Word Retrieval Measures with the AphasiaBank Stimuli: Test-Retest Reliability
Proposal for CAC 2014

Research into treatment for improving word retrieval ability in aphasia is increasingly focused on assessing outcomes at a discourse level. For example, the AphasiaBank project (http://talkbank.org/AphasiaBank/) uses a number of tasks to elicit discourses from individuals with aphasia. The discourses can then be analyzed with a set of analysis tools from the Computerized Language Analysis (CLAN) system. MacWhinney, Fromm, Holland, Forbes, & Wright (2010) have suggested that the AphasiaBank tools can be used to study recovery from aphasia and the effects of aphasia treatments. The AphasiaBank protocol is promising because of its ability to quickly and accurately perform a number of analyses that are time-consuming, cumbersome, and vulnerable to error when performed manually. However, except for a report on VOCD, a measure of lexical diversity that is part of the CLAN system (Boyle, 2013), there have been no reports about the test-retest reliability of the various language measures included in CLAN when they are used with the elicitation stimuli that are part of the AphasiaBank protocol.

Test-retest reliability refers to the assessment of whether a test produces the same results on repeated application when the participants who are being tested have not changed on the domain that is being measured (Fitzpatrick, Davey, Buxton, & Jones, 1998). Before a measure is used as an outcome assessment, its test-retest reliability must be established, otherwise it is impossible to assert that changes on the measure are related to treatment rather than to spurious, day-to-day variability inherent in the measurement or the behavior it is measuring (Brookshire & Nicholas, 1994; Herbert, Hickin, Howard, Osborne, & Best, 2008). Test-retest reliability is as important for measures used to evaluate impairments as it is for those that measure change, since measures that are not stable will not provide valid or reliable assessments of impairments.

Several measures available in the CLAN System can be used to assess word retrieval difficulty. To use CLAN, the discourses must first be transcribed and coded for errors and other behaviors of interest using a format specified in the CHAT Manual (http://talkbank.org/AphasiaBank/). CLAN can then be used to analyze the transcripts for the occurrence of the coded errors as well as for other language parameters. Word-finding problems that can be coded in CHAT include phonemic paraphasias, semantic paraphasias, neologisms, false starts, time fillers, and repetitions. The purpose of this investigation was to provide preliminary information about the test-retest reliability of these measurements in narrative discourses elicited with the AphasiaBank stimuli from speakers with aphasia.

Method

Participants

The participants were 7 right-handed English-speaking individuals with aphasia recruited from a university clinic and a community-based aphasia center. None had other history of neurologic impairment. One (P6) had a mild apraxia of speech in addition to aphasia. Table 1 contains demographic information and Table 2 contains test results. Discourses from 2 additional participants are currently being analyzed, and data from additional participants continues to be collected.

Procedures

Discourse samples were elicited in two sessions (separated by 2 to 7 days) without intervening treatment using stimuli and procedures developed for the AphasiaBank project

(MacWhinney et al., 2011). The discourses were transcribed and coded by a trained graduate student using procedures described in the CHAT Manual (http://talkbank.org/AphasiaBank/). Transcripts and their associated videos were reviewed independently by the investigator, and all transcription and coding discrepancies were resolved by consensus. Opening and closing comments (e.g., “okay”, “that’s that”) unrelated to the task were eliminated from further analysis. CLAN commands were written so that only the participants’ utterances were analyzed. (The AphasiaBank protocol limits the examiner’s speech to a minimum.) The EVAL command was used to derive the duration of the discourse sample and the number of repetitions that occurred. The FREQ command was used to determine the number of phonemic paraphasias, semantic paraphasias, neologisms, false starts, and time fillers. To compare across discourses of different lengths, each of the measures was calculated as a proportion of time (occurrence per minute). Because sample size can affect the stability of a measure (Brookshire & Nicholas, 1994), the narrative stimuli from AphasiaBank were analyzed in different ways: all narrative tasks combined versus tasks divided by narrative sub-genre (story retell, picture sequence description, and complex picture description).

To assess the extent to which scores in the first session were related to scores in the second session, the Pearson product-moment correlation coefficient and the standard error of measurement (SEM) were calculated. Using recommendations by Fitzpatrick and colleagues (1998), a correlation value of 0.70 or above was considered adequately reliable for group research studies, and a value of 0.90 or above was considered adequately reliable for clinical decision making about individuals. To determine the minimum change necessary to ensure a confidence level of 90% that a change would not be related to measurement error, the Minimal Detectable Change (MDC) value was calculated with the formula $\text{MDC}_{90} = \text{SEM} \times \sqrt{2} \times 1.65$ (Stratford, 2004).

Results & Discussion

Table 3 contains the results. When the discourses were grouped by narrative sub-genres, only one word finding measure in only two of the three Sub-genres yielded a correlation value greater than 0.70. In the Story Retell condition, the number of semantic paraphasias produced per minute yielded a correlation value of 0.85 across the two sessions, indicating that its stability is adequate for use in group research studies. In the Complex Picture Description condition, the number of phonemic paraphasias produced per minute yielded a correlation value of 0.95, indicating that its stability is adequate for use in group research studies and for clinical decision making about individuals. An individual’s score would have to decrease by at least 0.44 phonemic paraphasias per minute in order to attribute the change to intervention or language recovery, rather than to normal variability.

When all narrative tasks were combined, three word finding measures yielded correlation values greater than 0.70. The number of time fillers produced per minute and the number of repetitions produced per minute yielded correlation values of 0.73 and 0.86, respectively, indicating that their stability is adequate for use in group research studies. The number of semantic paraphasias produced per minute yielded a correlation value of 0.95, indicating that its stability is adequate for use in group research studies and for clinical decision making about individuals. An individual’s score would have to decrease by 0.42 semantic paraphasias per minute in order to attribute the change to intervention or language recovery, rather than to normal variability.

It appears that combining the narrative tasks resulted in more stable measurements across sessions. Given the small number of mildly and moderately impaired individuals with aphasia included thus far, these results are promising, since adding participants to the analysis is likely to improve the stability of the correlations. However, other word retrieval measures which are often used to measure change in aphasic word-retrieval impairment, notably measures of phonemic paraphasias and neologisms, were not adequately stable to recommend their use even in group research studies. Discussion will focus on factors that contribute to the instability of the measures and possibilities for improving their stability.

References

Brookshire, R.H. and Nicholas, L.E. (1994). Test-retest stability of measures of connected speech in aphasia. *Clinical Aphasiology, 22,* 119-133.

Fitzpatrick, R., Davey, C., Buston, M.J., & Jones, D.R. (1998). Evaluating patient-based outcome measures for use in clinical trials. *Health Technology Assessment, 2*(14).

Herbert, R., Hickin, J., Howard, D., Osborne, F., & Best, W. (2008). Do picture-naming tests provide a valid assessment of word retrieval in conversation in aphasia? *Aphasiology, 22,* 184-203.

MacWhinney, B., Fromm, D., Forbes, M., & Holland, A. (2010). Automated analysis of the Cinderella story. *Aphasiology, 24,* 856-868.

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

Stratford, P.W. (2004). Getting more from the literature: Estimating the standard error of measurement from reliability studies. *Physiotherapy Canada, 56,* 27-30.

Table 1. Demographic information about the participants.

Participant	Age	Gender	Race	Years Education	Occupation	Months Post Stroke
P1	80	M	C	18	Social worker	12
P2	59	F	C	17	Teacher	18
P3	84	M	C	12	Police officer	24
P4	72	M	AA	13	Tile setter	162
P5	80	M	C	14	Medical technologist	27
P6	72	M	C	12	Truck driver	86
P7	51	M	C	18	Attorney	6

AA = African-American, C = Caucasian

Table 2. Results of language testing for participants.

Test	P1	P2	P3	P4	P5	P6	P7
Western Aphasia Battery							
Aphasia Quotient (max = 100)	89.6	94.8	72.4	84	77.4	68.2	90.4
Fluency (max = 10)	9	9	8	9	9	6	9
Comprehension (max = 10)	9.1	9.5	7.9	7.9	9.4	9.3	9.3
Repetition (max = 10)	9.3	10	6.2	7.7	5	8	8.4
Naming (max = 10)	7.9	8.9	8.1	8.4	9.3	5.8	10
Type	anomic	anomic	anomic	anomic	conduction	Broca’s	anomic
Boston Naming Test Short Form (max = 15)	9	15	7	11	12	5	13

Table 3. Pearson product-moment correlation coefficients (*r*), standard errors of measurement (SEM), and Minimal Detectable Change (MDC_{90}) values for performance across two sessions on the measures of word retrieval problems. To assess the influence of sample size on reliability, the AphasiaBank tasks were analyzed in two different ways: narrative sub-genres (Cinderella retell; sequence picture description (N=2); complex picture description (N=1)) versus all narrative tasks combined.

	Phonemic paraphasias per minute	**Semantic paraphasias per minute**	**Neologisms per minute**	**False starts per minute**	**Time fillers per minute**	**Repetitions per minute**
Story Retell						
r	-0.29	0.85*	-0.32	0.36	0.28	0.46
SEM	0.57	0.66	0.30	1.51	3.85	1.69
MDC_{90}	1.33	1.53	0.70	3.51	8.98	3.95
Sequence Picture Descriptions						
r	-0.31	0.49	0.49	0.12	0.63	0.34
SEM	0.78	0.35	0.26	1.01	2.62	1.93
MDC_{90}	1.81	0.81	0.61	2.36	6.11	4.51
Complex Picture Description						
r	0.95**	0.34	-0.04	0.11	0.44	0.43
SEM	0.19	2.20	1.34	0.99	2.91	1.60
MDC_{90}	0.44	5.13	3.13	2.30	6.78	3.74
All Narrative Tasks						
r	0.32	0.95**	0.60	0.01	0.73*	0.86*
SEM	0.22	0.18	0.20	0.87	1.69	0.65
MDC_{90}	0.51	0.42	0.46	2.04	3.95	1.51

* = correlations indicating sufficient stability for use in group research studies.

** = correlations indicating sufficient stability for use in group research studies and for individual clinical decision making.