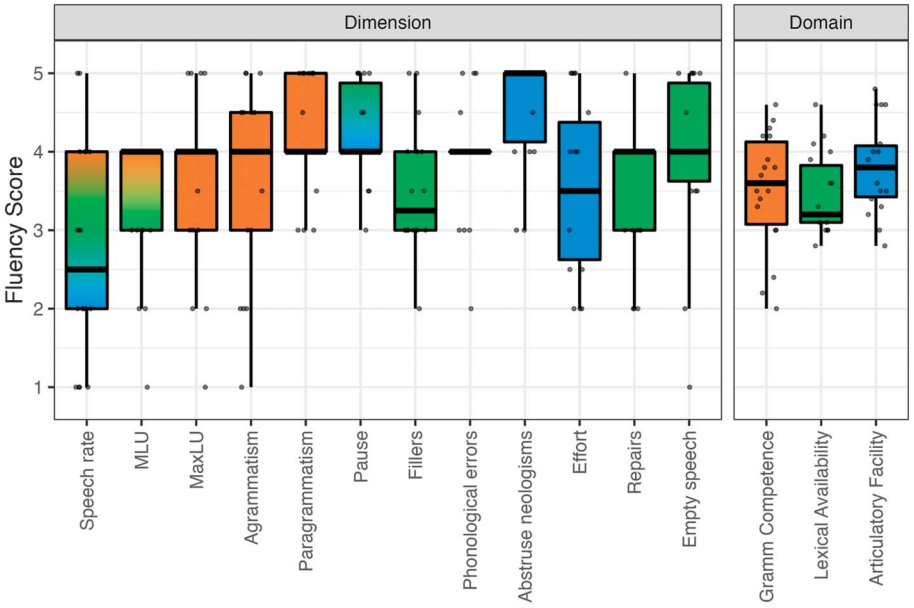
Interrater Reliability Analyses

To estimate interrater reliability, we calculated percent agreement using the *agree* function and ICCs using the *icc* function of the *irr* package (Gamer et al., 2019) in R. For ICCs, we used two-way random-effects models, of single-rater type (since this is how the measure would typically be used in the clinic), and with absolute agreement as the criterion. Benchmark criteria were used from Koo and Li (2016). We report percent agreement and ICC values for the total sample and each individual dimension in Table 4.

Percent Agreement

Percent agreement for the raw scores was expected to be low, as these values are on continuous scales; they are therefore not shown in Table 4. Across all 12 dimension scores (binned values from 1 to 5) for each of the 18 PwA (n = 216), the percent agreement between the two coders was 86.1%. The agreement level for Set A (86.7%) was only slightly higher than that for Set B (85.4%), indicating that the recoding process for Set A did not substantially inflate agreement levels. There were four dimensions that had agreement values below 80%. The most problematic, showing disagreements for half the sample, was the Effort dimension (discussed below). Although Agrammatism, Pausing, and Empty Speech had agreements ranging from 72% to 78%, many of these disagreements

Figure 4. Box-and-whisker plots for each fluency dimension and summary domain. The color of the box for each dimension represents the corresponding fluency domain (orange for Grammatical Competence, green for Lexical Availability, and blue for Articulatory Facility). Individual data points represent the mean of the two coders' fluency scores for each of the 18 people with aphasia in the reliability sample. Thick horizontal lines represent medians. Upper and lower boundaries of boxes represent third and first quartiles, respectively. Whiskers extend to the smallest and largest values no further than 1.5 times the interquartile range. MLU = mean utterance length; MaxLU = maximum utterance length.



Downloaded from: https://pubs.asha.org University of Rhode Island on 10/10/2024, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

Table 4. Percent agreement and intraclass correlations (ICCs) between the two coders for each fluency dimension for the sample of 18 people with aphasia.

4a. ICCs for raw scores				
Dimension	**% agree**	**ICC(A,1)**	**Lower CI**	**Upper CI**
Speech Rate	NA	1.00	.99	1.00
Mean Utterance Length	NA	1.00	.99	1.00
Max Utterance Length	NA	.99	.98	1.00
Agrammatism	NA	.98	.95	.99
Paragrammatism	NA	.92	.79	.97
Pausing	NA	.89	.73	.96
Fillers	NA	.95	.81	.98
Repairs	NA	.99	.96	.99
Empty Speech	NA	.89	.71	.96
Phonological Errors	NA	.97	.93	.99
Abstruse Neologisms	NA	1.00	.99	1.00
Effort	NA	.78	.45	.91
All dimensions	NA	.95	.93	.96
4b. Percent agreement and ICCs for dimension scores				
Dimension	**% agree**	**ICC(A,1)**	**Lower CI**	**Upper CI**
Speech Rate	100.0	1.00	1.00	1.00
Mean Utterance Length	100.0	1.00	1.00	1.00
Max Utterance Length	94.4	.98	.95	.99
Agrammatism	72.2	.91	.78	.97
Paragrammatism	88.9	.91	.77	.96
Pausing	77.8	.73	.41	.89
Fillers	83.3	.88	.70	.95
Repairs	100.0	1.00	1.00	1.00
Empty Speech	77.8	.91	.77	.96
Phonological Errors	94.4	.95	.89	.98
Abstruse Neologisms	94.4	.94	.86	.98
Effort	50.0	.70	.36	.88
All dimensions	86.1	.93	.91	.94

Note. All p values of intraclass correlations are < .001; CI = confidence interval; NA = not applicable.

corresponded to quite modest differences in raw scores, which is reflected in the ICC scores below.

ICCs

The ICC calculated across all raw scores was excellent ICC(A,1) = .95, 95% CI [.93, .96]. The ICC across all dimension scores was also excellent, ICC(A,1) = .93, 95% CI [.91, .94]. As illustrated in Table 4, all ICCs for raw scores (see Table 4a) were in the range of good (ICCs = .75–.90) to excellent (ICCs > .90) reliability. However, taking into account the lower bounds of the confidence intervals (CIs), Pauses and Empty Speech dropped to moderate reliability (ICCs = .50–.75) and Effort to poor reliability (ICCs < .50). For dimension scores (see Table 4b), all dimensions but two showed good to excellent reliability; Effort and Pauses showed moderate reliability. Taking into account the lower bound of the CIs, both of these dropped to poor reliability and Fillers dropped to moderate reliability.

Validity Analyses

Construct Validity: Intercorrelations Among Dimensions Within Domains

To assess construct validity, we examined the coherence of our domains using intercorrelation charts. Table 5 shows the Spearman rank-order correlations between each dimension score and each domain score and intercorrelations among the fluency dimensions within each domain. Examination of the table highlighted several potential problems with our domain scores.

First, within the GC domain, the four included dimension scores were all strongly associated with their domain score. As expected, Paragrammatism showed small negative relationships with GC and with Overall Fluency, supporting its exclusion from the GC domain. For the LA domain, the included dimensions all showed positive contributions to the domain, but Repairs did not

Downloaded from: https://pubs.asha.org University of Rhode Island on 10/10/2024, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

Table 5. Spearman rank-order correlations between fluency dimensions and fluency domains.

Fluency dimensions	Grammatical Competence	Lexical Availability	Articulatory Facility	Overall Fluency
Speech Rate	**.818*** **	**.852*** **	**.796*** **	.941***
MLU	**.754*** **	**.249**	.341	.555*
MaxLU	**.651** **	.362	.594**	.650**
Agrammatism	**.736*** **	.153	.522*	.554*
Paragrammatism	–.213	–.397	–.340	–.414
Pauses	.213	**.716*** **	**.402**	.431
Fillers	.043	**.440**	–.015	.045
Repairs	–.236	**.220**	–.073	–.155
Empty Speech	–.090	–.291	–.336	–.326
Phonological Errors	.435	.130	**.612** **	.471*
Abstruse Neologisms	–.049	–.367	**.127**	–.095
Effort	.672**	.345	**.858*** **	.710***
Fluency domains				
Grammatical Competence	1.000	.615**	.760***	.929***
Lexical Availability		1.000	.599**	.742***
Articulatory Facility			1.000	.897***
Overall Fluency				1.000

Note. Boldfaced values indicate which dimensions are included in which domain scores. MLU = mean utterance length; MaxLU = maximum utterance length.

*Values > .469 are significant at $p < .05$. **Values > .590 are significant at $p < .01$. ***Values > .709 are significant at $p < .001$.

appear to make a meaningful contribution to the LA domain score or to the Overall Fluency domain score. In addition, although Fillers made a meaningful contribution to the LA score, they did not correlate with Overall Fluency. Finally, the MLU score did not contribute as strongly as the MaxLU score, suggesting that the latter might be a better reflection of LA. Despite these apparently low contributions at a group level, we retained these dimensions in the LA domain to capture these potential sources of disfluency at an individual level. The AF domain showed meaningful contributions of all included scores with the exception of Abstruse Neologisms, which also did not show a significant correlation with Overall Fluency. The strongest contributions to the AF domain were Speech Rate, Effort, and Phonological Errors. Note that there were other significant relationships between scores in the GC domain and scores in the AF domain (notably Agrammatism and MaxLU with AF; Effort with GC), reflecting the common co-occurrence of grammatical and motor speech impairments in nonfluent aphasia.

Convergent Validity: Comparison With Clinical Ratings

To examine convergent validity, we compared profiles generated by the Flu-ID to the subjective ratings generated for these 18 PwA in our previous study (Gordon & Clough, 2022). Five of the eight ratings made by clinicians had direct correlates among our measures: SPEECH RATE, PAUSING, EFFORT, PHRASE LENGTH, and OVERALL FLUENCY. To compare to PHRASE LENGTH ratings, we averaged scores for MLU and MaxLU (Combined LU), as we suspect both contribute to subjective judgments of phrase length. To compare to the GRAMMATICALITY and LEXICAL RETRIEVAL ratings, we used our GC and LA domain scores. There was no corresponding measure for the MELODY rating, so we left this out of the analysis. Table 6 shows the correlations between clinician ratings and the corresponding ranked dimension/domain scores. These relationships are also illustrated using scatter plots in Supplemental Material S7.

All correlations were significant ($p < .05$) and met Cohen's (1988) criterion for large effects ($r > .50$). The

Table 6. Spearman correlations between ranked dimension/domain scores in the current study and corresponding Fluency subscale ratings by clinicians (Gordon & Clough, 2022; see Supplemental Material S7 for scatter plots).

Dimension/domain	Fluency subscale rating
Speech Rate	.909***
MLU	.645**
Pauses	.616**
Effort	.840***
Grammatical Competence	.691**
Lexical Availability	.531*
Overall Fluency	.909***

Note. MLU = mean utterance length.

*Values > .469 are significant at $p < .05$. **Values > .590 are significant at $p < .01$. ***Values > .709 are significant at $p < .001$.

Downloaded from: https://pubs.asha.org University of Rhode Island on 10/10/2024, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

strongest relationships were observed between the measurement and rating of Speech Rate (r = .91), between the composite measure of Overall Fluency and its corresponding rating (r = .91), and between the Effort dimension and its corresponding rating (r = .84). It is notable that the Effort dimension in the current study showed such a strong relationship with EFFORT ratings by clinicians, despite the relatively low reliability of these judgments discussed above. This is likely due to the fact that effort was subjectively judged in both studies, although by different judges. Despite individual differences in the ratings, ratings averaged across judges showed a strong correspondence, as we previously noted for clinician ratings (Gordon & Clough, 2022).

To further illustrate the similarities and differences between the subjective clinical ratings and the objective measures of the Flu-ID, we compared the profiles generated by each approach for two pairs of speakers. This comparison is illustrated and discussed in Supplemental Material S8.

Discussion: Clinical Utility and Further Development of the Flu-ID

In developing the Flu-ID, our goals were to ensure that the measures included were varied enough to capture the range of underlying deficits that might contribute to disruptions in fluency across individuals with aphasia, but specific enough to identify underlying impairments in a given individual. In addition, we aimed to develop a tool that improved the reliability of fluency measurement, while still being feasible to conduct in both research and clinical contexts. The value added by the Flu-ID is in the analysis rather than synthesis of individual contributors to fluency, unlike, for example, the Fluency scale in the WAB-R (Kertesz, 2006) or the use of a single proxy measure such as speech rate (Nozari & Faroqi-Shah, 2017) or MLU (Helm-Estabrooks, 1992). Unlike purely subjective rating scales as in the Auditory-Perceptual Rating of Connected Speech in Aphasia (APROCSA; Casilio et al., 2019) or the BDAE-3 (Goodglass et al., 2001a), we aim to facilitate the use of quantitative scores where possible. In addition, the Flu-ID is more specific in its purpose than other aphasia batteries, focusing on fluency of production, with less attention to the accuracy or specificity of meaning. (That said, counts of empty speech and phonological and grammatical errors are included, as they can help contribute to a broader understanding of the linguistic context contributing to the fluency profile.)

Reliability Analyses

To summarize the reliability findings, seven of the 12 dimension scores showed excellent reliability, with few disagreements and ICCs over .90. Another three dimensions showed moderate-to-good reliability, with either more frequent disagreements but high ICCs (Agrammatism, Empty Speech) or few disagreements but a lower ICC (Fillers). The remaining two dimensions—Effort and Pauses—showed poor-to-moderate reliability between coders. One drawback of binning raw scores into dimension scores was that, occasionally, a small difference in raw scores generated different dimension scores. For example, one coder identified 33.3% of the utterances of one PwA as containing Fillers, while the other coder identified 36.7%; because these numbers straddled the cutoff between two dimension scores, the first coder obtained a dimension score of 4 and the second obtained a dimension score of 3. For this reason, ICCs for raw scores tended to be slightly higher than those for dimension scores.

Notably, the Effort dimension was the only dimension that was perceptually rated rather than counted. Despite attempting to clarify what should be included and trying different rating methods (e.g., rating the whole sample rather than each utterance), our agreement remained low. Like fluency itself, "effort" is poorly defined and may be affected by difficulties in a wide range of underlying linguistic skills—constructing sentences, retrieving words, or formulating and implementing phonological and motor plans of connected speech. This ambiguity in distinguishing between contributors to production difficulty was not unexpected (e.g., see Haley & Jacks, 2023; Hybbinette et al., 2021; Wambaugh et al., 2019). However, despite our best efforts to focus on the phonological and articulatory aspects of production, our perceptions still frequently diverged. Related to this issue, the Flu-ID also lacks a specific motor speech measure. For some individuals, it may be useful to supplement the Flu-ID measurement with a motor speech assessment such as the Apraxia of Speech Rating Scale (Duffy et al., 2023; Strand et al., 2014). Despite these shortcomings, we decided to retain the Effort rating as part of the Flu-ID because of its importance to impressions of fluency. Nevertheless, we recommend caution in using Effort ratings to compare ratings by different clinicians.

One inevitable challenge is that reliability of some measures, particularly those related to accuracy, relies on inferences about the intended target utterance (Saffran et al., 1989). Such inferences are often difficult in interpreting aphasic speech, especially when output is severely affected. In particular, we found that the presence of abstruse neologisms, paragrammatic utterances, perseveration, and even empty speech often rendered the meaning and structure of utterances ambiguous and reduced reliability of coding. For this very reason, the authors of the QPA (Saffran et al., 1989, citing Menn, 1990, p. 338) cautioned against making such inferences. Still, the goal of the Flu-ID—to identify

Downloaded from: https://pubs.asha.org University of Rhode Island on 10/10/2024, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

impairments underlying fluency disruption—requires some degree of inference about target words and grammatical structures, and that introduces another source of variability among coders.

Validity Analyses

Our analysis of construct validity examined intercorrelations among dimensions within the GC, LA, AF, and Overall Fluency domain scores and demonstrated a lack of coherence in some domain scores. The goal of these domain scores is to provide a quantification of the extent to which underlying impairments in a given domain affect fluency. However, averaging the scores for different measures is an admittedly crude method of quantifying an impairment. Different measures may vary in the extent to which they affect fluency (or even whether they do), as suggested by our correlational analysis. One example of this is the finding that Paragrammatism had a negative relationship to the other grammatical dimensions and to Overall Fluency. Because of this, we decided to exclude Paragrammatism from the calculation of the GC domain score in the final version of the Flu-ID, for the same reasons that Empty Speech is excluded from the LA domain score.

We tested convergent validity by comparing Flu-ID scores to clinician ratings from our prior work (Gordon & Clough, 2022). The lower correlations for MLU, Pauses, and LA may be due to differences between the ratings in the work of Gordon and Clough (2022) and the measured scores used here. PHRASE LENGTH shows a wider distribution across the subjectively rated scale than Combined LU, which might reflect that the clinician raters weighted, for example, pausing more heavily than syntactic criteria in identifying utterances, resulting in the perception of shorter utterances for some PwA. Alternatively, the clustering of scores of 4 on the 5-point scale (and paucity of scores of 3) might suggest that the MLU ranges in the corresponding dimension scores should be adjusted. The lower correlation between pause measures may arise from our inclusion of only within-utterance pauses, whereas clinicians were probably attending to both within-utterance and between-utterances pauses when making their ratings. For the LA score, variability might be attributed to the wide range of behaviors that can signal word retrieval difficulty (pausing, repairs, empty speech, errors). Notably, our LA domain score focused on word retrieval difficulties that affect fluency, while the clinicians were not instructed to limit their lexical retrieval ratings in this way. For example, they may have based their judgments on the occurrence of empty speech, which was specifically excluded from our LA domain score.

Perceptual ratings of fluency are questionable as a gold standard, since they are susceptible to halo effects (Thorndike, 1920) across dimensions (i.e., the tendency for judgments on one dimension or characteristic to color judgments on other dimensions) and often have low reliability (Gordon, 1998). Thus, we find ourselves in the awkward situation of trying to interpret whether convergence or divergence with clinician ratings is desirable. For some measures (e.g., speech rate, MLU), it is clear that a quantification is preferable to a rating because the construct is inherently quantitative. For some (e.g., LEXICAL RETRIEVAL vs. Lexical Availability), divergence is understandable, because the ratings and dimension scores served different purposes, as discussed above. For other measures, the dimension scores may be missing something of importance to clinicians. One example may be the Pause dimension, which was intentionally restricted to within-utterance pauses exceeding a specific threshold. However, it is likely that between-utterances pauses also contributed to clinician ratings and perhaps should be counted in future versions of the Flu-ID. Another example is the Effort rating; despite its low reliability, it may not be possible to accurately and consistently quantify what is salient to listeners across different speakers.

Other Challenges and Recommendations

There are several limitations to the Flu-ID in its present form—some of these are necessary consequences of the aims to render fluency diagnosis more efficient and reliable; some are related to the lack of a current gold standard of fluency measurement; others are avenues for further development. First, transcription and coding are time-consuming and may not be feasible in some clinical contexts. On the other hand, in the interests of efficiency and improved reliability, the assessment involves some simplification and, consequently, a potential loss of information. However, this is defensible under the assumption that some simplification is necessary to achieve a much-needed gain in reliability. Ceiling effects were also noted in some of the measures. For example, the Flu-ID profile for PwA 2B is somewhat flattened relative to the clinician rating profile (see the figure in Supplemental Material S8).

Another aspect of simplification was that, to achieve efficiency, not all measures that might be informative are included. As noted above, content–function ratios and propositional density were excluded as too time-consuming to code. The Flu-ID also lacks a true measure of syntactic complexity, which has been shown to be highly predictive of fluency diagnoses across the spectrum of aphasia severity (Clough & Gordon, 2020; Gordon & Clough, 2020; Nozari & Faroqi-Shah, 2017). However, the measure of syntactic embedding we initially included was not sensitive enough to contribute useful information at an individual level. Other measures of grammatical complexity may provide more

Downloaded from: https://pubs.asha.org University of Rhode Island on 10/10/2024, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

sensitivity, such as the Developmental Sentence Score (Lee, 1974) that is built into CLAN (MacWhinney, 2000). With respect to our clinical feasibility goal, though, we determined that the additional manual coding required to count these structures would be too time-consuming. For more fluent PwA, syntactic embeddings could be coded fairly easily and included in the GC domain score; however, the measurement of fluency is generally less important for PwA who can already produce basic grammatical structures.

A drawback of collapsing several dimensions into one domain score (as with fluency itself) is that there can be many different overt manifestations of the same underlying impairment. For example, in the Flu-ID, a PwA who produces a variety of overt behaviors (e.g., pausing, fillers, repairs) to attempt to overcome lexical retrieval difficulty may be penalized more heavily than one who consistently produces fillers. Contributing to this is the simplification we adopted of counting each utterance containing one or more instances of a given behavior, rather than each instance of a given behavior. This decision means that we capture the proportion of utterances with reduced fluency, but not the extent to which each utterance is affected. It remains an open question whether counting instances of a given behavior is a more effective way of reflecting the degree of fluency disruption.

In its current form, we retain the domain scores in the Flu-ID, primarily as a conceptual cue to users that documenting surface manifestations is insufficient (Feyereisen et al., 1991). The clinician must go beyond overt behaviors to draw inferences about the underlying impairment for a given individual. However, because individual dimensions might dissociate, as discussed earlier, the summary chart generated by the Flu-ID contains scores for the individual dimensions rather than summary domains. If dimension scores within a given domain are consistently reduced, inferences about the underlying impairment may be quite straightforward. In other cases, additional evidence may be recommended. For example, inferences about Effort ratings being attributable to motor speech impairments can be supported by a motor speech assessment, as noted above. Inferences about Pauses reflecting word retrieval difficulties may be confirmed by observing omission errors or delayed responses on a confrontation naming test.

Comparing the Flu-ID to Automated Measures of Fluency

There have been other recent efforts to measure fluency in aphasia more objectively. Metu et al. (2023) compared the reliability of machine-learning algorithms to clinical judgments in distinguishing between fluent and nonfluent aphasia based on internet videos. Clinicians showed relatively poor agreement when using the WAB-R Fluency scale, better agreement with dichotomous judgments (fluent vs. nonfluent), and the highest agreement with trichotomous judgments (fluent vs. nonfluent vs. mixed). The accuracy of the machine-learning algorithms varied widely, depending on the benchmark to which they were compared, but were reported to be less accurate than the trichotomous classification by clinicians. One problem in interpreting these findings is that it is unknown how well the speakers with aphasia represented the range of fluency in aphasia; internet videos illustrating aphasia subtypes tend to reflect prototypically fluent and nonfluent types. Furthermore, no data were provided on how many of the video samples were judged by clinicians to belong to each category (e.g., how many were judged to have "mixed fluency").

Fontan et al. (2023) tested the ability of a signal-processing algorithm to predict clinical judgments of fluency on a continuous (5-point) scale. The algorithm detected energy envelopes in samples of read-aloud speech to automatically generate several measures related to the ratio of silence to speech. These low-level predictors accounted for a high proportion of the variance in clinician ratings, demonstrating that clinical perceptions of fluency are strongly driven by temporal factors such as speech rate. The dominance of more superficial measures such as speech rate and pausing is consistent with previous findings (e.g., Clough & Gordon, 2020; Gordon, 2020; Gordon & Clough, 2020, 2022), and is perhaps a more general feature of perceptual judgments of speech (e.g., see Gordon et al., 2019). As the authors acknowledged, however, this method did not account for other potential contributors to disfluency, such as repairs or grammatical errors, and it is unclear how well the models might perform with spontaneous, as opposed to read, speech.

Fromm et al. (2023) explored a semi-automated procedure within AphasiaBank to characterize fluency in aphasia. This approach begins, as does the Flu-ID, with transcripts that have been manually transcribed and coded, then applies the FLUCALC command (MacWhinney, 2000), generating a set of measures originally designed to capture dysfluency behaviors in children who stutter (Bernstein Ratner & MacWhinney, 2018). A subsequent principal components analysis (PCA), conducted on all Cinderella story samples in AphasiaBank from the aphasia and control databases, generated two principal components, one representing quantity and rate of speech (similar to what Gordon's [2020] factor analysis characterized as Narrative Productivity), and one representing fluency behaviors such as revisions and repetitions (comparable to Gordon's Repair factor). The small number of components (compared to six factors identified by Gordon) likely reflects the restricted focus of FLUCALC on measures of pausing and repetitions/repairs, which are most applicable to stutter-like dysfluencies. Grammatical and lexical sources of

Downloaded from: https://pubs.asha.org University of Rhode Island on 10/10/2024, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

dysfluency, which are more relevant to aphasia, are not represented. That said, the analysis of pauses in particular is more detailed, capturing both frequency and duration, and more objective, being derived directly from the audio sample, than in the Flu-ID. Optimally, a measure such as the Flu-ID, with its broader focus, could be integrated into the robust computational framework of AphasiaBank, enhancing its power and improving reliability of timing measures such as Pauses. In contrast to machine-learning methods, the transparent relationship between input and output in FLUCALC (and the Flu-ID) allows inferences about the factors contributing to fluency breakdown.

Such developments create exciting possibilities for speech and language analysis for both clinical and research purposes. However, the most important limitation of machine-learning and many other automated methods that are trained to match clinical fluency judgments is that they provide little information, if any, about the underlying deficits contributing to a breakdown in fluency. It is unknown whether machine-learning models can be trained to reliably identify the source of disruptions to fluency as assessed in the Flu-ID. In addition, automated prediction methods may perform well at a group level (e.g., Fontan et al., 2023) but are not yet sufficiently reliable and informative at an individual level to replace clinician ratings. For example, one outcome of the big-data approach conducted by Fromm and colleagues (over 500 samples) is that the PCA and cluster model approach generated relatively crude distinctions: between aphasia and no aphasia on one hand, and between more fluent and less fluent types of aphasia on the other. Finer-grained distinctions are needed for clinical use.

Conclusions

Understanding the underlying impairment is of critical importance to guiding therapy. A clinician's theory of therapy, including the hypothesized mechanism of action by which the therapeutic approach enacts change in the target (Hart et al., 2019), must be based on a hypothesis about what is causing the targeted behavior (or disruption of the targeted behavior). This process is particularly important in the treatment of aphasia for several reasons (Boyle et al., 2021): First, the nature of cognitive impairments is often difficult to infer on the basis of overt behaviors; second, those behaviors reflect the interaction of a number of linguistic and cognitive subskills, as well as poststroke communicative adaptations to those impairments; third, despite the long-standing history of aphasia therapy, the field has been more focused on the efficacy of therapy than its mechanisms. To facilitate this process in the assessment of fluency, the Flu-ID considers—but does not conflate—the many potential sources of dysfluency for a given individual with aphasia.

Just as word retrieval errors have helped inform models of lexical retrieval (e.g., Dell et al., 1997, 2004) and sentence production (e.g., Garrett, 1980), an enhanced understanding of fluency disruptions is also an important consideration for models of language production. For example, studies of typical adult speech suggest that different types of dysfluencies indicate difficulty at different levels, such as *um*s and *uh*s reflecting syntactic and lexical breakdown, respectively (Clark & Fox Tree, 2002). If the Flu-ID delivers on its aim of identifying underlying causes of fluency disruption in aphasia, assessment results can be used in more precise investigations of the nature of fluency breakdown across populations and its neurological substrates (Cordella et al., 2024). Ginzburg et al. (2014) likened dysfluencies to friction, arguing that "Some of the time it is useful to ignore the effects of friction, but the theory of motion is required to explicate the existence and quantitative effects of friction" (p. 10). Identifying the reasons for dysfluency is a critical first step to understanding the nature of the friction and thereby facilitating a smoother flow of language production in individuals with aphasia.

Data Availability Statement

The data sets generated and/or analyzed during the current study are available to qualified researchers from the corresponding author upon reasonable request.

Acknowledgments

This research was generously supported by a New Century Scholar grant from the American Speech-Language-Hearing Foundation awarded to both authors in 2017. The authors would also like to acknowledge the developers and contributors to AphasiaBank for providing such a rich and informative database for the study of aphasia.

References

Albert, M. L., Sparks, R. W., & Helm, N. A. (1973). Melodic Intonation Therapy for aphasia. *Archives of Neurology, 29*(2), 130–131. https://doi.org/10.1001/archneur.1973.00490260074018

Berndt, R. S., Wayland, S., Rochon, E., Saffran, E., & Schwartz, M. (2000). *Quantitative Production Analysis: A training manual for the analysis of aphasic sentence production*. Psychology Press.

Bernstein Ratner, N., & Brundage, S. B. (2022). *A clinician's complete guide to CLAN and Praat*. Carnegie Mellon University. https://aphasia.talkbank.org/

Bernstein Ratner, N., & MacWhinney, B. (2018). FluencyBank: A new resource for fluency research and practice. *Journal of Fluency Disorders, 56,* 69–80. https://doi.org/10.1016/j.jfludis.2018.03.002

Bose, A., Dutta, M., Dash, N. S., Nandi, R., Dutt, A., & Ahmed, S. (2022). Importance of task selection for connected speech

Downloaded from: https://pubs.asha.org University of Rhode Island on 10/10/2024, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

analysis in patients with Alzheimer's disease from an ethnically diverse sample. *Journal of Alzheimer's Disease, 87*(4), 1475–1481. https://doi.org/10.3233/JAD-220166

Boyle, M., Gordon, J. K., Harnish, S. M., Kiran, S., Martin, N., Rose, M. L., & Salis, C. (2021). Evaluating cognitive–linguistic approaches to interventions for aphasia within the Rehabilitation Treatment Specification System. *Archives of Physical Medicine & Rehabilitation, 103*(3), 590–598. https://doi.org/10.1016/j.apmr.2021.07.816

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

Butterworth, B., & Howard, D. (1987). Paragrammatisms. *Cognition, 26*(1), 1–37. https://doi.org/10.1016/0010-0277(87)90012-6

Carragher, M., Mok, Z., Steel, G., Conroy, P., Pettigrove, K., Rose, M. L., & Togher, L. (2023). Towards efficient, ecological assessment of interaction: A scoping review of co-constructed communication. *International Journal of Language & Communication Disorders, 59*(3), 831–875. https://doi.org/10.1111/1460-6984.12957

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. https://doi.org/10.1044/2018_AJSLP-18-0192

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

Clark, H. H., & Fox Tree, J. E. (2002). Using *uh* and *um* in spontaneous speaking. *Cognition, 84*(1), 73–111. https://doi.org/10.1016/S0010-0277(02)00017-3

Clark, H. H., & Wilkes-Gibb, D. (1986). Referring as a collaborative process. *Cognition, 22*(1), 1–39. https://doi.org/10.1016/0010-0277(86)90010-7

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.

Cordella, C., Di Filippo, L., Kolachalama, V. B., & Kiran, S. (2024). Connected speech fluency in poststroke and progressive aphasia: A scoping review of quantitative approaches and features. *American Journal of Speech-Language Pathology, 33*(4), 2091–2128. https://doi.org/10.1044/2024_AJSLP-23-00208

Curtis, S., Nicholas, M. L., Pittmann, R., & Zipse, L. (2020). Tap your hand if you feel the beat: Differential effects of tapping in Melodic Intonation Therapy. *Aphasiology, 34*(5), 580–602. https://doi.org/10.1080/02687038.2019.1621983

De Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching, 54*(2), 113–132. https://doi.org/10.1515/iral-2016-9993

Dell, G. S., Lawler, E. N., Harris, H. D., & Gordon, J. K. (2004). Models of errors of omission in aphasic naming. *Cognitive Neuropsychology, 21*(2–4), 125–145. https://doi.org/10.1080/02643290342000320

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review, 104*(4), 801–838. https://doi.org/10.1037/0033-295X.104.4.801

Doedens, W. J., & Meteyard, L. (2022). What is functional communication? A theoretical framework for real-world communication applied to aphasia rehabilitation. *Neuropsychology Review, 32*(4), 937–973. https://doi.org/10.1007/s11065-021-09531-2

Duffy, J. R., Martin, P. R., Clark, H. M., Utianski, R. L., Strand, E. A., Whitwell, J. L., & Josephs, K. A. (2023). The Apraxia of Speech Rating Scale: Reliability, validity, and utility. *American Journal of Speech-Language Pathology, 32*(2), 469–491. https://doi.org/10.1044/2022_AJSLP-22-00148

Fergadiotis, G., & Wright, H. H. (2011). Lexical diversity for adults with and without aphasia across discourse elicitation tasks. *Aphasiology, 25*(11), 1414–1430. https://doi.org/10.1080/02687038.2011.603898

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., 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

Fontan, L., Prince, T., Nowakowska, A., Sahraoui, H., & Martinez-Ferreiro, S. (2023). Automatically measuring speech fluency in people with aphasia: First achievements using read-speech data. *Aphasiology, 38*(5), 939–956. https://doi.org/10.1080/02687038.2023.2244728

Fridriksson, J., Basilakos, A., Boyle, M., Cherney, L. R., DeDe, G., Gordon, J. K., Harnish, S. M., Hoover, E. L., Hula, W. D., Pompon, R. H., Johnson, L. P., Kiran, S., Murray, L. L., Rose, M. L., Obermeyer, J., Salis, C., Walker, G. M., & Martin, N. (2021). Demystifying the complexity of aphasia treatment: Application of the Rehabilitation Treatment Specification Systemx. *Archives of Physical Medicine & Rehabilitation, 103*(3), 574–580. https://doi.org/10.1016/j.apmr.2021.08.025

Fridriksson, J., Hubbard, H. I., Hudspeth, S. G., Holland, A. L., Bonilha, L., Fromm, D., & Rorden, C. (2012). Speech entrainment enables patients with Broca's aphasia to produce fluent speech. *Journal of Neurology, 135*(12), 3815–3829. https://doi.org/10.1093/brain/aws301

Fromm, D., MacWhinney, B., Chern, S., Geng, Z., Kim, M., & Greenhouse, J. (2023, May 30–June 3). *Automated analysis of fluency behaviors in aphasia* [Paper presentation]. Clinical Aphasiology Conference, Atlantic City, NJ, USA.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). *irr: Various coefficients of interrater reliability and agreement* (R package Version 0.84.1). https://CRAN.R-project.org/package=irr

Garrett, M. F. (1980). Levels of processing in sentence production (Chapter 8). In B. Butterworth (Ed.), *Language production: Speech and talk* (Vol. 1, pp. 177–220). Academic Press.

Ginzburg, J., Fernandez, R., & Schlangen, D. (2014). Disfluencies as intra-utterance dialogue moves. *Semantics & Pragmatics, 7,* 1–64. https://doi.org/10.3765/sp.7.9

Goldman-Eisler, F. (1958). The predictability of words in context and the length of pauses in speech. *Language and Speech, 1*(3), 226–231. https://doi.org/10.1177/002383095800100308

Goodglass, H., Christiansen, J. A., & Gallagher, R. E. (1993). Comparison of morphology and syntax in free narrative and structured tests: Fluent vs. nonfluent aphasics. *Cortex, 29*(3), 377–407. https://doi.org/10.1016/S0010-9452(13)80250-X

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

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

Goodwin, C. (1979). The interactive construction of a sentence in natural conversation. In G. Psathas (Ed.), *Everyday language: Studies in ethnomethodology* (pp. 97–121). Irvington Publishers.

Downloaded from: https://pubs.asha.org University of Rhode Island on 10/10/2024, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

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., Andersen, K., Perez, G., & Finnegan, E. (2019). How old do you think I am? Speech-language predictors of perceived age and communicative competence. *Journal of Speech, Language, and Hearing Research, 62*(7), 2455–2472. https://doi.org/10.1044/2019_JSLHR-L-19-0025

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., & Clough, S. (2022). How do clinicians judge fluency in aphasia? *Journal of Speech, Language, and Hearing Research, 65*(4), 1521–1542. https://doi.org/10.1044/2021_JSLHR-21-00484

Gordon, J. K., Peters, E. L., Westbrook, K. D., Wickre, A. A., & Mansour, A. D. (2022). *Grammatical structures and errors in paragrammatism*. Academy of Aphasia

Grande, M., Meffert, E., Schoenberger, E., Jung, S., Frauenrath, T., Huber, W., Hussmann, K., Moormann, M., & Heim, S. (2012). From a concept to a word in a syntactically complete sentence: An fMRI study on spontaneous language production in an overt picture description task. *NeuroImage, 61*(3), 702–714. https://doi.org/10.1016/j.neuroimage.2012.03.087

Haley, K. L., & Jacks, A. (2023). Three-dimensional speech profiles in stroke aphasia and apraxia of speech. *American Journal of Speech-Language Pathology, 32*(4S), 1825–1834. https://doi.org/10.1044/2022_AJSLP-22-00170

Harmon, T. G., Jacks, A., Haley, K. L., & Faldowski, R. A. (2016). Listener perceptions of simulated fluent speech in nonfluent aphasia. *Aphasiology, 30*(8), 922–942. https://doi.org/10.1080/02687038.2015.1077925

Hart, T., Dijkers, M. P., Whyte, J., Turkstra, L. S., Zanca, J. M., Packel, A., Van Stan, J. H., Ferraro, M., & Chen, C. (2019). A theory-driven system for the specification of rehabilitation treatments. *Archives of Physical Medicine & Rehabilitation, 100*(1), 172–180. https://doi.org/10.1016/j.apmr.2018.09.109

Hartsuiker, R. J., & Notebaert, L. (2010). Lexical access problems lead to disfluencies in speech. *Experimental Psychology, 57*(3), 169–177. https://doi.org/10.1027/1618-3169/a000021

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

Hengst, J. A. (2020). *Understanding everyday communicative interactions: Introduction to situated discourse analysis for communication sciences and disorders*. Routledge. https://doi.org/10.4324/9781003034537

Hula, W., Donovan, N. J., Kendall, D. L., & Gonzalez Rothi, L. J. (2010). Item response theory analysis of the Western Aphasia Battery. *Aphasiology, 24*(11), 1326–1341. https://doi.org/10.1080/02687030903422502

Hybbinette, H., Ostberg, P., & Schalling, E. (2021). Intra- and interjudge reliability of the Apraxia of Speech Rating Scale in early stroke patients. *Journal of Communication Disorders, 89,* Article 106076. https://doi.org/10.1016/j.jcomdis.2020.106076

Kent, R. D., & Rosenbek, J. C. (1983). Acoustic patterns of apraxia of speech. *Journal of Speech and Hearing Research, 26*(2), 231–249. https://doi.org/10.1044/jshr.2602.231

Kershenbaum, A., Galassi, M., Shattuck-Hufnagel, S., Bachan, S., & Zipse, L. (2023). The effect of prosodic timing structure on unison production in people with aphasia. *American Journal of Speech-Language Pathology*. Advance online publication. https://doi.org/10.1044/2023_AJSLP-22-00304

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

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

Lee, J., Huber, J., Jenkins, J., & Fredrick, J. (2019). Language planning and pauses in story retell: Evidence from aging and Parkinson's disease. *Journal of Communication Disorders, 79,* 1–10. https://doi.org/10.1016/j.jcomdis.2019.02.004

Lee, L. (1974). *Developmental sentence analysis*. Northwestern University Press.

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

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

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

McNeil, M., Ballard, K. J., Duffy, J. R., & Wambaugh, J. L. (2017). Apraxia of speech theory, assessment, differential diagnosis, and treatment: Past, present, and future. In P. van Lieshout, B. Maassen, & H. Terband (Eds.), *Speech motor control in normal and disordered speech: Future developments in theory and methodology* (pp. 195–221). ASHA Press.

Metu, J., Kotha, V., & Hillis, A. E. (2023). Evaluating fluency in aphasia: Fluency scales, trichotomous judgements, or machine learning. *Aphasiology, 38*(1), 168–180. https://doi.org/10.1080/02687038.2023.2171261

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

Pistono, A., Pariente, J., Bezy, C., Lemesle, B., Le Men, J., & Jucla, M. (2019). What happens when nothing happens? An investigation of pauses as a compensatory mechanism in early Alzheimer's disease. *Neuropsychologia, 124,* 133–143. https://doi.org/10.1016/j.neuropsychologia.2018.12.018

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

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

Schnur, T. T., & Wang, S. (2023). Differences in connected speech outcomes across elicitation methods. *Aphasiology, 38*(5), 816–837. https://doi.org/10.1080/02687038.2023.2239509

Shea, C., & Leonard, K. (2019). Evaluating measures of pausing for second language fluency research. *The Canadian Modern Language Review, 75*(3), 216–235. https://doi.org/10.3138/cmlr.2018-0258

Sluis, R. A., Angus, D., Wiles, J., Back, A., Gibson, T. A., Liddle, J., Worthy, P., Copland, D. A., & Angwin, A. J. (2020). An automated approach to examining pausing in the speech of people with dementia. *American Journal of Alzheimer's Disease and Other Dementias, 35*. https://doi.org/10.1177/1533317520939773

Stark, B. (2019). A comparison of three discourse elicitation methods in aphasia and age-matched adults: Implications for language assessment and outcome. *American Journal of*

Downloaded from: https://pubs.asha.org University of Rhode Island on 10/10/2024, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

Speech-Language Pathology, 28(3), 1067–1083. https://doi.org/10.1044/2019_AJSLP-18-0265

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. https://doi.org/10.1016/j.jcomdis.2014.06.008

Tavakoli, P. (2016). Fluency in monologic and dialogic task performance: Challenges in defining and measuring L2 fluency. *International Review of Applied Linguistics in Language Teaching, 54*(2), 133–150. https://doi.org/10.1515/iral-2016-9994

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, May 20–24). *Reliability of rating spontaneous speech in the Western Aphasia Battery: Implications for classification* [Paper presentation]. Clinical Aphasiology Conference, Seabrook Island, SC, USA.

Wambaugh, J. L., Bailey, D. J., Mauszycki, S. C., & Bunker, L. D. (2019). Interrater reliability and concurrent validity for the Apraxia of Speech Rating Scale 3.0: Application with persons with acquired apraxia of speech and aphasia. *American Journal of Speech-Language Pathology, 28*(2S), 895–904. https://doi.org/10.1044/2018_AJSLP-MSC18-18-0099

Weisenberg, T., & McBride, K. E. (1935). *Aphasia*. Commonwealth Fund.

Zanca, J. M., Turkstra, L. S., Chen, C., Packel, A., Ferraro, M., Hart, T., Van Stan, J., Whyte, J., & Dijkers, M. P. (2019). Advancing rehabilitation practice through improved specification of interventions. *Archives of Physical Medicine & Rehabilitation, 100*(1), 164–171. https://doi.org/10.1016/j.apmr.2018.09.110

Downloaded from: https://pubs.asha.org University of Rhode Island on 10/10/2024, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions