Single word intelligibility has been used to estimate one aspect of the magnitude of articulatory involvement in individuals with coexisting aphasia and apraxia of speech (AOS). Preliminary work indicates that word intelligibility is sensitive to production variation in this population and can be assessed with satisfactory speaker-, listener-, and test-retest reliability [1,2]. However, the parameters for valid and reliable quantification have yet to be determined.

In light of the traditional view that AOS is associated with variability across repeated attempts at similar utterances [3], it is particularly important to understand the conditions necessary for establishing test-retest reliability. The purpose of the present study was to estimate test-retest reliability in speakers with a range of AOS and aphasia severity and, if it found to be satisfactory, determine whether individual speaker and listener performance supported the overall test stability. Because recent characterizations of AOS uncomplicated by significant aphasia have suggested that errors are consistent rather than variable in location and type [4,5], it is possible that speakers' short word productions and listeners' perceptions of them are more consistent than traditionally thought. On the other hand, the typical coexistence of AOS with aphasia and the mixed constellation of both production variability and consistency in acoustic and auditory-perceptual reports across individual speakers and tasks [6, 7, etc] limit the likelihood of this scenario.

Method

Speaker Participants

Speech samples were obtained from 8 men and 5 women, who were 10 months or more post onset of a cerebrovascular accident (table 1). All speakers were diagnosed with AOS based on traditional criteria [3] through the consensus of three experienced clinicians. AOS severity was rated on a 7-point scale from 1 (least severe) to 7 (most severe). The mean severity rating was 3.1 with a representative spread from 1 to 5. The speakers had either no dysarthria or a mild unilateral upper motor neuron dysarthria and all were diagnosed with coexisting aphasia. The mean Western Aphasia Battery [8] Aphasia Quotient was 69.8, with a range from 29.3 to 94.8 and including both fluent and nonfluent aphasia types.

Speech Sample

An experimental protocol for speech intelligibility testing [9] was used to elicit 70 monosyllabic words from each speaker. Each word was presented in writing and as a live model produced by an experimenter facing the participant. The same word list was elicited twice, approximately one hour apart. Audio-recordings of each attempted word production were digitized (22 kHz sampling rate, 10 kHz low-pass filter) and stored on disk as separate files.

Listening task

Eleven graduate students in speech language pathology served as listeners. They had normal hearing and no history of speech or language impairment. The perceptual experiments were conducted individually in a sound-treated IAC booth. Productions were presented in random order from a laptop computer through headphones and responses were collected on the same laptop through custom software.

Each listener was tested over two one-hour sessions, scheduled approximately 7 days apart. During each session, they listened to 50 words, randomly selected from the first or second recording for each of the 13 speakers. To minimize listener learning effects, there was a separate random selection of utterances for each speaker. However, to enable examination of speaker consistency, the same 50 words were selected from each speaker's first and second recording. Upon hearing each production, the listener typed in the word he/she thought the speaker was attempting to produce. Only real English words were accepted as responses. The software matched listener responses to target words and computed an overall intelligibility score, accounting for homophones and normal spelling variations. The order of presentation for the speakers and the individual words were randomized for each listening session. The order of each speaker's first or second recording was randomized for each listener.

Results

Overall Intelligibility

Figure 1 shows the overall intelligibility scores for each speaker, averaged across the 11 listeners. Mean intelligibility was 71 percent for both the first and the second recording sessions, with scores ranging from 47% to 92% across individual speakers. Estimates of test-retest reliability were consistently high. The Pearson product moment correlation was .97, the point-to-point agreement within 7 percentage points was 100%, and the point-to-point agreement within 5 percentage points was 77%.

Intra-speaker production consistency

To examine each speaker's word production consistency, the perceptual analysis results for the speakers' first and second production of each target word were compared. In these comparisons, the word typed by a majority of listeners was considered the best representation of the produced utterance. Figure 2 shows the range of production consistency across speakers and the correlation between this measure and the overall speech intelligibility. When considering all utterances, the mean point-to-point intra-speaker agreement was 69%. As the figure illustrates, this measure of consistency was highly correlated with overall intelligibility ($r=.94$; $p<.05$), indicating that there was chance agreement for highly intelligible utterances. When restricting the analysis to target words where the majority of listeners heard an error on at least one of the two recording sessions, intra-speaker consistency dropped to 16% and the correlation with overall intelligibility approached zero.

Inter-Listener Agreement

As an operational definition of inter-listener consistency, 9 of the 11 listeners were required to be in agreement about the perceived word. The mean percent of target words perceived with inter-listener consistency was 70% for the first recording and 69% for the second recording. Again, there was a significant correlation between inter-listener consistency and overall speech intelligibility ($r = .92$; $p<.05$). In a second analysis, considering only productions where the majority of listeners agreed that a word different from the target was produced, inter-listener consistency was 35% and 33% for the first and second recording sessions, respectively.

Conclusions

The results replicate previous investigations by demonstrating satisfactory test-retest reliability for speech intelligibility testing in speakers with aphasia and AOS of mild to moderate severity [1,2]. An item-by-item analysis showed that most words were perceived to be identical across listeners and repeated recordings. However, further analysis showed that this agreement was inflated by accurate productions, and when the analysis focused on utterances that were not fully intelligible, listeners' agreement on the words they heard and the perceived consistency across recording sessions were both low. Thus, satisfactory test-retest reliability appears contingent on multiple repetitions and not secondary to production consistency for speakers with aphasia and AOS. This finding facilitates the generation of assessment guidelines for future use and refinement of evolving intelligibility measures.

The critical speech sample and listener group sizes are unknown and should be determined empirically. Fifty monosyllabic words and 10-11 graduate student listeners appear sufficient for an orthographic transcription application, but fewer may be sufficient, particularly if additional performance training or screening is provided for the listeners. The size requirements pose many practical challenges for efficient test administration, particularly in clinical settings, and these challenges must be addressed to ascertain clinical viability. Possible software solutions will be discussed in this regard.

References

1. Haley, K. L., Wertz, R. T., & Ohde, R. N. (1998). Single word intelligibility in aphasia and apraxia of speech. *Aphasiology, 12,* 715-730.

2. Haley, K.L., & Diakaki, A. (2002). Reliability and effectiveness of computer-mediated single word intelligibility testing in speakers with aphasia and apraxia of speech. *Journal of Medical Speech-Language Pathology, 10,* 265-269.

3. Wertz, R. T., LaPointe, L. L., & Rosenbek, J. C. (1984). *Apraxia of Speech in Adults: The Disorder and Its Management.* San Diego, CA: Allyn & Bacon.

4. McNeil, M. R., Odell, K. H., Miller, S. B., & Hunter, L. (1995). Consistency, variability, and target approximation for successive speech repetitions among apraxic, conduction aphasic, and ataxic dysarthric speakers. *Clinical Aphasiology, 23*, 39-55.

5. McNeil, M. R., Robin, D. A., & Schmidt, R. A. (1997). Apraxia of speech: Definition, differentiation, and treatment. In M. R. McNeil (Ed.). *Clinical Management of Sensorimotor Speech Disorders,* Thieme: New York.

6. Wambaugh, J. L., Nessler, C., Bennet, J., & Mauszycki, S. C. (2004). Variability in apraxia of speech: A perceptual and VOT analysis of stop consonants. *Journal of Medical Speech Language Pathology, 12*, 221-227.

7. Shuster, L., & Wambaugh, J. L. (2003). Consistency of speech sound errors in apraxia of speech accompanied by aphasia. Paper presented at the Clinical Aphasiology Conference, Orcas Island, Washington.

8. Kertesz, A. (1982). *Western Aphasia Battery*, New York: Grune & Stratton.

9. Kent, R. D., Weismer, G., Kent, J. F., and Rosenbek, J. C. (1989). Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders, 54*, 482-499.

Table 1. Participant demographics and clinical test results. MPO = Months post onset of a cerebrovascular accident; AOS = Apraxia of speech severity rating (0 no AOS, 7 most severe AOS); WAB AQ: Aphasia Quotient on the Western Aphasia Battery [8]; Aphasia = Aphasia classification according to the WAB guidelines.

	Gender	**Age**	**MPO***	**AOS***	**WAB AQ***	**Aphasia***
S01	F	52	24	5	29.3	Broca
S02	M	57	45	5	59.5	Broca
S03	M	64	11	4	73.9	Broca
S04	F	45	30	4	72.3	Broca
S05	M	47	63	1	67.5	Broca
S06	M	57	23	4	80.6	Anomic
S07	M	63	180	2	94.8	Anomic
S08	F	74	13	2	79.5	Anomic
S09	F	56	44	4	71.9	Conduction
S10	M	55	24	2	78.8	Conduction
S11	M	45	17	3	29.9	Broca
S12	F	77	24	3	82.4	Anomic
S13	M	36	10	1	86.5	Anomic
Mean		*56*	*39*	*3.1*	*69.8*	

Figure 1. Intelligibility scores derived from the first (filled bars) and the second (open bars) recording session for individual speakers.

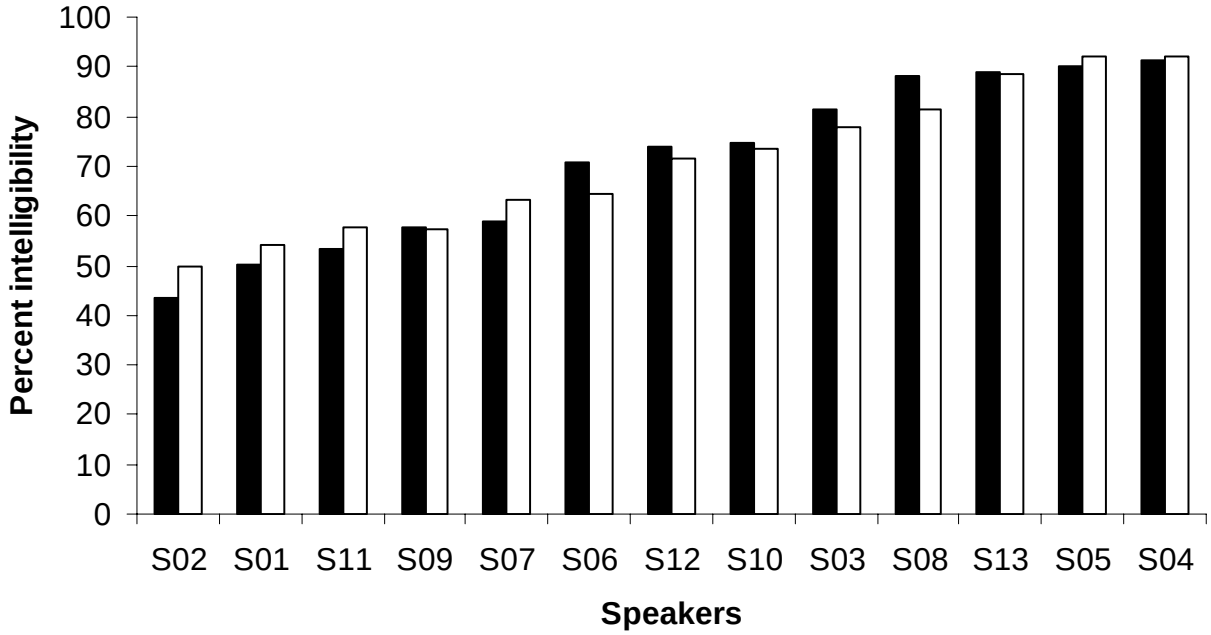

Figure 2. Correlation between intelligibility and intra-speaker production consistency. Filled diamonds represent consistency when all words were considered. Open circles represent consistency for productions where the majority of listeners heard a different word than the target word for at least one of the two recording session.