Stability of a Measure of Lexica Diversity (*D*) in Narrative Discourse
Proposal for CAC 2013

Research into treatment for improving word retrieval ability in aphasia is increasingly focused on assessing outcomes at a discourse level. One of the challenges in this regard is choosing a measure to assess word retrieval in discourse. Some researchers (Fergadiotis & Wright, 2011; MacWhinney, Fromm, Forbes, & Holland, 2011; Rider, Wright, Marshall, & Page, 2008; Wright & Capiluto, 2009; Wright, Silverman, & Newhoff, 2003) have proposed using a measure of lexical diversity (*D*) as a proxy measure of word retrieval in aphasic discourse, reasoning that as word retrieval ability improves, a wider variety of words should be produced. Fegadiotis and Wright (2011) define lexical diversity as the range of vocabulary deployed in a discourse sample by a speaker, reflecting the speaker's capacity to access and retrieve target words. *D* is a measure of lexical diversity that is robust to length variation, allowing comparison of discourses over time or between participants (MacWhinney et al., 2011).

One concern about using a measure is its session-to-session stability when it is applied to the discourse of people with aphasia. Bennett and Miller (2010) asserted that reliability of measurements forms the foundation of any scientific enterprise, and noted that reliability varies depending on the measure being used and the thing being measured. Herbert and colleagues (Herbert, Hickin, Howard, Osborne, & Best, 2008) stated that establishing stability in a measure is an essential prerequisite to its use as an outcome assessment for the evaluation of therapy. Brookshire and Nicholas (1994) cautioned that without knowing the stability of the outcome measures we use, "spurious differences generated by test-retest instability may be misconstrued as the effects of treatment" (p.129). Thus, we need information about the test-retest stability of an outcome measure in order to make appropriate decisions about its use as a valid metric of treatment effects. Equally important, the test-retest stability of a measure is important if it is used to describe and analyze aspects of an individual's language impairment. A measure that is not reasonably stable from session to session will not provide a valid, reliable assessment of an individual's impairment.

To date, there are no published reports that investigate the test-retest reliability of *D* in people with aphasia. Information about the stability of this measure (without intervening treatment) is essential before it can be used as a valid and reliable assessment of word retrieval abilities or of treatment-related changes in aphasic speakers. The aim of this investigation was to provide preliminary information about the reliability of *D* in narrative discourse production of aphasic speakers.

Method

Participants

The participants were 7 right-handed English-speaking aphasic individuals 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.

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 (MacWhinney et al., 2011). The discourses were transcribed and coded using procedures

developed by MacWhinney and colleagues (2011). *D* was calculated with the voc-*D* program in CLAN using a command code developed by MacWhinney and colleagues to examine lemmas (i.e., inflected forms of the same base are treated as the same lexical item) and eliminate false starts, neologisms, and other aphasic errors. 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.

Results & Discussion

The Pearson product-moment correlation coefficient was 0.84, representing a strong relationship between the values obtained for *D* in the two sessions. This suggests that *D* was sufficiently stable across two separate sessions to serve as a reliable assessment of lexical diversity for this group of aphasic individuals. The standard error of measurement was 6.66, which is acceptably stable for groups of scores with means of 56. Examination of individual participant scores in sessions 1 and 2, however, reveals that only 3 (P2, P5, and P6) of the 7 participants had difference scores that were within the SEM of 6.65 (Table 3). The remaining 4 participants showed changes in their lexical diversity scores that exceeded the SEM across two sessions with no intervening treatment. Thus, investigators who use *D* as an outcome measure in single-subject treatment designs should demonstrate its stability (or variability) across several baseline sessions before implementing treatment in order to make valid conclusions about any treatment-related changes.

References

Bennett, C.M. & Miller, M.B., (2010). How reliable are the results from magnetic resonance imaging? *Annals of the New York Academy of Sciences, 1191*, 133-155.

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

Fergadiotis, G. & Wright, H.H. (2011). Lexical diversity for adults with and without aphasia across discourse elicitation tasks. *Aphasiology, 25*, 1414-1430.

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

Rider, J.D., Wright, H.H., Marshall, R.C., & Page, J.L. (2008). Using semantic feature analysis to improve contextual discourse in adults with aphasia. *Aphasiology*, *17*, 161-172.

Wright, H.H. & Capilouto, G.J. (2009). Manipulating task instructions to change narrative discourse performance. *Aphasiology*, *23,* 1295-1308.

Wright, H.H., Silverman, S.W. & Newhoff, M. (2003). Measures of lexical diversity in aphasia. *Aphasiology*, *17*, 443-452.

Table 1. Demographic information about the participants.

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

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. Participants *D* scores in sessions 1 and 2, Pearson product-moment correlations and standard error of measurement (SEM) for the group data, and difference scores for each participant between sessions 1 and 2.

Participant	Session 1	Session 2	Difference (Session 2 – Session 1)
P1	54.50	43.37	-11.13
P2	50.25	43.62	-6.63
P3	49.67	39.63	-10.04
P4	45.18	67.23	22.05
P5	69.00	66.33	-2.67
P6	39.36	38.11	-1.25
P7	87.57	96.50	8.93
Mean	56.60	56.40	
Standard Deviation	16.50	21.50	
Pearson *r*	0.84		
SEM	6.66		