Perceptual cues used by listeners to discriminate fluent from nonfluent narrative discourse

Hyejin Park[1], **Yvonne Rogalski**[2], **Amy D. Rodriguez**[3,4], **Zvinka Zlatar**[4], **Michelle Benjamin**[4], **Stacy Harnish**[1,5], **Jeffrey Bennett**[6], **John C. Rosenbek**[1,5], **Bruce Crosson**[4,5], and **Jamie Reilly**[1]

[1]Department of Speech, Language, & Hearing Sciences, University of Florida, Gainesville, FL, USA

[2]Department of Speech-Language Pathology and Audiology, Ithaca College, Ithaca, NY, USA

[3]Department of Neurology, University of Florida, Gainesville, FL, USA

[4]Department of Clinical and Health Psychology, University of Florida, Gainesville, FL, USA

[5]Brain Rehabilitation Research Center, Malcom Randall Veterans Affairs Memorial Hospital, Gainesville, FL, USA

[6]Department of Radiology, University of Florida College of Medicine, Gainesville, FL, USA

Abstract

Background—Language fluency is a common diagnostic marker for discriminating among aphasia subtypes and improving clinical inference about site of lesion. Nevertheless, fluency remains a subjective construct that is vulnerable to a number of potential sources of variability, both between and within raters. Moreover, this variability is compounded by distinct neurological aetiologies that shape the characteristics of a narrative speech sample. Previous research on fluency has focused on characteristics of a particular patient population. Less is known about the ways that raters spontaneously weigh different perceptual cues when listening to narrative speech samples derived from a heterogeneous sample of brain-damaged adults.

Aim—We examined the weighted contribution of a series of perceptual predictors that influence listeners' judgements of language fluency among a diverse sample of speakers. Our goal was to sample a range of narrative speech representing most fluent (i.e., healthy controls) to potentially least nonfluent (i.e., left inferior frontal lobe stroke).

Methods & Procedures—Three raters blind to patient diagnosis made forced choice judgements of fluency (i.e., fluent or nonfluent) for 61 pseudorandomly presented narrative speech samples elicited by the BDAE Cookie Theft picture. Samples were collected from a range of clinical populations, including patients with frontal and temporal lobe pathologies and non-brain-damaged speakers. We conducted a logistic regression analysis in which the dependent measure was the majority judgement of fluency for each speech sample (i.e., fluent or non-fluent). The statistical model contained five predictors: speech rate, syllable type token ratio, speech productivity, audible struggle, and filler ratio.

Outcomes & Results—This statistical model fit the data well, discriminating group membership (i.e., fluent or nonfluent) with 95.1% accuracy. The best step of the regression model

Address correspondence to: Hyejin Park MS, University of Florida, Department of Speech, Language, & Hearing Sciences, 336 Dauer Hall, P.O. Box 117420, Gainesville, Florida 32610, USA. hj.park@ufl.edu.

included the following predictors: speech rate, speech productivity, and audible struggle. Listeners were sensitive to different weightings of these predictors.

Conclusions—A small combination of perceptual variables can strongly discriminate whether a listener will assign a judgement of fluent versus nonfluent. We discuss implications for these findings and identify areas of potential future research towards further specifying the construct of fluency among adults with acquired speech and language disorders.

Keywords

Fluency; Perception; Listener judgement; Nonfluent aphasia

Aphasiology has employed a classic distinction between "fluent" and "nonfluent" verbal output as the basis for discriminating between aphasia subtypes (Broca, 1863; Wernicke, 1874; see also Eggert, 1977). Goodglass (1993) argued that fluency provides a useful first cut towards diagnostic classification. Others have emphasised the additional diagnostic clinical value by suggesting that language fluency can also provide a common behavioural target for treatment (Kerschensteiner, Poeck, & Brunner, 1972). Today there remains little agreement regarding how to operationally define language fluency (Feyereisen, Pillon, & De Partz, 1991). Consequently there is a great deal of latitude in how fluency is characterised both across and between disciplines. For example, Wagenaar, Snow, and Prins (1975) classified a fluent speaker as one who produces many words within the context of a syntactically complex narrative. Others have emphasised the importance of semantic rather than formal/syntactic structure as a primary factor contributing to fluent speech. Fillmore (1979), for example, defined fluency as the facility to speak in a coherent, creative, semantically dense, and contextually appropriate manner in order to "fill time with talk". Kreindler, Mihailescu, and Fradis (1980) presented an altogether different perspective, arguing that fluent speech is composed of many words produced over a long time period with a rapid rate. Finally, Goodglass and Kaplan (1983) defined language fluency as the ability to produce long, uninterrupted, and grammatically diverse runs of words that are easily articulated.

Construct validity of aphasic language fluency remains poor due to the high degree of variability in how researchers have historically weighted perceptual and linguistic features. Furthermore, a number of empirical differences exist across different studies, reflecting deviations in the use of terminology, scales for measurement, and derivation of values (e.g., speech rate calculation). The features that have differentiated fluent and nonfluent speech in previous studies include speech rate (e.g., Feyereisen, Verbeke-Dewitte, & Seron, 1986; Gordon, 2006; but see Wagenaar et al., 1975; Kreindler et al., 1980), proportion of words in sentences (e.g., Gordon, 2006), the number of words produced (e.g., Kreindler et al., 1980), phrase length (e.g., Goodglass, Quadfasel, & Timberlake, 1964), total speaking time (e.g., Feyereisen et al., 1986), speech initiation latency and overt struggle (e.g., Benson, 1967), and melodic line (e.g., Gordon, 1998).

A number of researchers have also pursued a psycholinguistic approach to narrative fluency. For example, Gordon (2006) found that the ratio of verbs to verbs+nouns discriminated fluent from nonfluent aphasic people. Other work in this vein examined the relative proportions of closed- to open-class words and surprisingly found that these ratios were not predictive of clinical judgements of fluency (Feyereisen et al., 1986). Finally, other investigators have examined various syntactic contributions to fluency. Goodglass and colleagues, for example, found that grammaticality of sentence structures strongly discriminated fluent from nonfluent speakers (Goodglass, Christiansen, & Gallagher, 1993).

SOME APPROACHES TOWARDS DECONSTRUCTING FLUENCY

The greatest obstacle towards agreement on a metric of narrative language fluency is the extreme variability introduced primarily by two sources: raters and patients. Gordon (1998) remarked that the consistent finding of poor inter-rater and inter-instrument reliability renders cross-study inference practically meaningless. Difficulties in reliability at the rater level are further compounded by variability both between and within patients. For over a century researchers have attempted to control these principal sources of variability in a variety of ways. In general, studies of fluency have tended to fall under two broad classifications with respect to experimental control. The first, and most frequent, strategy has involved control of patient variability with the aim of evaluating rater or instrument characteristics. Studies that have employed this approach have tended to examine the variability that emerges when a relatively large number of independent raters gauge a smaller, fixed corpus of patient samples. Exemplifying this approach were Holland, Fromm, and Swindell (1986) who asked a panel of 22 experienced clinicians to rate the narratives of one patient, and Trupe (1984) who examined concordance and reliability among 20 raters on scoring the spontaneous speech section of the Western Aphasia Battery (Kertesz, 1982). In a departure from this particular method, Gordon (1998) assumed a somewhat hybrid approach to the assessment of narrative fluency in aphasia, by simultaneously assessing rater ($n = 24$) and patient ($n = 10$) variability. Paradoxically, the most consistent finding across each of these studies was the *inconsistency* with which raters and instruments tend to classify patients along the fluency dimension.

A second, less-common, approach to fluency, has involved asking a relatively small number of expert raters to gauge a comparatively larger corpus of narrative samples. However, there are exceptions. Through factor analyses and multivariate statistical procedures, Wagenaar and colleagues (1975) examined prediction of fluency among a large sample of stroke aphasia patients ($n = 74$), finding that raters' dichotomous classifications of fluency could be predicted by two principal components: speech tempo and mean length of utterance (for recent MLU lesion correlation work see Borovsky, Saygin, Bates, & Dronkers, 2007). Vermeulen, Bastiaanse, and Van Wageningen (1989) pursued a similar factor reduction approach examining the intercorrelation between production components (e.g., syntax, articulatory struggle, phonological paraphasia) among a large sample of patients with stroke aphasia ($n = 121$).

Virtually all investigations of fluency have limited inference to a specific chronic aphasia aetiology (i.e., chronic left hemisphere stroke) rather than introducing the heterogeneity associated with progressive (e.g., Alzheimer's disease) or subclinical aphasia (e.g., traumatic brain injury). However, there are many reasons to extend this multivariate approach to a larger and more heterogeneous sample of patients and controls. Perhaps the most compelling reason involves ecological validity. That is, clinicians must often make judgements of fluency for a variety of different patients without comprehensive a priori knowledge of disease pathology and in many cases where pathologies are indeed mixed (e.g., stroke aphasia evolving to vascular dementia). Thus, since fluency is a diagnostic marker for many different conditions, there exists a need to evaluate fluency within the context of a diverse range of disorders.

Our aim here was to examine the weighted contribution of perceptual cues towards listeners' judgements of language fluency. More specifically we extended previous multivariate approaches to fluency within the context of a small number of experienced judges ($n = 3$) rating a relatively large number of language samples ($n = 61$). Through regression analyses, we analysed the weighted contribution of five specific features towards influencing the consensus fluency judgement for each language sample. These included measures

describing: (1) how quickly one speaks (i.e., speech rate); (2) the diversity of one's narrative in terms of segmental elements (i.e., syllable type token ratio); (3) how much one speaks during a given interval (i.e., speech productivity); (4) the amount of articulatory struggle evident during one's production (i.e., audible struggle); and (5) how often one pauses during narrative production (i.e., filled and unfilled pauses). We hypothesised that these five factors together would be strongly predictive of whether a listener would judge a narrative sample as representative of fluent or nonfluent discourse. Moreover, we hypothesised that listeners would be differentially sensitive to the weighting of these features and that such differences will be elucidated through simultaneous logistic regression analysis.

METHOD

Participants

Speakers—We retrospectively obtained spoken narrative samples from a range of patients and non-brain-damaged (NBD) older adults in an effort to capture substantial variance in the dimension of interest (i.e., fluency). Since this study aimed to identify which variables influence listeners' judgements of language fluency, regardless of aetiology, we included a heterogeneous sample expected to produce a wide variety of perceptual characteristics. This sample included 61 right-handed, native English-speaking adults. Relevant neuropsychological and demographic information appear in Table 1.

A variety of nonfluent aphasias have been associated with frontal lobe pathologies (Broca, 1863; Croot, Patterson, & Hodges, 1998; Crosson et al., 2007). Thus 26 speakers with aphasia secondary to unilateral left hemisphere stroke with predominant frontal lobe damage confirmed by structural magnetic resonance imaging or computed tomography were included in the sample to represent the more nonfluent end of the fluency continuum. We also included samples from five speakers with semantic dementia (SD), seven speakers with Alzheimer's disease (AD), one speaker with mixed dementia, one speaker with traumatic brain injury (TBI), and one speaker with posterior cerebral artery stroke that damaged a posterior and inferior region of the temporal lobe (e.g., fusiform gyrus) resulting in visual and naming disturbances but well-preserved speech production. These samples were intended to represent the variety of temporal lobe pathologies whose production has generally been described as fluent but empty, or in the case of these dementia variants, progressive fluent aphasia (Reilly & Peelle, 2008; Wernicke, 1874). Finally we also recruited 20 NBD older adults to represent normal language fluency.

Listeners—There were two groups of listeners. The first group was tasked with rating the amount of audible struggle perceived in each narrative. This group was composed of three licensed speech-language pathologists who also authored the paper (A.R., Y.R., and J.R.). We also employed a second listener group tasked with making dichotomous judgements of language fluency for each narrative sample. This second listener group was also composed of three licensed doctoral-level speech-language pathologists, each with more than 5 years experience working with adult neurogenic speech and language disorders. The second listener group was blind to study aims and also received no diagnostic clinical detail regarding the participant who generated each language sample.

Perceptual features: Coding and derivation

We identified five variables that potentially influence a listener's judgement of fluent or nonfluent production. These include factors such as speech rate, type–token ratio, and speech productivity that have wide precedence as predictors of fluency in past work (Borovsky et al., 2007; Gordon, 1998; Kerschensteiner et al., 1972). With this mix we also

analysed two variables whose effects are less clear: audible struggle and prevalence of filled pauses (i.e., filler ratios). We describe these variables in turn.

Speech rate—We defined speech rate as the number of discrete syllables produced per second. We derived speech rate by dividing the total number of syllables by the pure speaking time (i.e., total speaking time stripped of silences). Syllable productions included fillers, neologisms, paraphasias, and repetitions of syllables. We counted syllables using syllabification rules for Standard American English.

Syllable type token ratio (TTR)—Syllable TTR, a measure of the number of unique syllables a speaker produced, was calculated as a ratio of the total number of unique syllables to the total number of syllables. We used syllable TTR as opposed to word TTR because it is more sensitive to repetitions at the syllable level and because it captures production of neologisms. For example, if a speaker attempted to produce "mother" but began with repetitions of the first syllable "mo, mo, mother" we counted the utterance as four total syllables and two unique syllables. These numbers provide more information than word TTR, which would ignore "mo" because it is not a word. Calculating TTR in this way was important because we expected that non-fluent speakers would produce more repetitions and/or self-corrections at the syllable level.

Speech productivity—Speech productivity, which reflects the amount of time spent speaking, was calculated by dividing the pure speaking time without silences by the total speaking time with silences.

Audible struggle—Audible struggle represents the degree of vocal tension and articulatory effort apparent in a speaker's productions. Three certified speech-language pathologists listened to the complete digital speech samples and independently made perceptual judgements of audible struggle (1 = most struggle, 5 = least struggle). We then averaged these responses.

Filler ratio—Fillers, a feature of both fluent and nonfluent speech, represent the effort in continuing sentences. A filler ratio was calculated by dividing the total amount of time producing fillers by the pure speaking time. Fillers included word types (e.g., you know, like, etc.) and sound types (e.g., uh, um, etc.).

Procedure

Language samples—Speakers were asked to describe the Cookie Theft picture from the Boston Diagnostic Aphasia Examination (BDAE; Goodglass & Kaplan, 1983). The examiner did not interrupt the speakers during their descriptions; however, if a speaker provided only a brief response the examiner encouraged the speaker with the cue, "Do you have anything more to add?" Responses were digitally recorded in a quiet room. Collected language samples were used for dichotomous fluency judgements and analysed in the context of the five perceptual measures described above.

Data preparation—We transcribed the language samples orthographically and then calculated the total number of syllables and the total number of unique syllables using the Systematic Analysis of Language Transcripts (SALT; Miller & Iglesias, 2008). We used the Goldwave program (version 5.57) to segment the digital speech samples. "Total speaking time" was obtained by deleting the examiner's voice from a measure of time beginning immediately after the examiner gave a cue for initiating the picture description and ending after production of speaker's final word. From this total speaking time sample we omitted silences/pauses by deleting the portion of the language sample that represented less than

10% of the wave form amplitude, including environmental noise and non-speech sounds (e.g., sighing, laughing, coughing) to acquire a "pure speaking time". Fillers (e.g., "you know", "uh") were extracted from the original speech sample and placed consecutively in a new sound file in order to measure the length of time each speaker spent producing fillers.

Rating procedures—We asked three SLPs blind to patient diagnosis to listen to each of the 61 narrative speech samples presented in pseudorandom order. After listening to each sample, the SLPs made a dichotomous judgement of "fluent" or "nonfluent". We biased the listeners by informing them that their task was to judge each patient as fluent/nonfluent. We provided no additional constraining detail (e.g., listen for dysarthria).

Listeners' judgements were consistent with previous work demonstrating that language fluency is a highly variable construct vulnerable to individual bias. Absolute agreement (i.e., agreement among all three listeners) was obtained for 67.21% of the sample. In cases of disagreement, we accepted the majority judgement (i.e., agreement for two of three listeners). All of the NBD older adult samples were judged as fluent. The listeners classified 60% of the temporal lobe patients as fluent and 61% of the frontal lobe patients as nonfluent. In total 39 (64%) speakers were assigned fluent and 22 (36%) speakers were nonfluent.

Statistical design—Our aim was to examine the weighted contribution of a series of perceptual features on judgements of language fluency. The nature of the categorical dependent variable made binary logistic regression an ideal model for analysis of these data. The dependent variable was fluency judgement (fluent or nonfluent) as assessed by majority rating. The independent variables were the five perceptual variables. Prior to analysis we converted all individual raw values of variables those five variables into z-scores. This step allowed us to obtain standardised scores and match the scales of each variable. Then we used SPSS-17's forward stepping algorithm, entering all five predictors simultaneously without specifying an a priori order of variable entrance. SPSS then selects forward entrance of variables based on the strength of each variable's partial correlation. We then continued an iterative stepping procedure until attaining a negative R^2 change.

RESULTS

We first summarise characteristics of the individual predictors contrasted against the groups classified as fluent or nonfluent via majority consensus. Independent samples t-tests comparing the fluent and nonfluent samples were Bonferroni corrected for multiple comparisons. We follow with a description of the cumulative model, describing the weighted proportion of each predictor to the categorical outcome variable.

Speech rate

Speakers who were judged as fluent produced 3.71 (SD = .74) syllables per second, whereas speakers judged as nonfluent produced an average of 2.31 (SD = .89) syllables per second. This statistically significant difference in speech rate, $t(59) = 6.62$, $p < .001$, indicates that fluent speakers spoke more rapidly than nonfluent speakers. Figure 1 reflects differences as a function of group (stroke, other, or control).

Syllable type token ratio

Syllable TTR, a measure of syllabic diversity, failed to discriminate fluent from non-fluent speakers, $t(59) = 1.16$, $p = .25$. Fluent samples showed an average syllable TTR of .62 (SD = .14), whereas samples rates as nonfluent had an average TTR of .58 (SD = .10). See Figure 2.

Speech productivity

Speakers who were judged as fluent produced speech without pauses for 67.3% (SD = 17%) of their narrative time, whereas nonfluent speakers spoke without pauses for 37.43% (SD = 16%) of their narratives. This statistically significant difference, $t(59)$ = 6.87, $p < .001$, indicates that fluent speakers verbalised almost twice that of their nonfluent counterparts. See Figure 3.

Audible struggle

All controls were judged at ceiling on audible struggle on the 5-point Likert scale. The remainder of fluent speakers were also judged near ceiling (mean = 4.75, SD = .49). In contrast, nonfluent speakers were rated as showing significantly more audible struggle (mean = 3.44, SD = .99) relative to their fluent counterparts. This statistically significant difference in rated audible struggle, $t(27.03)$ = −5.83, $p < .001$, is reflected in Figure 4.

Filler ratio

The duration spent producing filled pauses did not discriminate fluent from nonfluent narrative samples, $t(59)$ = .36, $p = .71$. Speakers who were judged as fluent produced an average of 6.3% of their narratives producing filled pauses, whereas nonfluent speakers spent 5.7% of their narrative time producing filled pauses. This difference in filler ratios was not statistically significant (see Figure 5).

Cumulative model

Table 2 summarises results of the simultaneous logistic regression. The best-fitting model, obtained after three steps, explained 62% of the variance (R^2 = .62), and also showed strong discrimination of fluency status (i.e., 95.1%). Comparable to the interpretation of B-weights in a standard regression analysis, the observed differences in the standardised odds ratios suggests that listeners did not weigh all of the perceptual cues equally.

Listeners were 9.8 times more likely to judge speakers who spoke approximately more than three syllables-per-second as fluent compared to those speakers who spoke less than three syllables-per-second. Listeners were approximately 18 times more likely to judge narratives wherein speakers verbalised for more than 50% of the picture description as denoting a fluent sample. In addition, listeners were almost three times more likely to judge speakers who showed minimal audible struggle as fluent relative to narratives assigned struggle scores of 4.3 or lower on a 5-point scale. That is, listeners were the most sensitive to speech productivity in judging language fluency, then to speech rate and finally to audible struggle.

Two of the original five variables (i.e., syllable TTR and filler ratio) failed to discriminate fluent from nonfluent samples, as evidenced by their exclusion from the regression model (both $p > .05$). Table 2 summarises parameters at each stage of this model, and Figure 1 demonstrates the magnitude of these differences for each predictor.

Several predictors in this regression model are correlated. For example, when one produced many hesitations, that person was also likely to show a diminished syllable TTR. Highly correlated predictors potentially violate the multicollinearity assumption of regression. Table 3 represents a correlation matrix detailing the strength of the relationships among our predictors. Field (2009) suggests a threshold of .80 for potential multicollinearity violation. Although none of our predictors exceeded this threshold, the correlation between speech rate and audible struggle ($r = .74$) approached it. Thus, one must interpret the exclusion of audible struggle from the final model with caution.

GENERAL DISCUSSION

The purpose of this study was to examine which perceptual features of speech contributed most to listeners' perceptual judgements of language fluency. We have shown that as few as three variables can discriminate perceptual judgements of fluency with high accuracy. More specifically, listeners were sensitive to a weighted combination of how much patients said, how quickly they said it, and how much they struggled to say it. Other perceptual features we hypothesised would contribute to fluency (i.e., fillers and repeated syllables) were not significant predictors in our model.

With respect to these results, speech productivity was the most influential variable when listeners judged language fluency, which suggests that the absence of pauses, rather than increased speaking rate or less effortful speech, was a more influential factor in judging language fluency. Pausing is a common sign of word retrieval difficulty (Benson, 1967; Goldman-Eisler, 1963) and reflects a speaker's effort with respect to phonation, articulation, and error correction (Benson, 1967). Although pausing itself does not specify speech production problems, the absence of pause can be a clear sign of fluent speech (Benson, 1967).

Speech rate, the second most influential among our five features, has been identified as a significant variable affecting fluency judgement in many other studies (e.g., Benson, 1967; Gordon, 2006; Kerschensteiner et al., 1972; Wagenaar et al., 1975). However, previous studies (Benson, 1967; Gordon, 2006; Kerschensteiner et al., 1972; Kreindler et al., 1980) investigated speech rate at the *word* level, suggesting a cutoff of 50 words per minute for nonfluent and 90 words per minute for fluent speech (Kerschensteiner et al., 1972). On the other hand, our study explored speech rate at the syllable level, since distorted words and neologisms would potentially eliminate some of the utterances (Feyereisen et al., 1986). Based on the means for fluent and nonfluent speakers, our findings suggest a cut-off of approximately three syllables per second for fluent versus nonfluent. However, this assessment of speech rate in isolation must be interpreted with caution, as other parameters such as struggle and productivity also tended to co-vary with speech rate.

Audible struggle, previously described as an effort in initiation (Benson, 1967; Kerschensteiner et al., 1972) and in articulatory agility (Gordon, 1998), has been identified as a significant variable in fluency classification (Kerschensteiner et al., 1972). However, one issue that complicates the use of audible struggle in fluency classification is that its rating is often subjective, and studies have not been consistent in the use of rating scales. For example, a 3-point scale was used by Benson (1967) and Kerschensteiner et al. (1972), a 5-point scale was used in our study, and a 7-point scale was used by Gordon (1998). Additionally, subjective judgements like judging audible struggle are easily influenced by environmental variations (e.g., visual observation and listening vs. listening alone stimuli, etc.), which results in variability among listener ratings (Kent, 1996). Kent (1996) and Osberger (1992) also argued that perceptual judgements are influenced by listener's experience with disordered speech. Since experienced listeners are more familiar with atypical utterances, they are more likely to have strong intra- and inter-rater reliabilities than inexperienced listeners. Our raters were experienced clinicians who were accustomed to listening to speakers with neurological disorders. Hence they may have been better able to identify audible struggle as a possible variable influencing their fluency ratings.

Although audible struggle was included in the best model, it was not significant, leaving some question as to whether it should be considered a predictor of fluency. However, Figure 4 showed that audible struggle significantly differentiated the fluent speakers from the nonfluent speakers, suggesting that with more participants, audible struggle could emerge as

a significant variable contributing to perceptual judgements of language fluency. Further, audible struggle by itself in step 1 classified participants as fluent vs nonfluent with an 86.9% accuracy. It is also possible that audible struggle shares considerable variance with speech rate, since the predictive power of audible struggle diminishes below a significant level only after speech rate is added to the model. In fact this possibility is supported by the correlation results (Table 3). The correlation coefficient between audible struggle and speech rate is relatively high and significant; whereas others are below .5 or non-significant. Therefore, if audible struggle was eliminated from the final model due to its multicollinearity with speech rate, it may be feasible to consider audible struggle a predictor of fluency judgements.

Although "using many words" and "uninterrupted speech" have been considered essential features of fluent speech (Goodglass & Kaplan, 1983; Wagenaar et al., 1975), the syllable TTR and filler ratio features did not contribute to listeners' perceptual judgements of fluency. It is known that nonfluent aphasic speakers (e.g., Broca's aphasia) often struggle with word retrieval, phonation and/or articulation, resulting in numerous attempts at word production (Goodglass, 1993). Consequently repetition of the same syllables during these attempts reduces the proportion of unique syllables, and producing more fillers increases filler ratio. However, we found that syllable TTR and filler ratio did not influence perceptions of language fluency. Our correlation results revealed that syllable TTR was only significantly correlated with speech productivity but the correlation coefficient was not high. Further, filler ratio was not significantly correlated with any variable, suggesting that syllable TTR and filler ratio were not multicollinear. Unexpectedly the direction of association was negative.

We hypothesised that speech would be judged as fluent if it was fast, and if it was characterised by many unique syllables, and if it was without silences, fillers, or struggle. However, the association of syllable TTR with other variables indicates that speakers producing more unique syllables also produced more silences. Although it was not significant, syllable TTR also showed a negative association with audible struggle, indicating that speakers producing more unique syllables also demonstrated greater struggle. One possibility is that speakers paused and struggled in an attempt to retrieve more unique syllables. This effort reduced language fluency so that listeners judged the speech as nonfluent regardless of how many syllables a speaker produced. Another possibility is that, although calculating every syllable allowed us to include neologisms and incomplete words, it might have inflated the number of unique syllables. Moreover, it is possible that syllable TTR is influenced by a number of speech features (i.e., self-corrections, repetitions, usage of pronouns and/or synonyms, etc.), and these features may not all affect language fluency. In fact previous studies have reported that self-corrections were related to aphasic severity but did not differentiate types of aphasia (i.e., fluent vs nonfluent) (Farmer, 1977; Farmer, O'Connell, & O'Connell, 1978; Marshall & Tompkins, 1982). As such, combining these speech features into one variable might have washed out significant effects. With regard to filler ratios, a lack of perceptual distinction between fluent and nonfluent speech may not be entirely surprising given the high occurrence of fillers in normal daily discourse. Evidence suggests that fillers, such as "you know" or "uh", are in fact conventional English words used as a purposeful means of alerting the listener of upcoming delays (Clark & Fox Tree, 2002).

Future directions

It is important to reiterate that our work addressed the subjective perception of fluency. Individual perception is vulnerable to many sources of bias, including effects of habituation and expertise. Nevertheless, the current results show promise for further refining the

construct of perceived language fluency. However, we note many unanswered questions and several potential future avenues of research. These include the following:

What are the neural correlates of impairment within each of the sub-domains (e.g., rate, productivity) that moderate listeners' perceptions of fluency?— Clinicians commonly employ a heuristic regarding an anterior–posterior distinction between fluent and nonfluent aphasias. That is, frontal lobe pathologies tend to manifest nonfluent speech production, whereas temporal lobe pathologies produce more fluent speech (see also Kertesz, 1979, 1982, for discussions of Cortical Quotient). This fluency segregation is evident in the classification of neurodegenerative disorders such as Alzheimer's disease and semantic dementia as progressive fluent aphasia and a similar designation of "fluent aphasia" applied to Wernicke's and transcortical motor aphasia (Adlam et al., 2006, Caspari, 2005, LaPointe, 2005). Likewise, many heterogeneous frontal lobe pathologies tend to be categorised as "nonfluent" aphasias (e.g., primary progressive nonfluent aphasia, Broca's aphasia, transcortical motor aphasia (Mesulam, 2003).

Probabilistic categorisation of fluency as a gross clinical marker for frontal or temporal lobe pathology is fraught with pitfalls. For example, two neurogenic language disorders whose diagnostic criteria include fluent speech production are semantic dementia and Wernicke's aphasia (Neary et al., 1998; Goodglass & Kaplan, 1983). Yet narratives produced by patients with these disorders tend to show striking divergence. Complicating matters even further is the fact that conditions often categorised as "fluent" or "nonfluent" based on their canonical distributions of brain damage sometimes present with paradoxical output. That is, some patients with extensive damage to the left inferior frontal cortex predicted to be severely nonfluent present do not show frank symptoms of Broca's aphasia (Basso, Lecours, Moraschini, & Vanier, 1985). The reverse dissociation has also been reported, where patients with intact frontal cortex who are accordingly predicted to remain fluent instead present with speech that is markedly dysfluent (see Borovsky et al., 2007; Dronkers, 1996).

Heterogeneity both within and between patients lends further support to the claim that fluency is a multi-dimensional construct (Gordon, 1998, 2006). Lesion correlation studies have approached this issue through a decomposition approach to fluency. That is, researchers have attempted to correlate specific aspects of fluency (e.g., MLU) with regional brain damage as first evident through postmortem analyses and later through neuroimaging techniques (e.g., computed tomography). Today, techniques such as voxel-based lesion mapping (VLSM) can offer powerful new insights into the neural components of fluency. VLSM, in particular, allows researchers to correlate behaviours on a continuous scale with lesion distributions traced on MR images with millimetre-level spatial resolution (see Bates et al., 2003; Dronkers, Plaisant, Iba-Zizen, & Cabanis, 2007; Dronkers, Wilkins, van Valin, Redfern, & Jaeger, 2004). Such correlations will soon prove valuable towards understanding the degree to which variables that comprise fluency are in fact neuroanatomically (and behaviourally) dissociable. Borovsky and colleagues (2007) recently laid the groundwork towards this aim by examining lesion correlates of fluency in conversational speech among a large, heterogeneous sample of patients with left hemisphere stroke aphasia. These authors specifically investigated factors such as mean length of utterance in morphemes, TTR, and overall tokens spoken, and their relation to temporal relative to frontal lobe pathology. More recently we have also attempted to correlate the specific aspects of fluency examined here in a subset of stroke patients with regional distributions of brain damage, finding that each measure is associated with a unique distribution of damage, some in the lateral frontal lobe (e.g., speech rate) whereas others (e.g., syllable TTR) have a more posteromedial peak (Reilly et al., 2011). Thus lesion mapping coupled with behavioural (multivariate) analyses will potentially allow us to better explain clinical outliers and more generally improve construct validity.

What is the nature of the interaction between linguistic and perceptual variables towards influencing judgements of language fluency?—In the present study we examined a range of perceptual variables related to rate and productivity. However, it is distinctly possible that higher-level lexical factors also play a role in listeners' perceptions of language fluency. For example, the presence of neologisms, however fluent or well formed, may influence a listener's impression of the overall fluency of a speech sample. However, given the relatively low rate of neologistic responses we observed here, it is not possible to statistically parse this effect within the current sample.

What are the primary sources of listener variance in making judgements of language fluency?—Research on the psychology of expertise has repeatedly demonstrated that skilled practitioners perceive target elements in qualitatively different ways from unskilled observers. For example, a skilled mountain climber may be more likely to perceive a pattern of ascending footholds where the casual observer sees a sheer rock face. Such individual differences, coupled with other features such as habituation, create a number of sources of variability. It is critical for both practical and theoretical reasons to elucidate such sources of variance both within the rater himself/herself and between disciplines. A limitation of the current study, and indeed others that have employed relatively few expert raters, involves the potential for small sample bias and the amplification of habituation effects. One way that future studies can better control this source of error is to employ a hybrid approach similar to that of Gordon (1998) who analysed both rater and patient variability within diverse samples.

How does an observer's perception of narrative fluency shift as a function of the unique neurological disorder of the patient?—From a standpoint of ecological validity, there is value in analysing a heterogeneous sample of patients. That is, clinicians must often evaluate fluency when the underlying disorder is either unknown or mixed (e.g., stroke and vascular dementia). From a methodological standpoint, however, such heterogeneity introduces a number of potential confounds with respect to rater reliability and internal consistency. It is unclear the extent to which raters spontaneously shift their criteria for fluency as a function of the neurological disorder that compromises a particular narrative sample. For example, a rater might be more likely to "forgive" word-finding pauses as a dysfluency marker in Alzheimer's disease relative to chronic stroke aphasia. Thus a critical future aim of fluency research will involve assessment of rater stability and elucidation of the differential weighting of perceptual cues across distinct neurological disorders.

What are the sources of variance with respect to thematic content of narratives?—Previous research has investigated fluency from a variety of linguistic standpoints (e.g., syntactic, semantic, prosodic). The thematic content of one's narrative likely plays a significant but unclear role in the judgement of language fluency. For example, narratives that place increased demands on complex semantic material (e.g., describe the events leading up to WWII) may be judged altogether differently than cursory everyday conversation because more difficult contents would require more efforts on organising sequences and/or word retrieval and so forth, resulting in production of slow speech rate, more self-corrections/repetitions, and/or more fillers. Here we controlled for a number of these factors by contrasting exclusively Cookie Theft samples. However, one can make an obvious argument against the ecological validity of this picture description as a means for capturing a truly representative sample.

Development of a formal scale of language fluency—The development of a formal fluency scale for aphasia has great potential for clinical utility. Yet the limiting factor towards development of such a scale is the variability in agreement on the nature of fluency

itself. The establishment of a formal operational definition of *fluency* is long overdue but certainly within the realm of plausibility. The success of such an endeavour must necessarily involve a substantial degree of cross-disciplinary investigation towards the ultimate aim of a consensus on how to assess and deconstruct fluency.

Acknowledgments

References

Adlam A, Patterson K, Rogers TT, Nestor PJ, Salmond CH, Acosta-Cabronero J, et al. Semantic dementia and fluent primary progressive aphasia: Two sides of the same coin? Brain. 2006; 129:3066–3080. [PubMed: 17071925]

Basso A, Lecours A, Moraschini S, Vanier M. Anatomoclinical correlations of the aphasias as defined through computerized tomography: Exceptions. Brain and Language. 1985; 26(2):201–229. [PubMed: 2417656]

Bates E, Wilson SM, Saygin AP, Dick F, Sereno MI, Knight RT, et al. Voxel-based lesion–symptom mapping. Nature Neuroscience. 2003; 6(5):448–450.

Benson D. Fluency in aphasia: Correlation with radioactive scan localization. Cortex. 1967; 3:373–394.

Borovsky A, Saygin A, Bates E, Dronkers N. Lesion correlates of conversational speech production deficits. Neuropsychologia. 2007; 45(11):2525–2533. [PubMed: 17499317]

Broca P. Localisation des fonctions cérébrales: Siége du langage articulé. Bulletin of the Society of Anthroplogy (Paris). 1863; 4:200–203.

Caspari, I. Wernicke's aphasia. In: LaPointe, LL., editor. Aphasia and related neurogenic language disorders. 3. Stuttgart: Thieme; 2005. p. 142-168.

Clark H, Fox Tree J. Using uh and um in spontaneous speaking. Cognition. 2002; 84(1):73–111. [PubMed: 12062148]

Croot K, Patterson K, Hodges J. Single word production in nonfluent progressive aphasia. Brain and Language. 1998; 61(2):226–273. [PubMed: 9468772]

Crosson B, Fabrizio K, Singletary F, Cato M, Wierenga C, Parkinson R, et al. Treatment of naming in nonfluent aphasia through manipulation of intention and attention: A phase 1 comparison of two novel treatments. Journal of the International Neuropsychological Society. 2007; 13(4):582–594. [PubMed: 17521480]

Dronkers NF. A new brain region for coordinating speech articulation. Nature. 1996; 384(6605):159–161. [PubMed: 8906789]

Dronkers NF, Plaisant O, Iba-Zizen MT, Cabanis EA. Paul Broca's historic cases: high resolution MR imaging of the brains of Leborgne and Lelong. Brain. 2007; 130(Pt 5):1432–1441. [PubMed: 17405763]

Dronkers NF, Wilkins D, van Valin RD, Redfern B, Jaeger JJ. Lesion analysis of the brain areas involved in language comprehension. Cognition. 2004; 92:145–177. [PubMed: 15037129]

Eggert, GH., editor. Wernicke's works on aphasia: A source-book and review. The Hague; Mouton: 1977.

Farmer A. Self-correctional strategies in the conversational speech of aphasic and nonaphasic brain damaged adults. Cortex. 1977; 13(3):327–334. [PubMed: 923272]

Farmer A, O'Connell P, O'Connell E. Sound error self-correction in the conversational speech of nonfluent and fluent aphasics. Folia phoniatrica. 1978; 30(4):293–302. [PubMed: 730094]

Feyereisen P, Pillon A, De Partz M. On the measures of fluency in the assessment of spontaneous speech production by aphasic subjects. Aphasiology. 1991; 5(1):1–21.

Feyereisen P, Verbeke-Dewitte C, Seron X. On fluency measures in aphasic speech. Journal of Clinical and Experimental Neuropsychology. 1986; 8(4):393–404. [PubMed: 2427539]

Field, A. Discovering statistics using SPSS. 3. Los Angeles: Sage Publications Limited; 2009.

Fillmore, C. On fluency. New York: Academic Press; 1979.

Goldman-Eisler, F. Hesitation, information, and levels of speech production. In: De Reuck, AVS.; O'Connor, M., editors. Disorders of language. London: CIBA Foundation Sumposium; 1963. p. 96-111.

Goodglass, H. Understanding aphasia. San Diego, CA: Academic Press, Inc; 1993.

Goodglass H, Christiansen J, Gallagher R. Comparison of morphology and syntax in free narrative and structured tests: Fluent vs nonfluent aphasics. Cortex. 1993; 29:377–403. [PubMed: 8258282]

Goodglass, H.; Kaplan, E. The assessment of aphasia and related disorders. 2. Philadelphia: Lea & Febiger Philadelphia; 1983.

Goodglass H, Quadfasel F, Timberlake W. Phrase length and the type and severity of aphaisa. Cortex. 1964; 1:133–153.

Gordon J. The fluency dimension in aphasia. Aphasiology. 1998; 12(7):673–688.

Gordon J. A quantitative production analysis of picture description. Aphasiology. 2006; 20(2):188–204.

Holland AL, Fromm D, Swindell CS. The labeling problem in aphasia: An illustrative case. Journal of Speech & Hearing Disorders. 1986; 51(2):176–180. [PubMed: 3702366]

Kent R. Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders. American Journal of Speech-Language Pathology. 1996; 5(3):7–23.

Kerschensteiner M, Poeck K, Brunner E. The fluency–nonfluency dimension in the classification of aphasic speech. Cortex. 1972; 8(2):233–247. [PubMed: 5043796]

Kertesz, A. Aphasia and associated disorders: Taxonomy, localization, and recovery. New York: Grune & Stratton; 1979.

Kertesz, A. Western Aphasia Battery. New York: Grune & Stratton; 1982.

Kreindler A, Mihailescu L, Fradis A. Speech fluency in aphasics. Brain and language. 1980; 9(2):199–205. [PubMed: 7363064]

LaPointe, LL. Aphasia and related neurogenic language disorders. 3. New York: Thieme New York; 2005.

Marshall R, Tompkins C. Verbal self-correction behaviours of fluent and nonfluent aphasic subjects. Brain and Language. 1982; 15(2):292–306. [PubMed: 7074346]

Mesulam MM. Primary progressive aphasia – a language-based dementia. New England Journal of Medicine. 2003; 349(16):1535–1542. [PubMed: 14561797]

Neary D, Snowden JS, Gustafson L, Passant U, Stuss D, Black S, et al. Frontotemporal lobar degeneration: A consensus on clinical diagnostic criteria. Neurology. 1998; 51(6):1546–1554. [PubMed: 9855500]

Miller, J.; Iglesias, A. Systematic Analysis of Language Transcripts (SALT), student version 2008 [computer software]. SALT Software, LLC; 2008.

Osberger, M. Speech intelligibility in the hearing impaired: Research and clinical implications. In: Kent, R., editor. Intelligibility in speech disorders: Theory, measurement, and management. Philadelphia: Jon Benjamins Publishing Co; 1992. p. 233-264.

Reilly, J.; Park, H.; Bennett, J.; Towler, S.; Benjamin, M.; Crosson, B. Broca and beyond: Decomposing the lesion correlates of speech fluency in aphasia. 2011. Manuscript in preparation

Reilly J, Peelle J. Effects of semantic impairment on language processing in semantic dementia. Seminars in Speech and Language. 2008; 29(1):32–43. [PubMed: 18348090]

Trupe, EH. Reliability of rating spontaneous speech in the Western Aphasia Battery: Implications for classification. Paper presented at the 14th Annual Clinical Aphasiology Conference; Seabrook Island, SC. 1984.

Vermeulen J, Bastiaanse R, Van Wageningen B. Spontaneous speech in aphasia: A correlational study. Brain and Language. 1989; 36(2):252–274. [PubMed: 2465808]

Wagenaar E, Snow C, Prins R. Spontaneous speech of aphasic patients: A psycholinguistic analysis. Brain and language. 1975; 2:281–303. [PubMed: 1182498]

Wernicke, C. Der aphasische symptomencomplex: Eine psychologische Studie auf anatomischer basis. Breslau: Cohn & Weigert; 1874.

Figure 1.
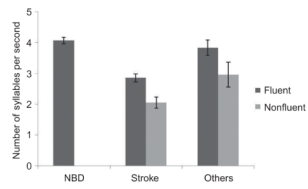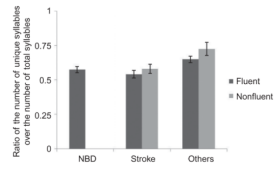Speech rate in syllables per second, ranged 0 to 5.

Figure 2.
Syllable type token ratio: ratio of unique syllables to total syllables, ranged 0 to 1.
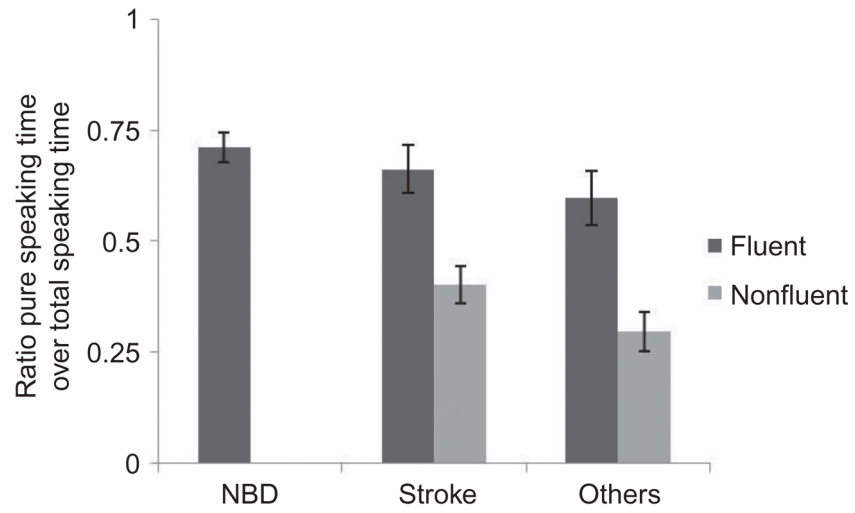
Figure 3.
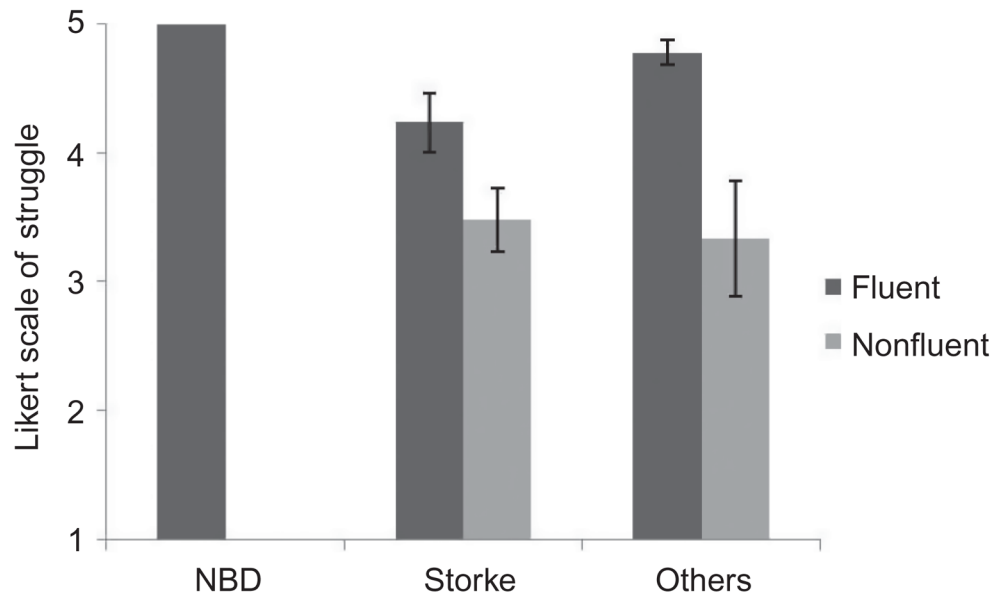Speech productivity: ratio of pure speaking time to total speaking time, ranged 0 to 1.

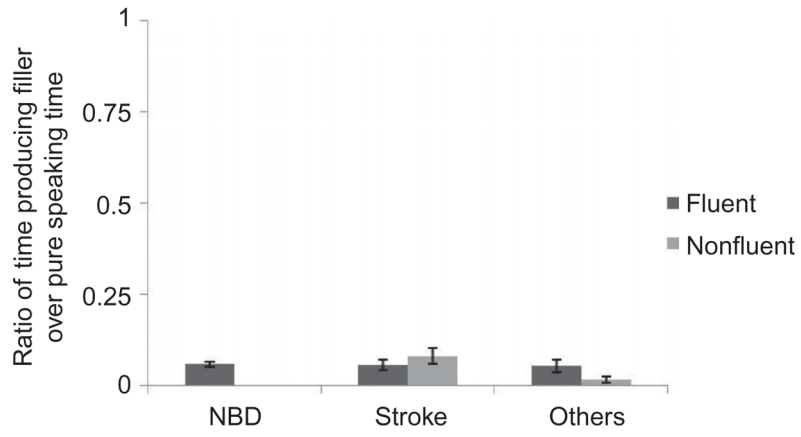**Figure 4.**
Audible struggle: Likert scale rating, ranged 1 to 5.

Figure 5.
Filler ratio: ratio of filled pause time to total speaking time, ranged 0 to 1.

TABLE 1

Demographic information

Participant number	Age	Sex	Years of education	Diagnosis	Consensus rated fluency
01	73	F	14	Stroke	Fluent
02	65	F	15	Stroke	Fluent
03	51	M	16	Stroke	Nonfluent
04	62	M	15	Stroke	Fluent
05	57	M	14	Stroke	Fluent
06	59	F	18	Stroke	Fluent
07	79	F	12	Stroke	Nonfluent
08	52	M	12	Stroke	Nonfluent
09	92	F	12	Stroke	Nonfluent
10	53	M	12	Stroke	Nonfluent
11	73	M	14	Stroke	Nonfluent
12	62	F	18	Stroke	Nonfluent
13	82	F	14	Stroke	Nonfluent
14	55	M	14	Stroke	Nonfluent
15	82	F	12	Stroke	Fluent
16	69	M	12	Stroke	Nonfluent
17	68	M	14	Stroke	Fluent
18	63	F	18	Stroke	Fluent
19	50	M	14	Stroke	Nonfluent
20	61	M	12	Stroke	Nonfluent
21	68	F	16	Stroke	Nonfluent
22	59	M	14	Stroke	Nonfluent
23	42	F	12	Stroke	Fluent
24	66	M	12	Stroke	Fluent
25	64	M	12	Stroke	Nonfluent
26	81	M	12	Stroke	Nonfluent
27	54	F	–	SD	Fluent
28	51	M	12	TBI	Nonfluent
29	58	M	12	AD	Fluent

Participant number	Age	Sex	Years of education	Diagnosis	Consensus rated fluency
30	74	M	–	PCA Stroke	Nonfluent
31	58	F	–	SD	Fluent
32	62	F	18	SD	Fluent
33	78	F	–	SD	Fluent
34	73	M	–	AD	Fluent
35	50	M	–	SD	Fluent
36	59	F	12	AD	Nonfluent
37	72	M	18	AD	Fluent
38	70	F	–	AD	Nonfluent
39	59	F	–	AD	Fluent
40	74	F	12	Mixed dementia	Nonfluent
41	69	F	–	AD	Nonfluent
42	78	F	18	NBD	Fluent
43	72	F	18	NBD	Fluent
44	71	M	14	NBD	Fluent
45	69	F	13	NBD	Fluent
46	74	M	12	NBD	Fluent
47	71	F	13	NBD	Fluent
48	73	F	12	NBD	Fluent
49	74	M	10	NBD	Fluent
50	77	F	17	NBD	Fluent
51	73	M	16	NBD	Fluent
52	78	F	15	NBD	Fluent
53	78	M	18	NBD	Fluent
54	81	F	14	NBD	Fluent
55	73	M	18	NBD	Fluent
56	55	M	18	NBD	Fluent
57	63	F	16	NBD	Fluent
58	70	F	15	NBD	Fluent
59	64	F	–	NBD	Fluent
60	60	F	–	NBD	Fluent

Participant number	Age	Sex	Years of education	Diagnosis	Consensus rated fluency
61	61	F	–	NBD	Fluent
Total	66.62 years	M = 28F = 33	14.31 years		FL = 39NF = 22

M = Male; F= Female; *SD* = Semantic dementia; TBI = Traumatic brain injury; AD = Alzheimer's disease; NDB = Non-brain-damaged control; FL = Fluent; NF = Nonfluent.

TABLE 2

Model summary and variable coefficients by step

Included variables	Model summary			Variable coefficients			
	Cox & Snell R²	−2 LL	Accuracy of classification	B	SE	Odds ratio	p
Step 1	.40	48.47***	86.9%				
Audible struggle				2.09	.53	8.06	.000
Step 2	.57	27.65***	93.4%				
Speech productivity				2.52	.81	12.37	.002
Audible struggle				2.01	.66	7.46	.002
Step 3	.62	21.08**	95.1%				
Speech rate				2.28	1.13	9.80	.043
Speech productivity				2.87	.94	17.70	.002
Audible struggle				.97	.81	2.63	.230
Step 4	.61	22.74	93.4%				
Speech rate				3.01	1.00	20.30	.003
Speech productivity				2.94	.93	18.98	.002

Omnibus test of changes of −2LL (−2 Log likelihood: lower value indicates better model fit) from the previous model was presented with *p*-value:

**
  p < .01,

***
  p < .001.

The output of the logistic regression is interpretable in terms of an odds ratio. This is an indicator of the change in odds resulting from a unit change in the predictor. It is similar to the *b*-coefficient in logistic regression.

TABLE 3

Correlation matrix for perceptual predictors of fluency

	Speech rate	Syllable TTR	Speech productivity	Audible struggle	Filler ratio
Speech Rate	1	.05	.36**	.74***	−.23
Syllable TTR		1	−.39**	−.09	−.17
Speech Productivity			1	.41**	−.14
Audible Struggle				1	.04

Correlation is significant

**
p < .01;

***
p < .001.