Published in final edited form as:
Aphasiology. 2011 ; 25(12): 1600–1620. doi:10.1080/02687038.2011.628379.

Computer-Mediated Assessment of Intelligibility in Aphasia and Apraxia of Speech

Katarina L. Haley[1], **Heidi Roth**[2], **Enetta Grindstaff**[1], and **Adam Jacks**[1]

[1]Division of Speech and Hearing Sciences, CB#7190; Department of Allied Health Sciences; The University of North Carolina at Chapel Hill, Chapel Hill NC 27599

[2]Department of Neurology, CB #7025; The University of North Carolina at Chapel Hill, Chapel Hill NC 27599

Abstract

Background—Previous work indicates that single word intelligibility tests developed for dysarthria are sensitive to segmental production errors in aphasic individuals with and without apraxia of speech. However, potential listener learning effects and difficulties adapting elicitation procedures to coexisting language impairments limit their applicability to left hemisphere stroke survivors.

Aims—The main purpose of this study was to examine basic psychometric properties for a new monosyllabic intelligibility test developed for individuals with aphasia and/or AOS. A related purpose was to examine clinical feasibility and potential to standardize a computer-mediated administration approach.

Methods & Procedures—A 600-item monosyllabic single word intelligibility test was constructed by assembling sets of phonetically similar words. Custom software was used to select 50 target words from this test in a pseudo-random fashion and to elicit and record production of these words by 23 speakers with aphasia and 20 neurologically healthy participants. To evaluate test-retest reliability, two identical sets of 50-word lists were elicited by requesting repetition after a live speaker model. To examine the effect of a different word set and auditory model, an additional set of 50 different words was elicited with a pre-recorded model. The recorded words were presented to normal-hearing listeners for identification via orthographic and multiple-choice response formats. To examine construct validity, production accuracy for each speaker was estimated via phonetic transcription and rating of overall articulation.

Outcomes & Results—Recording and listening tasks were completed in less than six minutes for all speakers and listeners. Aphasic speakers were significantly less intelligible than neurologically healthy speakers and displayed a wide range of intelligibility scores. Test-retest and inter-listener reliability estimates were strong. No significant difference was found in scores based on recordings from a live model versus a pre-recorded model, but some individual speakers favored the live model. Intelligibility test scores correlated highly with segmental accuracy derived from broad phonetic transcription of the same speech sample and a motor speech evaluation. Scores correlated moderately with rated articulation difficulty.

Conclusions—We describe a computerized, single-word intelligibility test that yields clinically feasible, reliable, and valid measures of segmental speech production in adults with aphasia. This tool can be used in clinical research to facilitate appropriate participant selection and to establish

Please address correspondence to: Katarina Haley, Division of Speech and Hearing Sciences, 3124 Bondurant Hall, University of North Carolina at Chapel Hill, Campus Box 7190, Chapel Hill, NC 27599-7190, Katarina_Haley@med.unc.edu, Telephone: 919-966-9460, Fax: 919-966-0100.

matching across comparison groups. For a majority of speakers, elicitation procedures can be standardized by using a pre-recorded auditory model for repetition. This assessment tool has potential utility for both clinical assessment and outcomes research.

Keywords

speech intelligibility; articulation; aphasia; computer mediated testing

After a left-hemisphere stroke, the production of consonant and/or vowel segments is often impaired. The resulting segmental speech errors may be caused by several different mechanisms, including sensory-motor, motor programming, motor planning, and/or phonologic impairments. They are typically attributed to aphasia, apraxia of speech (AOS), unilateral upper motor neuron dysarthria (UUMND), or a combination of these disorders. Depending on their quality and presumed etiology, errors are variably characterized as phonemic paraphasias, substitutions, distortions, or distorted substitutions. Many distortions, such as segment prolongations or inter-syllabic pauses, affect supra-segmental properties of speech (e.g. stress, rate), whereas others are perceived as indistinct in manner or place of articulation or in voicing. Segmental errors may also occur in the form of substitutions or phonemic paraphasias, where production appears precise, but generates perception of incorrect phonemes in listeners.

The purpose of this paper is neither to differentiate the source of these varied errors, nor to characterize their type. It is assumed that different individuals present with different error profiles, largely due to the site of lesion, the affected neural substrate, and the type of disorder. Instead, our focus is how to quantify the magnitude of the difficulties in the most psychometrically sound and economical way possible. Our discussion will be limited to the quantification of segmental consonant and vowel changes and, since only mild effects are expected with UUMND (Duffy, 2005), our main concern will be with aphasia and AOS.

Quantifying Segmental Speech Errors

There is no doubt that quantification of segmental speech errors is essential for both methodological and clinical reasons. Without such information, the validity of comparisons among individuals is questionable, whether in the clinic or in the literature. A person who produces few segmental errors is likely to behave very differently and respond very differently to treatment compared to an individual who produces a high density of errors. Moreover, examination of change over time, secondary to spontaneous recovery, treatment generalization, and/or disease progression is fundamentally contingent upon the quantification of severity, whether the scope of the analysis is an individual patient or a population with a particular disorder. Finally, the development of practice guidelines, as well as the clinical application of evidence-based practice, necessitates consideration of severity in formulating clinical questions and assessing the available evidence.

Typically, relatively crude quantification procedures are used to describe segmental speech errors in individuals with left hemisphere lesions. The most common approach is perceptual scaling with anywhere from three to seven categories, such as “mild,” “moderate,” or “severe.” At first glance, this approach seems economical and valid, requiring minimal preparation and data analysis and presumably integrating whatever information the listeners consider essential to severity. Often, perceptual scaling measures are embedded in more general estimates of disorder severity, so that segmental production difficulties cannot be differentiated from other production features (Aichert & Ziegler, 2008; Bartle-Meyer & Murdoch, 2010; Brendel & Ziegler, 2008; Dabul, 2000; Duffy, 2006). Perceptual scaling can also be conducted more specifically for segmental speech production by asking judges to rate the magnitude of phonemic paraphasia or articulatory difficulty (Goodglass, Kaplan, & Barresi, 2000; Robin, Jacks, Hageman, Clark, & Woodworth, 2008; Ziegler, Thelen, Staiger, & Liepold, 2008). All these perceptual scaling procedures have inherent limitations in sensitivity. Additionally, given the subjective nature of the rating task, reliability should be examined before such procedures are adopted by the research community. Preliminary empirical evidence indicates that agreement, even among highly experienced raters, may be limited for this type of rating scale metrics (Haley, Jacks, de Riesthal, Abou-Khalil, & Roth, 2010).

The use of an articulation test might provide more sensitive and objective quantification (Davis, Farias, & Baynes, 2009), but the lack of consistent phonological error patterns and the presence of anomic difficulties in most individuals with left hemisphere lesions reduce the validity of this approach. A more valid approach for aphasic individuals is to code a speech sample using phonetic transcription (Peach & Tonkovich, 2004; Shuster & Wambaugh, 2000) and derive an overall measure of production accuracy from these data. However, this approach requires significant time commitment by phonetically trained listeners and is not feasible for clinical application or large-scale data collection where a simple severity metric is sought. Other forms of error type quantification that yield frequency of occurrence, such as percent words with distortions or substitutions have also been used (Haley, et al., 2010; Staiger & Ziegler, 2008) and may be more efficient than transcription-based approaches.

Another approach to the assessment of overall production errors and severity of production deficits is to use a word intelligibility test. Intelligibility is defined as the degree to which a speaker’s intended message is understood by a listener. It has a rich history of use as an index of overall severity or information transfer in other disorders that affect speech production, such as dysarthria and hearing loss (Beukelman & Yorkston, 1979; Kent, 1992). Intelligibility tests based on single words are particularly sensitive to consonant and vowel distortions and substitutions. Thus, segmental production errors in aphasia, AOS, and UUMND might also be well indexed with a measure of intelligibility.

Single Word Intelligibility Assessment in Dysarthria

Several speech intelligibility tools are available for quantifying production severity in individuals with dysarthria. Depending on the purpose of the evaluation, the clinician or researcher may select tests that are composed of different units and according to different principles. The present discussion is restricted to single word intelligibility tests, since sentence production and more contextual sampling tasks are difficult for many people with aphasia, due to the prevalence of language impairments at lexical and syntactic levels.

In typical word intelligibility tests for dysarthric speakers, single words are produced by the speaker and presented, out of context, to unfamiliar normal-hearing listeners who respond by writing or selecting from several alternatives the words they think the speaker was trying to say. Because the index can be affected by factors that have nothing to do with the speaker, such as the quality of the signal transmission or the listeners’ skills and expectations, care must be taken to keep those factors consistent to the greatest degree possible. One of the more challenging factors to control is the possibility of listener learning effects. If the same listener is asked to score more than a single test, as is typically the case in clinical practice, his or her knowledge or expectations about the content of the speech sample can influence the test scores. The most common strategies to reduce this risk is to evaluate only a subset of potential target words during each new test administration and to assemble the test so that target words are phonetically similar to each other. The *Assessment of Intelligibility of Dysarthric Speech* (Yorkston & Beukelman, 1984) is one of the most well-known intelligibility tests for dysarthria. The purpose of this test is to obtain an estimate of overall severity. The dysarthric speaker reads 50 one- or two-syllable words, one at a time. The productions are audio-recorded and presented for identification to one or more listeners. The main strategy for limiting learning effects in listeners is that target words are selected at random from a larger set of potential words, so that two different speech sample recordings are rarely the same.

Other word identification tests for dysarthria have the dual purpose of estimating overall severity and identifying the phonetic patterns associated with reduced intelligibility. These tests are typically based on a multiple-choice response format and designed so that the listener must select one alternative from several phonetically similar words. The phonetic difference between the target and selected word is then examined to identify particularly challenging distinctions. For example, the *Multiple-Word Intelligibility Test* (Kent, Weismer, Kent, & Rosenbek, 1989) uses a four-alternative forced choice response format, and the computerized *Speech Intelligibility Test* (Yorkston, Beukelman, Hakel, & Dorsey, 1996) uses a six-alternative response format based on these principles. The approach has proved productive for several types of dysarthria, leading to the identification of phonetic patterns that affect intelligibility negatively and the demonstration of relationships between test profiles and acoustic markers of speech production (Bunton & Weismer, 2001; Kent, et al., 1989).

Use of the *Multiple-Word Intelligibility Test* (Kent, et al., 1989) as a clinical instrument is limited, due to the small number of target words and unavailability of strategies for random word list generation. Only four response alternatives are available and they are constructed so that only one word, the target, forms a minimal pair with all the other words. The single word test from the *Speech Intelligibility Test* (Yorkston, et al., 1996) handles potential listener learning with a similar approach as the *Assessment of Intelligibility of Dysarthric Speech* (Yorkston & Beukelman, 1984), by random selection of target words from a larger pool. Like the *Multiple-Word Intelligibility Test*, listeners are presented with phonetically similar response options, all monosyllabic words, but for the *Speech Intelligibility Test*, response alternatives typically differ in the same segment (e.g. initial consonant, vowel, or final consonant). This design permits random selection of target words but can in some cases encourage listeners to use a process of elimination in selecting the response.

Intelligibility Assessment in Aphasia and Apraxia of Speech

Word intelligibility testing has been applied to individuals with aphasia and AOS only to a very limited extent and with uncertain or impractical techniques. The fundamental problem is that available tools are based on the use of oral reading of written words presented on cards or a computer screen. Most people with aphasia have significant difficulties with this particular task due to language processing impairments that are separate from their production of spontaneous speech. They may display a range of behavior, including refusal to read the word, repeated unsuccessful attempts, and production of semantic paraphasias.

Repetition of a spoken model is a viable alternative elicitation procedure. AOS and aphasia syndromes associated with phonemic paraphasia or other segmental production errors usually involve significant repetition difficulties, but word repetition performance generally displays the same pattern as spontaneous speech. A possible exception may be conduction aphasia, which theoretically is associated with particularly salient repetition difficulties relative to spontaneous speech. However, even for these individuals, the repetition of single content words usually reveals significant phonemic paraphasia but is otherwise not problematic (Buchsbaum, et al., 2011; Goodglass, 1993). The most challenging complication of using a repetition format for intelligibility testing in speakers with aphasia and AOS is that it can be very difficult to record a speaker's production without recording the spoken cue. For obvious reasons, any such spoken models cannot be presented to the listeners charged with scoring the test.

Several recent studies have applied intelligibility tests for dysarthria to individuals with coexisting aphasia and AOS; however any procedures used to accommodate language impairment in the testing were not detailed. For example, the *Assessment of Intelligibility of Dysarthric Speech* (Yorkston & Beukelman, 1984) has been used in clinical studies of individuals with coexisting aphasia and AOS (Katz, McNeil, & Garst, 2010; Mauszycki & Wambaugh, 2008; Mauszycki, Wambaugh, & Cameron, 2010; Wambaugh & Mauszycki, 2010) and Katz and colleagues (2010) also used the *Multiple-Word Intelligibility Test* (Kent et al., 1989) to document characteristics of their study participant.

In previous studies in our laboratory, we have used various techniques to adapt intelligibility tests developed for dysarthria to individuals with aphasia. For example, Haley and Diakaki (2002) tested word intelligibility with the *Assessment of Intelligibility of Dysarthric Speech*, finding satisfactory test-retest reliability in speakers with aphasia and AOS. To accommodate elicitation via the repetition of a spoken model, two investigators combined efforts to provide salient models while simultaneously turning an audio-recording on and off, so that the cue would not be recorded and heard by listeners.

We also have examined the sensitivity and basic psychometric properties of the *Multiple-Word Intelligibility Test* for speech intelligibility testing in aphasic speakers with and without AOS (Haley & Martin, 2011; Haley, Wertz, & Ohde, 1998). In these studies we used different techniques to reduce listener learning, by adding dummy words or by scoring only a subset of randomly selected target words. Additionally, instead of using the standard multiple-choice response format for the test, we employed an orthographic response format. By asking listeners to write the words they thought the speakers produced rather than select them from a field of four, it was possible to reduce further the likelihood that they would learn and/or recognize the targeted speech sample. To elicit each word production, a written word cue was combined with a live speaker model. The entire session was audio-recorded, and individual word productions were digitized for intelligibility scoring by a listener. This procedure permitted elimination of the spoken model and randomization of the utterances for presentation to a listener. The results showed excellent sensitivity and satisfactory test-retest reliability and construct validity. However, the labor-intensive process would be impractical for clinical applications or large-scale data collection.

In summary, the application of single word intelligibility tests developed for dysarthria to individuals with aphasia and/or AOS is supported both rationally and empirically, but existing clinical tools have not been customized for this population. Further, existing procedures may promote listener learning effects and require time-consuming adaptations for testing individuals with impaired oral reading.

Purpose—The main purpose of this study was to examine the sensitivity, test-retest reliability, construct validity, and procedural efficiency for a new single word speech intelligibility test, developed for aphasic individuals with and without AOS. A related purpose was to examine the clinical feasibility of administering this test with custom software designed to automate all aspects of data collection that do not involve human interaction or judgment. We asked the following questions:

1. Are sensitivity, construct validity, and test-retest reliability for a new speech intelligibility test satisfactory in a clinical sample of aphasic individuals with and without AOS?

2. Does the use of custom software enable test administration within a clinically feasible time?
3. Is a live speaker model necessary to elicit the speech sample, or can a pre-recorded model be used, thereby providing a possibility to standardize administration procedures?

Method

Participants

The speaker participants in this study were 23 individuals who had sustained a stroke to the left cerebral hemisphere and carried a diagnosis of aphasia, and 20 neurologically healthy volunteers. Stroke survivors were recruited via study information shared by their treating neurologist or speech-language pathologist. To qualify for enrollment, they were required to be at least four weeks post stroke onset, have the ability to repeat single words, and not carry a diagnosis of progressive neurologic disorder. Demographics and clinical test results for aphasic participants are provided in table 1. The mean age for the 10 female and 13 male aphasic participants was 62 (SD = 13.7). Neurologically healthy speaker participants were recruited through word of mouth and posted flyers. Often, family members of participants with aphasia volunteered. This group included 7 males and 13 females and ranged in age from 29 to 94 (M = 63, SD = 13.6). Speakers in both groups were native speakers of English. They passed a hearing screening at 40 dB HL in the better ear for 1000 Hz and 2000 Hz (Ventry & Weinstein, 1983). The study was approved by the Institutional Review Board of the University of North Carolina at Chapel Hill, and all participants provided signed informed consent.

We attempted to retrieve brain MRI or CT scans and pertinent neurological reports for all aphasic participants. However, because several participants were many years post stroke onset, access to the records was inconsistent. In the end, we were able to access scans of adequate quality for 14 participants and review neuroradiologic reports for an additional five participants. Lesion localization and etiology are summarized in table 1. Neither scans nor neurologic reports were available for four of the participants. However, these participants gave a history consistent with stroke and identified date of onset with confidence.

To classify aphasia type and estimate severity of language impairment, we administered the *Aphasia Diagnostic Profiles* (ADP; Helm-Estabrooks, 1992). A range of severity scores was obtained, from the 9th to the 98th percentile. Seven participants profiled with a nonfluent type of aphasia, seven with a fluent type of aphasia, and nine were classified with borderline fluency. A structural-functional examination of the speech mechanism and a standard motor speech evaluation requiring repetition of nonsense syllables, words, and sentences were also administered (Duffy, 2005; Wertz, LaPointe, & Rosenbek, 1984). All aphasic participants demonstrated either no dysarthria or mild upper motor neuron dysarthria (UUMND). The motor speech evaluation was audio-recorded, and all 23 speech samples were presented to three certified speech-language pathologists with more than ten years experience in the differential diagnosis of motor speech disorders. The clinicians were unfamiliar with the participants and unaware of the results of the clinical testing. They were familiar with contemporary criteria for AOS (Duffy, 2005; McNeil, Robin, & Schmidt, 2009; Wambaugh, 2006), but asked to make their own judgments concerning diagnosis. They listened independently to the audio-recordings for each speaker and rated the presence of AOS on a three-point rating scale (1 = no AOS, 2 = possible AOS, 3 = AOS). Agreement within a single scale level for all three clinicians was 52%, and agreement among clinician pairs was 91%, 87%, and 61%. These low levels of agreement are clearly unsatisfactory and illustrate the problems with impressionistic diagnostic practices. For this reason, we offer the mean rating scores across the three clinicians instead of the judgment of a single clinician as an indication of AOS. As shown in table 1, there was unanimous agreement that four speaker participants had AOS and that one did not have AOS, whereas diagnosis for the remaining 18 participants was uncertain. Finally, a narrative was elicited based on the picnic scene from the Western Aphasia Battery (Kertesz, 2006). This narrative was audio-recorded for the purpose of rating severity of articulation difficulty, as will be described in a following section on validation measures.

Intelligibility Test and Recording

Software development—Custom software was developed to automate the preparation, administration, and scoring process for speech sample capturing and intelligibility testing. The software runs on a PC computer and is available for research purposes by contacting the first author. It is presently not available for commercial use. Six modules were included: (a) Preparation of the speech sample by custom or random selection of words within sets; (b) Presentation of images, written words, and spoken words in any combination of modalities and with single key controls for recording and storing each elicited utterance as a separate audio-file; (c) Acoustic analysis and editing capabilities via linked spectrographic and waveform displays; (d) Perceptual testing with orthographic transcription and multiple-choice response formats, each scored automatically; (e) Post-testing review of speaker recordings, listener responses, and error patterns; and (f) Data integration, analysis, and export.

Test development—A new single word intelligibility test was developed and integrated with the software. The complete test is provided in the appendix and is available on-line, along with versions in other languages (Haley, 2011). The test was constructed to allow generation of a new, but comparable, speech sample for each recording session. A pool of 600 different monosyllabic words was used to construct the test. These words were organized in 50 sets of 12 phonetically similar words, similar to the *Assessment of Intelligibility for Dysarthric Speech* (Yorkston & Beukelman, 1984).

Like intelligibility tests designed to identify phonetic or phonemic error patterns, the test was constructed so that all words in a set had the same number of syllables (monosyllabic in this case) and were maximally similar to each other (Kent, et al., 1989; Yorkston, et al., 1996). However, unlike other tests, the words within each set differed from each other in more than a single consonant or vowel. Thus, several different types of minimal pairs were included and the words were linked to each other in several different ways. This arrangement was selected to minimize response predictability and to increase sensitivity to a range of different types of segmental errors. Specifically, each word was required to form a minimal pair with at least one other word in the set and over ¾ of the words formed a minimal pair with nine or more words in the same set (range 9–12 words, mean = 10, SD = 0.95). At the same time, there were at least three different types of minimal pairs in each word set (range 3–9, mean = 6.3, SD = 1.4). Only one example in homophone word-pair was included (e.g. “flee” but not “flea”) and there were no homographs with different pronunciation in standard American English (e.g. “sow”). All words were common nouns, verbs, adverbs, or adjectives, and they occurred with a frequency of at least 1/1,000,000 according to standard frequency norms (Kučera & Francis, 1967; Leech, Rayson, & Wilson, 2001).

A spoken model for each word was recorded and stored as a separate audio-file within the software program. The speaker was a female in her twenties who used a General American dialect. She was asked to produce each word clearly, at a conversational rate, and separate in intonation from words recorded before or after. The audio-files were linked to the written words and to a homophone library of acceptable alternative spellings. For each recording, except those repeated to estimate test-retest reliability, unique word lists were generated by random selection of one of the twelve words from each word set.

Recording of the speech samples—Speech samples were recorded in a quiet room in either the laboratory or the participant's home. Responses were recorded directly on the internal soundcard of a laptop computer via a head-mounted microphone (AKG-C420). The pre-recorded auditory models were presented via external speakers with the volume set at a level that was determined comfortable by each participant. Throughout the recording task one written word at a time was presented in 112 point font on a 12″ × 9″ computer screen, a live or pre-recorded auditory model of this word was presented, and the speaker repeated the word.

A brief training, consisting of ten words, was introduced to familiarize the speaker with the recording procedure. Next, three sets of 50-word intelligibility speech samples were recorded. Two speech samples with identical word lists were recorded consecutively to allow examination of test-retest reliability. For these sets, the examining clinician provided the auditory model, by reading the word on the screen aloud, and recorded the speaker's response via a single keystroke or mouse click. A third speech sample was obtained using a pre-recorded auditory model, with a different random list of 50 words. Upon presentation of the word on the screen, the recorded model was played through the external speakers, and the participant repeated the word. The clinician's role in this condition was to activate the mouse or a key to stop the recording and advance to the next word. This speech sample was recorded either prior to or following the consecutive test-retest reliability recordings.

For both elicitation conditions, one re-stimulation was provided for each word upon request or if a response was not produced within approximately five seconds. Any partial or whole word repetitions or attempts to self-correct were recorded as part of the participant's attempt to produce the target word. The recording sessions for all aphasic participants were video-recorded and later coded for elicitation efficiency and overall duration. A measure of elicitation responsiveness was calculated as the percentage of words that were produced within five seconds or less on the first trial, as coded by a trained observer. A second observer coded a random selection of 37% of the sessions independently and obtained scores within two percentage points for all observations. Overall session duration was estimated from the video player counter and verified by a second observer.

Perceptual intelligibility testing

Listeners—Due to the large number of speaker participants, three groups of listeners were used to complete the intelligibility testing. Listener participants were graduate or undergraduate students taking preclinical or clinical coursework in communication sciences and disorders. They were recruited through posted advertisements on a student bulletin board and electronic listserv and received a small monetary compensation for their participation. One group of five graduate students (all female, mean age 25, range 21–37 years) listened to the neurologically healthy participants. A second group of ten graduate students (eight female, mean age 26, range 22–43 years) listened to the repeated speech samples from the aphasic speakers in the live clinician elicitation condition. Finally, ten undergraduate students (nine female, mean age 25, range 21–39 years) listened to the first live clinician and pre-recorded elicitation conditions for the aphasic speakers.

Graduate student listeners had completed at least one graduate level course in the clinical management of neurologic communication disorders. Undergraduate students had completed one or more pre-professional course in communication sciences and disorders and were planning to continue graduate studies in the field. All listeners were native speakers of

English. They passed a hearing screening at 25 dB HL for the octave frequencies from 500 Hz to 8000 Hz, reported normal corrected visual acuity, and had no history of speech or language impairment.

Listening sessions—Listening sessions included presentation of a single speech sample from each of the speakers and lasted approximately 1.5 hours. If multiple sessions were required, they were scheduled between 7 and 14 days apart. In these cases, the order of the speech samples was randomized for each speaker and each listener. The sessions were conducted in a sound-treated IAC booth, using the same laptop computers and software that were used for speech sample recordings. The output was presented through circum-aural headphones (Sennheiser HD 25-1) with the volume set at a level that the listener subjectively found most comfortable. One word at a time was presented in random order. The listeners were asked to identify the words they thought the speakers were trying to say and to guess if they were not sure. They were permitted to listen to each utterance only once. Responses were entered using the mouse or keyboard, with the listeners controlling the rate of presentation and managing the data collection independently. Listeners were encouraged to take a break at any time and required to break for at least 10 minutes midway through testing.

Response formats—The aphasic speakers’ recordings from the live clinician condition were scored with both an orthographic and a multiple-choice response format, while recordings from the pre-recorded elicitation condition were scored only with an orthographic response format. The order of the response conditions was counterbalanced across listeners on a session-by-session basis, so each listener used a consistent response format throughout the session. In the orthographic response format, a response window was positioned in the middle of the screen. Listeners were asked to type the word they thought the person was trying to say. They were assured that alternate but accurate spellings were acceptable but that they would be prompted to reenter their response if it was not a real word. If unable to guess the target word for any item, they were permitted to type the word “nothing” to indicate that they could not understand it. An electronic spell checking engine was integrated with the software to verify that all entries were real words and a library of possible homophones was developed for each of the 600 words in the test to account for alternate spelling. In the multiple-choice response condition, twelve phonetically similar words (representing a word set in the test design, see appendix) were presented on the screen in random order, and listeners were asked to select the word they thought the person was trying to say. Overall intelligibility was computed automatically by the software, accounting for homophones in the orthographic response condition.

For the 20 neurologically healthy speakers, a single speech sample, the first recording in the live clinician elicitation condition, was presented and listeners were tested with the orthographic response format only. Only this single speech sample was analyzed, even though these speakers participated in the same recording protocol as the aphasic participants. The purpose of the analysis was to provide a point of comparison with intelligibility scores from the aphasic participants.

A 10-minute training session was completed at the beginning of the test sessions to explain the test procedures and to ensure that all listeners were familiar with the task. This training included demonstration of the software, and practice on recordings from a different set of speakers with aphasia and AOS.

Validation Measures

To establish construct validity of the speech intelligibility test as an index of segmental speech production difficulties, we obtained three measures of degree of production difficulties for the aphasic speakers: phonetic transcription of an intelligibility test speech sample, phonetic transcription of the motor speech evaluation, and subjective rating of overall articulation.

Phonetic transcription of the intelligibility speech sample—Two phonetically trained listeners transcribed the first speech sample recording from the live clinician elicitation condition, using broad phonetic transcription. All 1150 productions (23 speakers × 50 utterances) were transcribed. To avoid potential bias from listener expectations, the transcribers were not informed what the target productions were.

For this measure, we used a consensus transcription procedure. First, both coders completed transcription of all words independently. This transcription resulted in a point-to-point agreement of 93% for consonants and 95% for vowels. One or more segments were transcribed differently by the two observers for 195 utterances (17% of the sample). These points of disagreements were reviewed during a joint listening session where the coders discussed their impressions and repeated listening until they agreed on the final transcription. The target words were then revealed and the consensus transcription was compared to the target words by the two coders working jointly. The percentage of perceived vowel and consonant segments that corresponded (in the correct order) to target word segments was computed.

Phonetic transcription of the motor speech evaluation—The audio-recorded motor speech evaluation consisted of repetition requests for 41 items, ranging from monosyllabic words to multisyllabic words and multi-word sentences (Duffy, 2005; Wertz, et al., 1984). All attempted items were transcribed phonetically. For this sample it was not possible to conceal the target words and sentences, so they were printed on the transcription sheets. One phonetically trained observer listened to the 943 utterance attempts (23 speakers × 41 items) and transcribed the productions using broad phonetic transcriptions. The percent of segments that was produced correctly was computed for each speaker by dividing the number of segment transcriptions that matched the target transcriptions by the number of segments in the attempted words. Due to difficulties repeating multisyllabic words and sentences, participants occasionally rejected items, and these rejected items were not included in the computation of production accuracy. A second observer transcribed the speech sample from 22 of the speakers independently and computed percent correctly produced segments. Inter-rater reliability, based on the Pearson product moment correlation, was .96.

Subjective rating of overall articulation—Audio-recordings of the *WAB* picnic scene narrative were presented to six raters for global rating of articulation difficulty. Three of the raters were certified speech-language pathologists who worked in a rehabilitation setting at a major medical center. Each had between 1.5 and 2.5 years of professional experience working with adults with neurologic communication disorders. Three additional untrained listeners were included. These listeners, two females and one male, were 36, 57, and 65 years old. Their professional backgrounds were in accounting, clinical child psychology, and high school education, and they had no personal or professional experience with stroke.

The raters completed the listening session individually in a single session. They were given the *WAB* picnic scene stimulus card and a CD with 23 tracks, arranged in a different random order for each listener. Each track corresponded to the recorded narrative from one aphasic speaker. The raters were told that the speakers who produced these speech samples had survived a stroke and that the stroke had affected their speech in different ways. They were asked to listen to the speech sample and rate overall articulation on a seven-point rating scale from profound articulation difficulties to no articulation difficulties. Articulation difficulties were defined as saying the wrong speech sounds, pronouncing sounds unclearly, saying only part of the word, or not saying words at all. Raters were allowed to listen to each speech sample as many times as they liked. Two subjective rating scores were obtained by calculating the mean from these ratings. One score was obtained for the three clinicians, and one was obtained from the three everyday listeners. Agreement within two scale levels for all three raters was 74% for the speech-language pathologists and 83% for the everyday listeners.

Results

The recording and listening sessions progressed at a clinically feasible rate. Recording and listening sessions for each 50-word speech sample were completed in less than six minutes for every speaker and listener, but often took only half that time. As expected, intelligibility scores were markedly different for the neurologically healthy speakers and the speakers with aphasia. Group results for the aphasic speakers within the different elicitation conditions, recording sessions, response formats, and listener groups are provided in table 2 and individual scores for the live elicitation condition are plotted in figure 1. The mean orthographic intelligibility from the live recording for the healthy speakers was 96% (SD = 3.6%) with a range from 87% to 100%, and the comparable score for the aphasic speakers was 69% (SD = 24.0%) with a range from 3% to 95%. As expected, an independent samples t-test showed that this difference was statistically significant [t(23.1) = 5.2, p<.001]. Having established construct validity at this very basic level, our attention turned to comparisons across conditions for the aphasic speakers.

Effect of Elicitation Condition

The mean duration of the entire recording session was slightly longer for the live elicitation mode (4 minutes and 12 seconds) than for the pre-recorded elicitation mode (3 minutes and 43 seconds; t(22) = 2.93, p = .008)). Elicitation responsiveness was 99% for the live model condition and 98% for the pre-recorded model condition. Thus, based on observations conducted at the time of recording, the two conditions appeared to have comparable clinical feasibility.

To examine whether the two elicitation conditions yielded different rates of intelligibility, the samples were presented to ten undergraduate students for identification in an orthographic response format. Due to a problem recording some of the utterances in the pre-recorded elicitation condition for one speaker (P20), comparisons between elicitation conditions were restricted to data from the remaining 22 speakers. As shown in table 3, mean intelligibility for these speakers was comparable for repetition of a live clinician model (66%) and repetition of a pre-recorded model (64%). A paired sample t-test confirmed that the difference was not statistically significant [t(21) = 1.37, p = .18].

Inspection of individual data showed that intelligibility scores for the two elicitation conditions differed by less than ten percentage points for the majority of the speakers. However, the live elicitation condition yielded higher scores than the pre-recorded model condition by 11 to 19 percentage points for five of the speakers (P03, P04, P05, P10, and P13). These participants did not stand out from the rest based on clinical test results. Three profiled with Broca’s aphasia and two with borderline fluent aphasia. Their overall aphasia severity ranged from the 30th to the 63rd percentile and their AOS diagnosis ratings were consistent with AOS or possible AOS. No participants showed an advantage for the pre-recorded condition.

Reliability of Intelligibility Testing

Test-retest reliability was strong. Scores for individual speakers are displayed in figure 1. As can be seen, intelligibility scores were highly correlated for both the multiple choice and the orthographic response formats (r = .97, p<.001). As shown in table 3, mean intelligibility was nearly identical for the first and second recordings in both the multiple choice and orthographic response formats. As expected, these differences were not statistically significant [t(22) = 0.17, p = .86; t(22) = 1.26, p = .22]. Moreover, 91% of scores were within six percentage points between the first and second recording for both response conditions. Only two speakers (P05 and P20) showed a greater difference, with a 12 to 14 percentage point difference in both response conditions.

Predictably, scores for the multiple choice response condition were higher than scores from the orthographic response condition for the group as a whole [t(22) = 10.9, p<.001] and for all speaker participants individually. The difference was larger for speakers with lower intelligibility, due to a ceiling effect for highly intelligible speakers in the multiple-choice condition (see figure 1).

Inter-observer reliability estimates were also strong. Intra-class correlations, based on a two-way mixed effects model, were .97 or greater (p<.001) in all four conditions for the graduate students and in both conditions for the undergraduate students. Furthermore, the Pearson product moment correlation for both recording sessions and response formats exceeded .92 for all possible pairs of listeners within each of the six conditions.

The availability of orthographic transcription test results for the live elicitation condition for both graduate and undergraduate students permitted a comparison between listener groups. The graduate students scored an average of only two percentage points higher than the undergraduate students, but the slightly higher performance was observed for 21 of the 23 speakers and reached statistical significance (t(22) = 4.28, p<.001).

Validity of Intelligibility Testing

The strength of the relationship between each of the validation measures and intelligibility for the speech sample elicited with a live clinician model is detailed in table 3. For comparisons involving only percentage data, we used the Pearson product moment correlation, and, for comparisons involving ratings, we used the Spearman rank order correlation.

There was a strong relationship between intelligibility scores derived from both orthographic and multiple choice response formats and percent segments correct as determined from phonetic transcription of the same speech sample (r=.97 and .98, p<.001). To establish construct validity, it is important to consider the relationship not only among different quantification methods, but also among different behavior samples. For this purpose, the intelligibility results were compared to the segmental integrity of a standard motor speech sample obtained during the same recording session. The content of this sample differed substantially from the intelligibility sample, but it was also elicited via repetition of sequentially unrelated utterances modeled by a clinician. The results showed that, for this variable, there was a statistically significant and moderately strong correlation between intelligibility scores and percent segments correct as determined by phonetic transcription (r=.74; p<.001).

As our two final estimates of segmental speech production difficulty, we turned to subjective ratings of a narrative speech sample. These measures were selected because they differed considerably from the intelligibility measure both in the nature of the speech task and in its quantification, and would conceptually capture similar speaker difficulties. The results showed a statistically significant moderate correlation with intelligibility scores (ρ = .55 to .64; p<.001).

Discussion

The results demonstrated excellent sensitivity, construct validity, test-retest reliability, and inter-observer agreement for a new monosyllabic intelligibility test developed for aphasic and apraxic speakers. The first criterion for an index designed to estimate magnitude of involvement must be sensitivity to individual differences. In this study, a greater range of scores was observed (from 3% to 95% in the orthographic response condition) than in previous reports (Haley & Diakaki, 2002; Haley & Martin, 2011; Haley, et al., 1998). This observation is likely due the use of a larger and more varied participant sample, and it is encouraging. The response format affected sensitivity in a predictable manner. Scores were significantly higher in the multiple-choice condition than in the orthographic transcription condition and the range of scores was smaller (23% to 99%). More importantly, a ceiling effect was obtained for the most intelligible speakers in the multiple-choice condition, with more than half the participants (N = 12) scoring above 90%, compared to only two participants in the orthographic response condition. Thus, we recommend the use of an orthographic response condition for individuals with aphasia and/or AOS. The recommendation is supported also by the anticipated objective of obtaining an overall measure of severity rather than a phonetic profile in this population.

The single word intelligibility test appears to be a valid index of magnitude of segmental production errors. At the most basic level of construct validity, the scores were significantly and substantially below those for an age-matched control group of neurologically healthy individuals. Within the aphasic participant group, intelligibility scores were significantly and strongly correlated with segmental accuracy scores derived from broad phonetic transcription of both the same and a different speech sample, a finding consistent with previous work (Haley, Bays, & Ohde, 2001; Haley & Martin, 2011). Significant correlations of moderate strength were also obtained between the intelligibility scores and subjective ratings of overall articulation. The moderate magnitude of the correlation with these perceptual scaling measures is not surprising, given their smaller range of scores, the subjectivity of the rating, and the influence of other factors unrelated to segmental speech production, as reported previously (Haley, et al., 1998).

Consistency of scores on repeated administrations is crucial for testing of individuals with aphasia and AOS, as both disorders have been associated with variable segmental speech sound production across repeated attempts. In this study, test-retest reliability was excellent (r = .97), consistent with a recent study (Haley & Martin, 2011) that examined both test-retest reliability and token-to-token variability of monosyllabic word production from the *Multiple-Word Intelligibility Test* (Kent, et al., 1989). Haley and Martin (2011) showed that even though test-retest reliability was strong (r = .98), perceived sound substitutions for the individual words varied considerably from recording to recording. Thus, converging evidence indicates that strong test stability can be achieved as long as the measure is based on whether or not the target word was understood and on the integration of data from multiple listeners and approximately 50 single words.

A major purpose for developing custom software was to allow simultaneous presentation of both written and spoken word stimuli and efficient and accurate recording of only the speaker's response. This objective was met, with recording sessions progressing smoothly and requiring only a single keystroke by the investigator for each word. The testing was further automated throughout all phases of the testing, allowing highly efficient data collection. The elicitation and recording of 50 single words was completed in approximately four minutes for the aphasic speakers, never exceeding six minutes for any participant. The efficiency of the procedure supports application not only to clinical practice, but also to large-scale data collection and examination of speech production during more acute stages of stroke recovery.

Similar to speech sample elicitation and recording, perceptual scoring of the intelligibility tests was also efficient. The listeners controlled the pace of their own testing and scored each intelligibility test in approximately three to four minutes. Although scheduled listening sessions were used in this study, the software design allowed listeners to log on at their convenience and complete any speech samples assigned to them. This process eliminates much of the practical obstacles involved with intelligibility scoring in a busy clinical environment, where listeners, who often are colleagues of the treating clinician, have limited availability for scheduled listening sessions. The use of computer-mediated testing provides opportunities to complete scoring remotely, not only in time, but also in place (Ziegler & Zierdt, 2008). This enabled the recruitment of larger pools of listeners potentially from anywhere in the world, so long as signal transmission factors are kept consistent (McHenry, 2011).

There was minimal difference in session recording duration between the live speaker model condition and the pre-recorded model condition, and no significant difference in intelligibility scores. Thus, the use of either a pre-recorded model or a live model would be appropriate, at least for large scale data collection. By using a pre-recorded model it is possible to standardize the auditory model and avoid potential effects of speaker dialect, rate, or other idiosyncratic speech patterns modeled by the clinician or investigator. Additionally, speech samples can be elicited remotely in both time and place, providing opportunities for more extensive data collection and monitoring of patients over time. However, it must be noted that five of the 23 participants showed an advantage of the live clinician elicitation condition whereas none displayed an advantage of the pre-recorded condition. It is possible that these five individuals benefited from the addition of visual cues provided by the investigator.

The scope of the current study was limited to investigating the sensitivity, test-retest reliability, and construct validity of this new test, leaving several important questions for further study. Of particular interest is whether subsequent administrations of the test using newly generated random word lists (i.e. parallel forms) will yield comparable intelligibility. The test was designed specifically with the intent to minimize listener learning by using random selection of target words, thus increasing the chance of unbiased intelligibility scores over repeated tests by different speakers. However, random selection of words might have unintended effects, resulting in variable levels of difficulty across tests that may affect speaker performance.

Lexical-semantic properties are among the factors that may have influenced word repetition, particularly since many of the participants presumably had impaired phonologic processing abilities. When assembling the test, care was taken to avoid prepositions, pronouns, contractions, and other function words with low semantic content. Moreover, the word pool included only words that were among the most frequently used in written and spoken English, so adult native speakers would be as familiar as possible with all words. In this way, we attempted to minimize the speakers' reliance on phonologic working memory when reproducing the words.

Phonetic complexity also may impact the likelihood of accurate segmental production. One of the primary reasons for including only monosyllabic words in the test was to limit variation in overall phonetic complexity of the word lists. However, it is possible in principle that two or more randomly generated tests may yield target word samples that differ in complexity of the onset, nucleus, and coda for the constituent words. Because different target word samples were used for the pre-recorded and live clinician elicitation conditions, and because comparable levels of intelligibility were found for the two conditions, parallel forms reliability is likely satisfactory for this test. However, further study is currently underway to examine reliability for alternate test forms and to explore the potential contribution of phonetic complexity on error rates.

Single-word intelligibility testing focuses by definition on segmental production, which is only one aspect of speech production in aphasia and AOS. Notably, suprasegmental variations are also important and should be quantified. Reduced speaking rate, segmental and inter-segmental prolongations, and abnormal stress are among such variations that are particularly likely to influence differential diagnosis. For example, while individuals with AOS and phonemic paraphasia both would be expected to have segmental speech errors, impairments of prosody are expected in the former but not the latter. In any case, we contend that single-word intelligibility testing provides an important tool for reliably quantifying segmental aspects of speech output in individuals with aphasia and/or AOS.

Conclusion

The results of the study have implications for both clinical practice and clinical research. A computerized single-word intelligibility test appears to be a viable tool for quantifying magnitude of segmental speech errors in people with left hemisphere lesions, thereby facilitating comparison among individual speakers. Monosyllabic word intelligibility is likely to be sensitive to both recovery and deterioration of speech production and may be useful for outcomes research and clinical documentation. Further research is needed to establish such properties empirically. However, we are hopeful that the measure may become as relevant to examining clinical course and response to treatment in individuals with aphasia with or without AOS as has intelligibility for the clinical management of dysarthria.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by grant R03DC006163 from the National Institute on Deafness and Other Communication Disorders. We gratefully acknowledge the contributions by Dr. Xuefeng Yang who wrote the software code and the many research assistants who helped with data collection, including Rachel Goff, Clay Hadden, Megan Jacobsen, Olivia Lysakowski, Gwenyth Martin, Kerry Mandulak, Betsy Neufeld, and Katherine Steere. Portions of this research were presented at the Annual Convention of the American Speech-Language-Hearing Association, in 2008 (Chicago, IL) and in 2009 (New Orleans, LA).

References

Aichert I, Ziegler W. Learning a syllable from its parts: Cross-syllabic generalisation effects in patients with apraxia of speech. Aphasiology. 2008; 22(11):1216–1229.

Bartle-Meyer CJ, Murdoch BE. A kinematic investigation of anticipatory lingual movement in acquired apraxia of speech. Aphasiology. 2010; 24(5):623–642.

Beukelman DR, Yorkston KM. The relationship between information transfer and speech intelligibility of dysarthric speakers. Journal of communication disorders. 1979; 12(3):189–196. [PubMed: 438358]

Brendel B, Ziegler W. Effectiveness of metrical pacing in the treatment of apraxia of speech. Aphasiology. 2008; 22(1):77–102.

Buchsbaum BR, Baldo J, Okada K, Berman KF, Dronkers N, D'Esposito M, et al. Conduction aphasia, sensory-motor integration, and phonological short-term memory - An aggregate analysis of lesion and fMRI data. Brain and Language. 2011

Bunton K, Weismer G. The relationship between perception and acoustics for a high-low vowel contrast produced by speakers with dysarthria. Journal of Speech Language and Hearing Research. 2001; 44(6):1215–1228.

Dabul, B. Apraxia Battery for Adults-2. Austin, TX: Pro-Ed; 2000.

Davis C, Farias D, Baynes K. Implicit phoneme manipulation for the treatment of apraxia of speech and co-occurring aphasia. Aphasiology. 2009; 23(4):503–528.

Duffy J. Apraxia of speech in degenerative neurologic disease. Aphasiology. 2006; 20(6):511–527.

Duffy, JR. Motor Speech Disorders: Differential Diagnosis and Management. 2. St. Louis: Mosby; 2005.

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

Goodglass, H.; Kaplan, E.; Barresi, B. Boston Diagnostic Aphasia Examination. 3. Pro-Ed; 2000.

Haley, KL. Chapel Hill Multilingual Intelligibility Test. 2011. Retrieved June 25, 2011, from http://www.med.unc.edu/ahs/sphs/card/chmit

Haley KL, Bays GL, Ohde RN. Phonetic properties of aphasic-apraxic speech: A modified narrow transcription analysis. Aphasiology. 2001; 15(12):1125–1142.

Haley KL, Diakaki A. Reliability and effectiveness of computer-mediated single word intelligibility testing in speakers with aphasia and apraxia of speech. Journal of Medical Speech-Language Pathology. 2002; 10(4):257–261.

Haley, KL.; Jacks, A.; de Riesthal, M.; Abou-Khalil, R.; Roth, HL. Quantifying the motor speech evaluation for aphasic and apraxic speakers. Paper presented at the Annual Convention of the American Speech-Language Hearing Association; Philadalphia, PA. 2010.

Haley KL, Martin G. Production variability and single word intelligibility in aphasia and apraxia of speech. Journal of Communication Disorders. 2011; 44(1):103–115. [PubMed: 20822776]

Haley KL, Wertz RT, Ohde RN. Single word intelligibility in aphasia and apraxia of speech. Aphasiology. 1998; 12(7/8):715–730.

Helm-Estabrooks, N. Aphasia Diagnostic Profiles. Pro-Ed; 1992.

Katz WF, McNeil MR, Garst DM. Treating apraxia of speech (AOS) with EMA-supplied visual augmented feedback. Aphasiology. 2010; 24(6–8):826–837.

Kent, RD. Intelligiblity in speech disorders: Theory, measurement, and management. Philadelphia: John Benjamin; 1992.

Kent RD, Kent JF, Weismer G, Sufit RL, Brooks BR, Rosenbek JC. Relationships between speech intelligibility and the slope of second-formant transitions in dysarthric subjects. Clinical Linguistics and Phonetics. 1989; 3(4):347–358.

Kent RD, Weismer G, Kent JF, Rosenbek JC. Toward phonetic intelligibility testing in dysarthria. Journal of Speech and Hearing Disorders. 1989; 54(4):482–499. [PubMed: 2811329]

Kertesz, A. Western Aphasia Battery – Revised. Pearson; 2006.

Kučera, H.; Francis, WN. Computational Analysis of Present-Day American English. Providence, RI: Brown University Press; 1967.

Leech, G.; Rayson, P.; Wilson, A. Word Frequencies in Written and Spoken English: Based on the British National Corpus. Harlow, England: Pearson Educational Limited; 2001.

Mauszycki SC, Wambaugh JL. The effects of rate control treatment on consonant production accuracy in mild apraxia of speech. Aphasiology. 2008; 22(7/8):906–920.

Mauszycki SC, Wambaugh JL, Cameron RM. Variability in apraxia of speech: Perceptual analysis of monosyllabic word productions across repeated sampling times. Aphasiology. 2010; 24(6–8):838–855.

McHenry M. An exploration of listener variability in intelligibility judgments. American Journal of Speech-Language Pathology. 2011; 20(2):119–123. [PubMed: 21317298]

McNeil, MR.; Robin, DA.; Schmidt, RA. Apraxia of speech. In: McNeil, MR., editor. Clinical management of sensorimotor speech disorders. New York: Thieme; 2009. p. 249-268.

Peach RK, Tonkovich JD. Phonemic characteristics of apraxia of speech resulting from subcortical hemorrhage. Journal of Communication Disorders. 2004; 37(1):77–90. [PubMed: 15013380]

Robin DA, Jacks A, Hageman C, Clark HM, Woodworth G. Visuomotor tracking abilities of speakers with apraxia of speech or conduction aphasia. Brain and Language. 2008; 106(2):98–106. [PubMed: 18558428]

Shuster LI, Wambaugh JL. Perceptual and acoustic analyses of speech sound errors in apraxia of speech accompanied by aphasia. Aphasiology. 2000; 14:635–651.

Staiger A, Ziegler W. Syllable frequency and syllable structure in the spontaneous speech production of patients with apraxia of speech. Aphasiology. 2008; 22(11):1201–1215.

Ventry I, Weinstein B. Identification of elderly people with hearing problems. ASHA. 1983; 25:37–42. [PubMed: 6626295]

Wambaugh J. Treatment guidelines for AOS: Lessons for future research. Journal of Medical Speech-Language Pathology. 2006; 14(4):317–321.

Wambaugh JL, Mauszycki SC. Sound Production Treatment: Application with severe apraxia of speech. Aphasiology. 2010; 24(6–8):814–825.

Wertz, RT.; LaPointe, LL.; Rosenbek, JC. Apraxia of Speech in Adults: The Disorder and Its Management. San Diego: Singular; 1984.

Yorkston, KM.; Beukelman, DR. Assessment of Intelligibility of Dysarthric Speech. Austin, TX: Pro-Ed; 1984.

Yorkston, KM.; Beukelman, DR.; Hakel, M.; Dorsey, M. Speech Intelligibility Test. Lincoln, NE: Institute for Rehabilitation Science and Engineering at Madonna Rehabilitation Hospital; 1996.

Ziegler W, Thelen AK, Staiger A, Liepold M. The domain of phonetic encoding in apraxia of speech: Which sub-lexical units count? Aphasiology. 2008; 22(11):1230–1247.

Ziegler W, Zierdt A. Telediagnostic assessment of intelligibility in dysarthria: a pilot investigation of MVP-online. Journal of Communication Disorders. 2008; 41(6):553–577. [PubMed: 18582894]

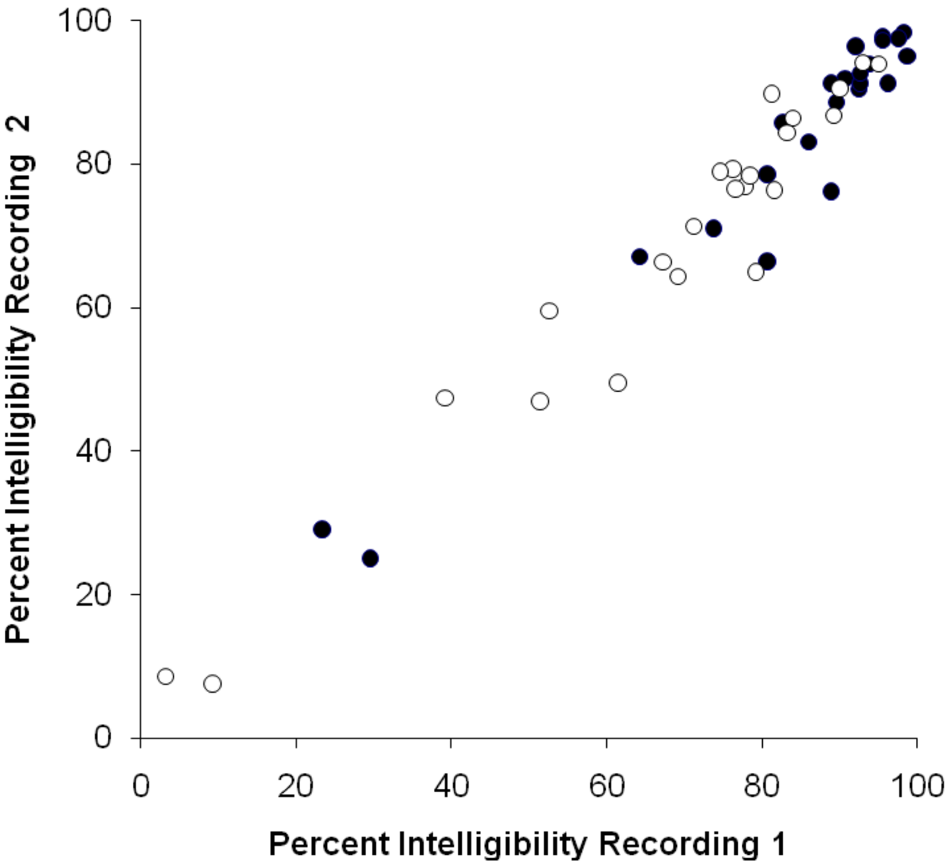

Figure 1.
Test-retest reliability for intelligibility scores from the orthographic response format (open circles; r=.97, p<.001) and the multiple choice response format (closed circles; r=.97, p<.001).

Table 1

Demographics, lesion localization, and clinical test results for participants with aphasia.

	Sex	Age	TPO	Hand	Lesion	Aphasia type	ADP	AOS
P01	F	67	5:11	R	*Not available*	Mixed nonfluent	9th	2.3
P02	M	76	0:7	R	L frontal, parietal, PVWM, prom. bilateral atrophy	Mixed nonfluent	30th	2.0
P03	F	66	0:5	R	L frontal, parietal, temporal, PVWM	Broca's	30th	2.3
P04	M	66	32:0	R	*Not available*	Broca's	58th	2.0
P05	M	43	0:5	R	L frontal, parietal, temporal BG, IC	Broca's	30th	2.7
P06	M	69	0:8	R	L frontal, temporal, some parietal, BG, IC	Broca's	30th	2.0
P07	F	63	1:10	R	L frontal, parietal, temporal	Broca's	32nd	2.3
P08	M	46	1:8	R	L frontal, parietal, temporal	Borderline fluent	61st	2.0
P09	F	68	0:8	R	L parietal, mid-temporal, some frontal, PVWM	Borderline fluent	63rd	3.0
P10	F	53	4:3	R	L frontal, parietal, temporal, IC	Borderline fluent	55th	3.0
P11	M	42	2:3	R	L parietal, some frontal, some temporal	Borderline fluent	19th	2.3
P12	M	73	6:1	R	L frontal, parietal, minimal temporal	Borderline fluent	84th	3.0
P13	F	95	0:5	R	L MCA	Borderline fluent	63rd	2.0
P14	M	69	45:0	L	L frontal, parietal, some temporal	Borderline fluent	55th	2.3
P15	F	73	3:7	R	*Not available*	Borderline fluent	58th	2.0
P16	F	46	1:7	R	L frontal, parietal, temporal,	Borderline fluent	35th	2.3
P17	M	56	6:3	R	L temporal, pre-frontal, BG, PVWM	Transcort. sensory	58th	2.3
P18	F	67	2:3	R	L parietal, mid-temporal,	Conduction	53rd	3.0
P19	M	51	1:10	R	L MCA, prom. bilateral atrophy	Conduction	58th	1.3
P20	M	78	0:4	R	L parietal (angular branch)	Conduction	61st	1.7
P21	M	72	3:7	R	*Not available*	Conduction	89th	1.7
P22	F	44	2:0	R	L mid-parietal, PVWM; R frontal supra-ventricular	Anomic	86th	2.0
P23	M	49	0:9	R	L BG, IC, PVWM	Anomic	98th	1.0

Note. TPO= time post onset in years:months; Hand=premorbid handedness; MCA= middle cerebral artery infarction not further specified; BG= basal ganglia; IC-internal capsule; PVWM= periventricular white matter abnormalities; prom=prominent; Aphasia type = classification on the Aphasia Diagnostic Profiles (Helm-Estabrooks, 1992); ADP=Overall severity percentile on the Aphasia Diagnostic Profiles; AOS=Mean score from three independent perceptual ratings (1=no AOS, 2=possible AOS, 3=AOS).

Table 2

Mean intelligibility scores for the tested response conditions, listener groups, and recording conditions.

Response condition, listener group	First recording, live elicitation	Second recording, live elicitation	Pre-recorded elicitation
Orthographic, graduate students	68.9 (24.0)	68.7 (23.7)	n/a
Multiple choice, graduate students	84.0 (19.7)	82.7 (20.0)	n/a
Orthographic, undergraduate students	66.5 (23.9) 65.9 (24.3)*	n/a	63.9 (25.7)*

Note. Standard deviations are reported in parenthesis. Scores marked with an asterisk (*) are based on 22 speakers and all other scores are based on 23 speakers.

Table 3

Relationship between intelligibility scores and other measures of segmental production accuracy.

Response condition[a]	**Phonetic transcription monosyllabic words**[b]	**Phonetic transcription motor speech evaluation**[b]	**Rated articulation, speech-language pathologists**	**Rated articulation, everyday listeners**
Intelligibility orthographic	r=.97**	r=.74**	ρ =.63**	ρ =.64**
Intelligibility multiple choice	r=.98**	r=.74**	ρ =.64**	ρ =.55**

Note. Correlations are expressed as Pearson product moment (r) and Spearman rank order (ρ) coefficients.

[a] First live clinician elicitation condition, graduate student listeners.

[b] Phonetic transcription was expressed as percentage of phonetic segments produced correctly.