AJSLP

Supplement Article

Stability of Word-Retrieval Errors With the AphasiaBank Stimuli

Mary Boyle[a]

Purpose: This study examined the test–retest reliability of select measures of word-retrieval errors in narrative discourses of individuals with aphasia assessed with the AphasiaBank stimuli.
Method: Ten participants with aphasia were video recorded during 2 sessions producing narratives elicited with pictures. Discourses were transcribed and coded using AphasiaBank procedures, then analyzed for the stability of rates of phonological errors, semantic errors, false starts, time fillers, and repetitions per minute. Values for correlation coefficients and the minimal detectable change score were used to assess stability for research and clinical decision making.
Results: There was poor test–retest reliability when the discourses were analyzed by each narrative subgenre. When the narrative discourses were combined for analysis, several measures appeared to be sufficiently stable across sessions for use in group studies, and 1 could be adequately stable for making clinical decisions about an individual.
Conclusions: Because the short speech samples yielded by the subgenre narrative analyses demonstrated poor test–retest reliability, it is recommended that all of the picture-based narrative discourse tasks be combined for analysis of word-retrieval impairments when the AphasiaBank stimuli are used. However, the confidence intervals associated with the reliability coefficients obtained in this study suggest caution in using the measures if they are based on performance in a single session. More investigations of the test–retest reliability of measures used to study language impairment in discourse contexts are essential.

Research into treatment for improving word-retrieval ability in aphasia is increasingly focused on assessing outcomes at a discourse level (Andreetta, Cantagallo, & Marini, 2012; Boyle, 2014; Fergadiotis, Wright, & West, 2013; MacWhinney, Fromm, Forbes, & Holland, 2011; Marini, Andreetta, del Tin, & Carlomagno, 2011). 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, and Wright (2010) have suggested that the AphasiaBank tools can be used to study recovery from aphasia and the effects of aphasia treatments. The 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, 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). It is important when evaluating impairments because measures that are not stable will not provide valid or reliable assessments of impairments. Furthermore, before a measure is used as an outcome assessment, its test–retest reliability must be established in order to provide confidence that changes on the measure are related to treatment or recovery rather than to the spurious, day-to-day variability inherent either in the instrument itself or in the behavior that it is measuring (Brookshire & Nicholas, 1994a, 1994b; Herbert, Hickin, Howard, Osborne, & Best, 2008). The ability to retrieve words can be affected by such factors as fatigue (Diamond, Johnson, Kaufman, & Graves, 2008), so it seems particularly important to attend to the test–retest reliability of the measures that are used to assess it.

Although procedures that have been used to assess aspects of successful word retrieval in connected speech have been tested for such factors as equivalence of alternate forms (Brookshire & Nicholas, 1994a; Doyle et al., 2000),

[a]Montclair State University, Bloomfield, NJ
Correspondence to Mary Boyle: boylem@mail.montclair.edu

Editor: Anastasia Raymer
Associate Editor: William Hula
Received September 15, 2014
Revision received February 7, 2015
Accepted June 2, 2015
DOI: 10.1044/2015_AJSLP-14-0152

Disclosure: *The author has declared that no competing interests existed at the time of publication.*

concurrent validity (McNeil, Doyle, Fossett, Park, & Goda, 2001; McNeil et al., 2007), and interjudge reliability (Boyle, 2014; Brookshire & Nicholas, 1994a, 1994b; Herbert, Best, Hickin, Howard, & Osborne, 2003; Hula, McNeil, Doyle, Rubinsky, & Fossett, 2003; McNeil et al., 2001, 2007; Nicholas & Brookshire, 1993, 1995), few have been examined for test–retest reliability (Boyle, 2014; Brookshire & Nicholas, 1994a, 1994b; Herbert et al., 2013). When test–retest reliability has been assessed, stability of the measurements has varied considerably. For example, Boyle (2014) reported correlation coefficients that ranged from .23 for a measure of inaccurate main concepts to .99 for the number of words produced per minute. Results reported by Brookshire and Nicholas (1994b) ranged from .32 for the percentage of words that were correct information units when only one stimulus was used to elicit the discourse, to .98 for the number of words produced per minute and for the percentage of words that were correct information units when five or 10 stimulus items were used. When discussing their results, Brookshire and Nicholas (1994b) noted that the test–retest stability of the measures in general increased as the size of the speech samples increased.

Several investigators have examined behaviors that signal word-finding difficulty in connected speech. Brookshire and Nicholas (1995) examined productions that they called performance deviations, which included part-word and unintelligible productions, nonword fillers, inaccurate words, false starts, unnecessary exact repetitions, nonspecific or vague words, fillers, the word "and," and off-task or irrelevant comments. Doyle et al. (2000) examined discourses elicited with a story retelling procedure for the number of mazes produced per minute, the percentage of all syllables that contained sound production errors, and the number of silent pauses per minute. McNeil et al. (2007) examined the percentage of total words that were in mazes and the number of abandoned sentences in story retells. Most recently, Boyle (2014) used the discourse elicitation procedures developed and standardized by Nicholas and Brookshire (1993, 1995) to examine the percentage of T-units that contained verbal paraphasias, initial sounds, phonemic paraphasias, neologisms, repetitions, reformulations, empty words, time fillers, delays, and comments on the task. Although all of these investigations reported on interjudge reliability for these behaviors, only the study by Boyle (2014) reported test–retest reliability. Correlation coefficients ranged from .04 for the percentage of T-units that contained neologisms to .97 for the percentage of T-units that contained delays longer than 6 s. Given this wide range of session-to-session variability, it seems prudent to examine the stability of word-retrieval errors in connected speech with other discourse elicitation procedures before using those procedures to assess impairment or to measure change related to treatment or recovery.

Several measures available in the CLAN system can be used to assess word-retrieval difficulty. Discourses must first be transcribed and coded for errors and other behaviors of interest using a format specified in the Codes for the Human Analysis of Transcripts (CHAT) Manual (MacWhinney, 2000; 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 phonological errors, semantic errors, false starts (called phonological fragments in the AphasiaBank coding system), time fillers, and repetitions. These word-finding problems, proposed for use in AphasiaBank by a group of "20 senior aphasia researchers" (MacWhinney, Fromm, Forbes, & Holland, 2011, p. 1287) who collaborated in the development of the elicitation stimuli and the computational analysis of the discourses, were chosen as the focus of this study because of their similarity to measures of word-retrieval difficulty in discourse that have been studied previously with different discourse elicitation protocols and analysis methods (Boyle, 2014; Brookshire & Nicholas, 1995; Doyle et al., 2000). In this study, the AphasiaBank measures of word-retrieval difficulty were calculated for occurrences per minute rather than as raw counts of occurrence for several reasons: (a) Nicholas and Brookshire (1993) found that calculated measures of successful word retrieval, such as words per minute and correct information units per minute, were more stable across sessions than raw count measures, such as number of words and number of correct information units; (b) several of the measures of word-retrieval difficulty investigated in earlier studies used calculated measures, such as the number of mazes per minute and the number of silent pauses per minute (Doyle et al., 2000); and (c) using the calculated measure allowed comparison across discourses of different lengths that were obtained in two different sessions. The purpose of this investigation was to provide preliminary information about the test–retest reliability of these measurements of word-retrieval difficulty in narrative discourses elicited with the AphasiaBank stimuli from speakers with aphasia.

Method

The methodology used in this investigation was approved by the Institutional Review Board of Montclair State University. All participants provided signed informed consent.

Participants

Ten right-handed English-speaking individuals with aphasia who were recruited from a university clinic and a community-based aphasia center participated in this study. Aphasia was secondary to a single, left-hemisphere cerebrovascular accident for all participants except for P8, whose aphasia was secondary to a traumatic brain injury. No participants had a history of other neurologic impairment. Two participants (P6 and P9) had a mild apraxia of speech in addition to aphasia, assessed by methods and criteria described by Duffy (2013). All participants were screened to ensure that they had adequate vision and hearing to perform the tasks.

Table 1 contains demographic information and test results. Ages ranged from 26 to 84 years (M = 64.5; SD = 18.6) and time poststroke ranged from 6 to 240 months (M = 63.3; SD = 78.2). All participants lived in their communities independently or with a family member. Two of the participants were African American (one woman and one man) and eight were White (three women and five men). All had completed high school and seven completed some postsecondary education (years of education, M = 14.8, SD = 2.5). Aphasia severity, assessed by the Aphasia Quotient from the Western Aphasia Battery–Revised (WAB-R; Kertesz, 2006), ranged from mild to moderate. Five participants had anomic aphasia, three had conduction aphasia, and two had Broca's aphasia. Naming impairment ranged from mild to moderate on the naming subtests of the WAB-R and the short form of the Boston Naming Test–Second Edition (Kaplan, Goodglass, & Weintraub, 2000).

Procedures

Stimuli and Elicitation Tasks

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). This article is focused on analysis of the picture-description narratives and the story retell narratives. For the picture descriptions, participants look at black-and-white drawings and tell a story about each with a beginning, middle, and end. The stimuli for these tasks may be viewed at http://talkbank.org/AphasiaBank/protocol/pictures. They consist of two picture sequences and one complex picture. One of the picture sequences shows a child kicking a soccer ball and breaking a window. The other picture sequence shows a child refusing an umbrella and getting caught in the rain. The complex picture shows a cat stuck in a tree and the aftermath of a man's attempt to rescue it. For the story retell task, participants look through a picture book of the Cinderella story that has the words covered; then the book is removed and the participants tell the story. All tasks were video recorded.

Data Analysis

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 author. Point-to-point interrater agreement exceeded 93% for transcribed words and exceeded 80% for each error code. All transcription and coding discrepancies were resolved by consensus discussion while reviewing the transcripts and the video recording. Opening and closing comments (e.g., "okay," "that's that") unrelated to the task were eliminated from further analysis. CLAN commands (MacWhinney, 2000) were written so that only the participants' utterances were analyzed. 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 phonological errors, semantic errors, false starts, and time fillers (i.e., "uh" and "um"). Operational definitions for the behaviors signaling word-retrieval difficulty can be found in the CHAT Manual (http://talkbank.org/AphasiaBank), and are summarized in the Appendix. 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, 1994b), the narrative discourses were analyzed in different ways: all narrative tasks combined versus tasks divided by narrative subgenre (picture-sequence narratives, complex-picture narratives, and story retell narratives).

Assessment of Session-to-Session Stability

To assess the extent to which scores in the first session were related to scores in the second session, the intraclass correlation coefficients (ICC; Cicchetti, 1994) were calculated using a one-way random model with measures of absolute agreement. Using recommendations by Fitzpatrick et al.

Table 1. Participants' demographic information and test results.

Variable	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Age	80	59	84	72	80	72	51	46	26	75
Gender	M	F	M	M	M	M	M	F	F	F
Race	W	W	W	AA	W	W	W	W	AA	W
Education (years)	18	17	12	13	14	12	18	12	16	16
MPO	12	18	24	162	27	86	6	240	15	43
WAB-R–AQ (100)	89.6	94.8	72.4	84	77.4	68.2	90.4	89	61.4	60.5
Fluency (10)	9	9	8	9	9	6	9	9	4	6
Comprehension (10)	9.1	9.5	7.9	7.9	9.4	9.3	9.3	9.1	6.3	9.75
Repetition (10)	9.3	10	6.2	7.7	5	7.9	8.4	8.8	5.6	2.5
Naming (10)	7.9	8.9	8.1	8.4	9.3	5.8	10	8.6	5.8	5
Aphasia type	Anomic	Anomic	Conduction	Anomic	Conduction	Broca's	Anomic	Anomic	Broca's	Conduction
BNT (15)	9	15	7	11	12	5	13	6	7	4

Note. AA = African American, W = White, MPO = months postonset, WAB-R–AQ = Western Aphasia Battery-Revised–Aphasia Quotient (Kertesz, 2006), BNT = Boston Naming Test–Second Edition (Kaplan, Goodglass, & Weintraub, 2000).

(1998), a correlation value of .70 or above was considered adequately reliable for group studies, and a value of .90 or above was considered adequately reliable for clinical decision making about individuals. To assess how accurately scores from session 1 could predict scores from subsequent sessions, the standard errors of measurement (*SEM*) were calculated with the formula $SEM = SD\sqrt{1 - r}$, where *SD* is the standard deviation for the obtained score distribution and *r* is the correlation coefficient. The standard error of measurement provides information that can be used to determine the range of scores likely to include an individual's true score. 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) was calculated with the formula $MDC_{90} = SEM \times \sqrt{2} \times 1.65$ (Stratford, 2004).

Results

Table 2 contains the range of the discourse sample lengths in minutes for each subgenre (picture-sequence narratives, complex-picture narratives, and story retell narratives) and for all narrative tasks combined. Each of the subgenre narrative types contained narratives that were under or just over a minute, whereas the shortest combined narrative tasks were 4 to 6 min long.

Table 3 contains the means, standard deviations, ICCs and associated 90% confidence intervals, standard errors of measurement, and minimal detectable changes for the discourse tasks analyzed by narrative subgenre (picture-sequence narratives, complex-picture narratives, and story retell narratives) and analyzed with all narrative tasks combined. Correlated *t* tests (two-tailed, $df = 9$, $\alpha \leq .05$) revealed that there were no statistically significant differences between the pairs of means for sessions 1 and 2 for any of the word-retrieval difficulty measures for any of the narrative subgenres or when all narrative tasks were combined. When the discourses were analyzed by each narrative subgenre, no more than one of the measures in each subgenre yielded a correlation value greater than .70, the criterion for stability in group studies. In the picture-sequence narrative condition, none of the measures met the criterion. In the complex-picture narrative condition, the number of phonological errors produced per minute yielded a correlation value of .95, indicating that its stability may be adequate for use in group research studies and for clinical decision making about individuals. However, the 95% confidence interval associated with this estimate suggests that plausible values range from .81 to .99. Thus, one can be very confident that the number of phonological errors produced per minute obtained from descriptions of the complex pictures is stable enough for use in group research studies, but it should be used with caution to make decisions about an individual's performance. If this metric were to be applied to an individual, he or she would have to decrease the rate of phonological errors by at least 0.56 per minute in order to attribute the change to intervention or language recovery rather than to normal variability. In the story retell condition, the number of semantic errors produced per minute yielded a correlation value of .94, indicating that its stability may be adequate for use in group research studies and for clinical decision making about individuals. The 95% confidence interval suggests that plausible values range from .77 to .98, providing confidence that the number of semantic errors produced per minute obtained from a retell of the Cinderella story is stable enough for group research studies but suggesting caution in using it for making decisions about an individual. An individual would have to decrease the rate of semantic errors by at least 1.35 per minute in order to attribute the change to intervention or language recovery.

Table 2. Range of discourse sample lengths in minutes for each subgenre and for the combined narratives across participants in each session.

	Range of sample length in minutes	
Subgenre	**Time 1**	**Time 2**
Sequence-picture narratives	1.15 to 6.30	0.87 to 12.33
Complex-picture narratives	0.42 to 2.97	0.43 to 3.42
Story retell narratives	0.73 to 6.38	1.10 to 6.23
Combined narratives	5.78 to 21.08	4.22 to 24.87

When the narrative discourse tasks were combined, four word-retrieval measures yielded correlation values greater than .70. The number of phonological errors produced per minute, the number of time fillers produced per minute, and the number of repetitions produced per minute yielded correlation values of .84 (with a 95% confidence interval from .50 to .96), .76 (confidence interval = .32 to .93), and .81 (confidence interval = .44 to .95), respectively, indicating that their stability might be adequate for use in group research studies or in making decisions about an individual. The ranges of the confidence intervals suggest that these metrics should be used very cautiously for either purpose. The number of semantic errors produced per minute yielded a correlation value of .96 with a 95% confidence interval from .85 to .99, indicating that its stability is adequate for use in group research studies and that it might be used, with caution, for clinical decision making about individuals. An individual would have to decrease the rate of semantic errors by at least 0.60 per minute in order to attribute the change to intervention or language recovery.

Discussion

The purpose of this study was to examine the test–retest stability of select measures of word-retrieval errors in the discourses of individuals with aphasia as assessed with the stimuli and procedures of AphasiaBank. Test–retest reliability coefficients are not fixed characteristics of a test, but are estimates that may vary depending on the sample being tested. Thus, readers should bear in mind that this discussion is based on results obtained with a single small

Table 3. Means (*M*), standard deviations (*SD*), intraclass correlation coefficients (ICC) with associated 95% confidence intervals, standard errors of measurement (*SEM*), and minimal detectable change scores with a 90% confidence interval (MDC_{90}) for participants' performance on the measures of word retrieval in discourse.

	PE/min		SE/min		FS/min		TF/min		Rep/min	
Sequence-picture narratives	**S1**	**S2**	**S1**	**S2**	**S1**	**S2**	**S1**	**S2**	**S1**	**S2**
M	1.07	0.66	0.94	1.70	1.88	1.90	6.94	3.85	2.29	1.34
SD	1.02	0.57	0.88	1.09	1.08	1.43	4.2	1.98	2.19	0.87
ICC (95% confidence intervals for ICCs)	.64 (.10 to .90)		.52 (–.09 to .85)		.20 (–.43 to .72)		.10 (–.51 to .66)		.18 (–.45 to .70)	
SEM	0.61		0.61		0.97		9.44		1.98	
MDC_{90}	1.43		1.42		2.26		22.02		4.62	
Complex-picture narratives										
M	0.07	0.80	2.74	2.33	2.23	2.74	4.20	5.49	2.77	3.57
SD	0.05	0.62	2.43	1.73	1.04	1.11	3.62	5.47	1.80	3.47
ICC (95% confidence intervals for ICCs)	.95[b] (.81 to .99)		.47 (–.16 to .83)		–.08 (–.63 to .55)		.40 (–.23 to .81)		.31 (–.33 to .77)	
SEM	0.24		1.77		1.09		2.81		1.49	
MDC_{90}	.56		4.13		2.53		6.55		3.48	
Story retell narratives										
M	0.67	0.59	2.21	1.92	3.21	5.59	6.11	6.45	3.23	3.07
SD	0.32	0.43	1.65	1.02	1.81	5.47	3.98	2.81	2.21	2.32
ICC (95% confidence intervals for ICCs)	.56 (–.05 to .86)		.94[b] (.77 to .98)		.22 (–.42 to .72)		.44 (–.20 to .82)		.40 (–.24 to .80)	
SEM	0.58		0.58		1.60		2.98		1.71	
MDC_{90}	1.37		1.35		3.72		6.95		4.00	
Combined narratives										
M	0.55	0.48	1.25	1.36	1.75	2.27	3.60	4.14	1.96	1.77
SD	0.27	0.14	0.94	0.99	0.88	1.79	2.91	2.56	1.54	1.21
ICC (95% confidence intervals for ICCs)	.84[a] (.50 to .96)		.96[b] (.85 to .99)		.34 (–.31 to .78)		.76[a] (.32 to .93)		.81[a] (.44 to .95)	
SEM	0.27		0.26		0.71		1.42		0.67	
MDC_{90}	0.63		0.60		1.66		3.32		1.57	

Note. PE/min = phonological errors per minute, SE/min = semantic errors per minute, FS/min = false starts per minute, TF = time fillers per minute, Rep/min = repetitions per minute, S1 = session 1, S2 = session 2.

[a]Adequate stability for group studies. [b]Adequate stability for clinical decision making about individuals; the MDC_{90} value is the amount of change required to be confident that the change is not the result of the measurement's variability.

sample of participants. In addition, the criteria for choosing reliability coefficients that are considered sufficiently stable for group research studies or for making clinical decisions about an individual should not be considered as invariable cutoff values, because test–retest reliability coefficients may vary based on factors such as the time between test administrations and the behavior that is being studied. With these caveats in mind, when the combined narrative discourses were analyzed, four word-retrieval measurements were sufficiently stable across sessions for use in group studies or, in one case, for making clinical decisions about an individual, although the confidence intervals associated with these results suggest that using the measures should be done very cautiously. The more stable measurements in the combined discourse condition included phonological and semantic errors, which are frequently analyzed by researchers who are interested in measuring changes in word retrieval. In contrast, none of the individual subgenre analyses yielded stable results for both of these frequently analyzed word-finding problems. Consequently, it does not seem prudent to make decisions about word-retrieval problems based on any one of the subgenres of AphasiaBank picture tasks alone. The decreased stability that was found in the subgenre analyses was probably due to the shorter discourse samples that they yielded. Brookshire and Nicholas (1994a) cautioned against making decisions on the basis of short speech samples, because they can be highly unstable from one session to the next, and stated that longer speech samples are more likely to have adequate test–retest reliability. That appears to be the case with the AphasiaBank's picture-based narrative discourse tasks.

Researchers who want to conduct a research study with a group of participants, and who want to take only one pretreatment measure of word-finding errors, need to use measures that have good test–retest reliability. In this study, the number of phonological errors per minute, semantic errors per minute, time fillers per minute, and repetitions per minute met this criterion when picture description and story retell tasks of the AphasiaBank stimuli were combined and analyzed together. Clinicians who are providing treatment and who don't have the luxury of collecting multiple pretreatment baseline measures could consider measuring changes in semantic errors per minute, as this measure was stable enough to base decisions about changes in an individual's word retrieval performance from a single administration of the combined narratives. However, because the confidence intervals associated with the measures were wide and indicated that the true correlations for the measures might be below those deemed adequately stable for group studies (.70) or for making decisions about an individual's performance (.90), caution must be recommended. The most prudent course of action for researchers and clinicians

would be to obtain narrative samples on more than a single day, without intervening treatment, so that the variability associated with each particular sample or client can be assessed and accounted for when evaluating change in performance.

Sources of Variability

Bennett and Miller (2010) stated that test–retest stability can vary depending on the measure being used, the thing being measured, and day-to-day variations in the participant's physiologic and cognitive states. In this study, the measures being used included the AphasiaBank protocol and the various manifestations of word retrieval problems. The thing being measured, word retrieval in discourse, is one of the sources of "normal" day-to-day variability. Discourse can vary from one day to another in lexical selection without compromising the accuracy of the information that is conveyed. These day-to-day changes may influence the errors that are produced. For example, if a person retrieves more items from the same semantic category during discourse production on one day than another, it might cause more semantic errors in that category, because greater semantic interference would be present (Oppenheim, Dell, & Schwartz, 2010). Similar to this, it might be that using more phonologically complex words on one day than another might make that day's discourse more susceptible to phonemic paraphasias (Brookshire, 2007; Goodglass, Kaplan, Weintraub, & Ackerman, 1976). Day-to-day variations in physiologic and cognitive states are another component of normal variability (Bennett & Miller, 2010). They could involve differences in hormone levels, heart rate, blood pressure, attention, anxiety, and other physiological variations. Bennett and Miller noted that tasks involving higher cognition had lower test–retest reliability than motor and sensory tasks in fMRI studies of participants without neurological impairment, and that test–retest reliability was even lower in participants with clinical disorders. They likened test–retest methodology involving human beings to attempts to hit a moving target. Given the variability that is related to sources that are difficult to control, such as the freedom of word choice inherent in discourse and the physiologic and cognitive states of an individual, it seems important that we try to use measures that have good test–retest stability.

Future Directions

Ten participants were included in this study, representing a small sample with a limited range of severity and aphasia types. In terms of severity, it would be useful to investigate whether test–retest reliability is affected differently in mild aphasia versus moderate aphasia. It would also be informative to ascertain whether discourses produced by people with more severe aphasia are more or less stable across sessions than those with milder impairments. Harvill (1991) noted that the standard error of measurement (and therefore, presumably, such measures as the MDC that are derived from it) are not the same for all score levels of a test. Scores at either extreme in a range of scores might have a lower standard error of measurement than scores in the middle of the range, meaning that smaller raw score changes at either extreme of the range might indicate a true change in score whereas they would not indicate a true change for middle-range scores (Harvill, 1991). Examining this possibility in groups of people with milder and more severe aphasia could inform researchers and clinicians about whether these measures might be useful with such individuals.

Despite the limitations of the small sample size, this study provides useful information. The AphasiaBank database continues to grow and now contains discourse samples from hundreds of people with aphasia. As we begin to mine this databank it is important for us to be aware that we have limited information about the test–retest reliability of many of the language behaviors that we can measure with the protocol. This study provides preliminary information that some of the behaviors that can be studied with the protocol may be sufficiently stable to draw sound conclusions, but perhaps others are not. More work regarding test–retest reliability is essential so that AphasiaBank's potential and, indeed, the potential of studying language impairment in discourse contexts, can be fulfilled.

Acknowledgments

This work was supported by a Separately Budgeted Research Award from Montclair State University. Sincere gratitude and appreciation are extended to the people with aphasia who participated and to the students who assisted in various stages of this project—namely, Keli Meyer, Ashley Leeshock, Elizabeth Kline, Julie Irwin, Kortney Babington, Melissa Burnham, and Jasmine Wallace.

References

Andreetta, S., Cantagallo, A., & Marini, A. (2012). Narrative discourse in anomic aphasia. *Neuropsychologia, 50,* 1787–1793.

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

Boyle, M. (2014) Test–retest stability of word retrieval in aphasic discourse. *Journal of Speech, Language, and Hearing Research, 57,* 966–978.

Brookshire, R. H. (2007). *Introduction to neurogenic communication disorders* (7th ed.). St. Louis, MO: Mosby.

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

Brookshire, R. H., & Nicholas, L. E. (1994b). Speech sample size and test–retest stability of connected speech measures for adults with aphasia. *Journal of Speech and Hearing Research, 37,* 399–407.

Brookshire, R. H., & Nicholas, L. E. (1995). Performance deviations in the connected speech of adults with no brain damage and adults with aphasia. *American Journal of Speech-Language Pathology, 4*(4), 118–123.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6,* 284–290.

Diamond, B. J., Johnson, S. K., Kaufman, M., & Graves, L. (2008). Relationships between information processing, depression, fatigue and cognition in multiple sclerosis. *Archives of Clinical Neuropsychology, 23,* 189–199.

Doyle, P. J., McNeil, M. R., Park, G., Goda, A., Rubenstein, E., Spencer, K., … Szwarc, L. (2000). Linguistic validation of four parallel forms of a story retelling procedure. *Aphasiology, 14,* 537–549.

Duffy, J. R. (2013). *Motor speech disorders: Substrates, differential diagnosis, and management* (3rd ed.). St. Louis, MO: Elsevier.

Fergadiotis, G., Wright, H. H., & West, T. M. (2013). Measuring lexical diversity in narratives of people with aphasia. *American Journal of Speech-Language Pathology, 22,* S397–S408.

Fitzpatrick, R., Davey, C., Buxton, M. J., & Jones, D. R. (1998). Evaluating patient-based outcome measures for use in clinical trials. *Health Technology Assessment, 2*(14), 1–74.

Goodglass, H., Kaplan, E., Weintraub, S., & Ackerman, N. (1976). The "tip-of-the-tongue" phenomenon in aphasia. *Cortex, 12,* 145–153.

Harvill, L. M. (1991). Standard error of measurement. *Educational measurement: Issues and practice*. *10,* 33–41.

Herbert, R., Best, W., Hickin, J., Howard, D., & Osborne, F. (2003). Combining lexical and interactional approaches to therapy for word finding deficits in aphasia. *Aphasiology, 17,* 1163–1186.

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

Hula, W. D., McNeil, M. R., Doyle, P. J., Rubinsky, H. J., & Fossett, T. R. D. (2003). Inter-rater reliability of the story retell procedure. *Aphasiology, 17,* 523–528.

Kaplan, E., Goodglass, H., & Weintraub, S. (2000). *Boston Naming Test–Second Edition*. Austin, TX: Pro-Ed.

Kertesz, A. (2006). *Western Aphasia Battery–Revised*. San Antonio, TX: Pearson.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ: Erlbaum.

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

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

Marini, A., Andreetta, S., del Tin, S., & Carlomagno, S. (2011). A multi-level approach to the analysis of narrative language in aphasia. *Aphasiology, 25,* 1372–1392.

McNeil, M. R., Doyle, P. J., Fossett, T. R. D., Park, G. H., & Goda, A. J. (2001). Reliability and concurrent validity of the information unit scoring metric for the story retelling procedure. *Aphasiology, 15,* 991–1006.

McNeil, M. R., Sung, J. E., Yang, D., Pratt, S. R., Fossett, T. R. D., Doyle, P. J., & Pavelko, S. (2007). Comparing connected language elicitation procedures in persons with aphasia: Concurrent validation of the story retell procedure. *Aphasiology, 21,* 775–790.

Nicholas, L. E., & Brookshire, R. H. (1993). A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *Journal of Speech and Hearing Research, 36,* 338–350.

Nicholas, L. E., & Brookshire, R. H. (1995). Presence, completeness, and accuracy of main concepts in the connected speech of non-brain-damaged adults and adults with aphasia. *Journal of Speech and Hearing Research, 38,* 145–156.

Oppenheim, G. M., Dell, G. S., & Schwartz, M. F. (2012). The dark side of incremental learning: A model of cumulative semantic interference during lexical access in speech production. *Cognition, 114,* 227–252.

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

Appendix

Operational Definitions of Behaviors Indicating Word-Retrieval Difficulty

Repetition

Words or phrases that are repeated without change are included in this error category. Repetitions that were clearly produced for rhetorical effect were not coded as errors.

Phonological Errors

For one-syllable CVC words, the error word must match the target on two elements (e.g., the same initial consonant plus the same vowel, the same vowel plus the same final consonant) and the remaining element may be a substitution, an addition, or an omission.

For one-syllable CV or VC target words, the absence of an initial consonant or a final consonant in the error word would also count as a match.

For multisyllabic words, the error must have complete matches on all but one syllable, and the error syllable must meet the requirements for one-syllable words stated above.

In this study, common colloquial forms or dialectal variations were not counted as phonological errors.

Semantic Errors

The error is a real word that may be semantically related or unrelated to the target, or a real word for which the target is not known.

Note, however, that if an error qualifies as both phonological and semantic, it receives codes for both errors. Morphological and part-of-speech errors are not scored as semantic errors.[a]

False Starts

This category includes partial productions of a word or of an attempt at a word; a phonological fragment.

Time Fillers

This category includes the production of "uh," "um," or similar nonword verbalizations; filled pauses.

[a]The CLAN system is capable of analyzing many kinds of errors other than word-retrieved errors. The complete scoring system is available at https://www.talkbank.org/AphasiaBank.