JSLHR

Research Article

Test–Retest Stability of Word Retrieval in Aphasic Discourse

Mary Boyle[a]

Purpose: This study examined the test–retest stability of select word-retrieval measures in the discourses of people with aphasia who completed a 5-stimulus discourse task.
Method: Discourse samples across 3 sessions from 12 individuals with aphasia were analyzed for the stability of measures of informativeness, efficiency, main concepts, noun and verb retrieval, word-finding difficulty, and lexical diversity. Values for correlation coefficients and the minimal detectable change score were used to assess stability for research and clinical decision making.
Results: Measures stable enough to use in group research studies included the number of words; the number of correct information units (CIUs); the number of accurate-complete, accurate-incomplete, and absent main concepts; the percentage of T-units that had word-finding behaviors of any kind; the percentage of T-units that contained empty words; and a lexical diversity measure. Words per minute, CIUs per minute, and the percentage of T-units that contained time fillers or delays were sufficiently stable to use when making clinical decisions about an individual.
Conclusion: Although several of the measures demonstrated acceptable stability for group research studies, relatively few were sufficiently stable for making clinical decisions about individuals on the basis of a single administration.

Key Words: aphasia, anomia, discourse analysis

Clinicians and researchers typically assess word-retrieval impairments in aphasia at the single-word level, whether using a comprehensive assessment like the Western Aphasia Battery (Kertesz, 1982) or a focused test of naming like the Boston Naming Test (Kaplan, Goodglass, & Weintraub, 2001) or the Test of Adolescent/Adult Word Finding (German, 1990). There is evidence, however, that aphasic anomia manifests differently in confrontation naming contexts than in connected speech contexts. For example, Williams and Canter (1982, 1987) reported that people with aphasia (PWA) performed differently on single-word confrontation naming tasks than on picture description tasks, and Mayer and Murray (2003) and Pashek and Tompkins (2002) reported that lexical retrieval in connected speech tasks was superior to that in single-word confrontation-naming tasks for individuals with aphasia. At the same time, research on treatment of word-retrieval impairment increasingly is focused on treating the impairment and measuring outcomes at the discourse level (see Boyle, 2011, for a review). Currently, there is no standardized test of word retrieval in discourse for PWA, which makes it challenging to measure change in performance. Herbert and colleagues (Herbert, Hickin, Howard, Osborne, & Best, 2008) argued that confrontation-naming scores are valid assessments of word retrieval in connected speech because they found moderate-to-strong correlations between the two tasks. Despite this argument, investigators have used a variety of methods to assess participants' word-retrieval abilities in connected speech.

Some researchers (Antonucci, 2009; Boyle, 2004; Boyle & Coelho, 1995; Cameron, Wambaugh, Wright, & Nessler, 2006; Coelho, McHugh, & Boyle, 2000; Falconer & Antonucci, 2012; Wambaugh & Ferguson, 2007) have measured the amount of information a person with aphasia is able to convey by using the measures of informativeness (correct information units [CIUs]), efficiency, and main concepts developed by Nicholas and Brookshire (1993, 1995). These measures serve as proxies for word-retrieval abilities, with the assumption that as word-retrieval ability improves, more CIUs and main concepts will be conveyed more efficiently. Other researchers have developed their own measures of successful word retrieval. For example, Mayer and Murray (2003) developed a measure to assess the percentage of nouns and verbs that are successfully retrieved (%WR) during discourse production, and this has been used to measure the outcome of treatment aimed at improving word retrieval (Antonucci, 2009; Falconer & Antonucci, 2012). Some researchers (Boyle, 2004; Peach & Reuter, 2010) have

[a]Montclair State University, Bloomfield, NJ

Correspondence to Mary Boyle: boylem@mail.montclair.edu

Editor: Rhea Paul
Associate Editor: Kristine Lundgren

Received July 1, 2013
Revision received September 2, 2013
Accepted October 14, 2013
DOI: 10.1044/2014_JSLHR-L-13-0171

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

focused on measuring word-retrieval problems, rather than successes, by modifying a test of word-finding difficulty in discourse developed for children (Test of Word Finding in Discourse; German, 1991). The assumption behind using this measure is that as word retrieval improves, the behaviors that signal word-finding difficulty will decrease. Still others (Fergadiotis & Wright, 2011; MacWhinney, Fromm, Forbes, & Holland, 2011; Rider, Wright, Marshall, & Page, 2008; Wright & Capilouto, 2009; Wright, Silverman, & Newhoff, 2003) have proposed using a measure of lexical diversity (*D*) as an assessment of word retrieval in aphasic discourse, reasoning that as word retrieval ability improves, a wider variety of words should be produced.

Test–Retest Stability

One concern about using a new or an adapted measure is its session-to-session stability. The terms *test–retest stability, session-to-session stability,* and *test–retest reliability* are interchangeable. They refer to the assessment of whether a test produces the same results on repeated applications when the participants who are being tested have not changed on the domain that is being measured (Fitzpatrick, Davey, Buston, & Jones, 1998). Schiavetti, Metz, and Orlikoff (2011) noted that estimating test–retest stability requires performing a complete repetition of the exact measurement on at least two occasions. There is not agreement about the length of time that should elapse between testing sessions, but the typical range of time is between 2 and 14 days (Fitzpatrick et al., 1998).

Bennett and Miller (2010) asserted that reliability of measurements forms the foundation of any scientific enterprise and noted that test–retest reliability varies depending on the measure being used (e.g., the various methods of assessing word retrieval in discourse), the thing being measured (e.g., the ability to retrieve and produce words in discourse), and day-to-day variations in the participant's physiologic and cognitive states (e.g., a participant feeling more tired or distracted on one day than another). Herbert and colleagues (2008) stated that establishing the stability of a measure is an essential prerequisite to using it as an outcome assessment for the evaluation of therapy, and 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, researchers and clinicians need information about the test–retest stability of the various measures of word retrieval in discourse. The test–retest stability of these measures is important if they are used to describe and analyze aspects of an individual's language impairment because a measure that is not reasonably stable from session to session will not provide a valid, reliable assessment of an individual's impairment. Furthermore, measures that lack acceptable test–retest stability may not be the best measures for assessing treatment-related changes. If outcome measures vary by large amounts before treatment commences, then even larger changes pre- to posttreatment are necessary to provide convincing evidence that those changes are related to the treatment rather than to the inherent variability of the measurement or of the behavior being measured.

Interpretation of Test–Retest Stability for Research and Clinical Purposes

Levels of test–retest stability that are acceptable for group research studies are not the same as levels that are acceptable for making decisions about an individual's impairment or change in performance. Fitzpatrick and colleagues (1998) noted that an instrument used to assess individuals should have a higher degree of reliability than one used to assess groups. They recommended a reliability level of at least 0.90 for instruments that are used to make clinical decisions about an individual, and they considered a reliability level of at least 0.70 acceptable for measures that assess groups of participants in clinical research. The more stringent requirements for individual assessment are related to the fact that confidence intervals around an individual's score are wide at reliability levels less than 0.90, so that the true score of the individual falls within a wide range of possible scores. For this reason, Donoghue and Stokes (2009) argued that the standard error of measurement (*SEM*), which indicates the extent to which a score varies on repeated measurements (Stratford, 2004), is better than the test–retest correlation coefficient for clinical applications, such as deciding whether an individual's performance has changed. The standard error of measurement denotes the amount of random error that is likely to be associated with a particular score. Knowing the standard error of measurement allows one to estimate a range of scores that indicates where a person's true score lies. For example, a person's observed score ±1 *SEM* indicates the range of scores that gives an examiner confidence that, 68 out of 100 times, the person's true score will be within that range of scores. The observed score ±1.65 *SEM* indicates the range of scores that will include the true score 90% of the time, and the observed score ±1.96 *SEM* indicates the range of scores that will include the true score 95% of the time. The standard error of measurement can be used to calculate another clinically useful measurement, the minimal detectable change (MDC) value. The MDC estimates the amount by which an individual's score must change on an assessment instrument in order to be sure that the change is a real one and not simply a reflection of measurement error (e.g., test–retest instability; Donoghue & Stokes, 2009). The MDC can be calculated to a particular level of confidence. MDC_{90} reflects a 90% confidence level, which is the level recommended by Donoghue and Stokes (2009) for decisions regarding the effectiveness of interventions for individuals.

Test–Retest Stability of Discourse-Level Word-Retrieval Measures

Of the measures of word retrieval in discourse outlined above, most is known about the test–retest stability of the informativeness, efficiency, and main concept measures.

Nicholas and Brookshire (1993, 1995) developed these measures for discourses elicited from PWA by using a set of 10 stimulus items. Recognizing the considerable amount of time necessary to transcribe and analyze such large samples of discourse, they sought to determine whether stable session-to-session results could be achieved using discourses from five of the stimulus items by dividing the stimuli into two equivalent sets of five items designated as Set A and Set B (Brookshire & Nicholas, 1994). Each set consisted of two complex picture narratives, one picture sequence narrative, one procedural discourse, and one autobiographical discourse. They reported that good test–retest stability was achieved for the number of words per minute (wpm), the number of CIUs per minute (CIUs/min), and the percentage of words that were CIUs (%CIUs). However, they did not analyze the stability of the other measures commonly used in their analysis systems (i.e., the number of words; the number of CIUs; and the number of accurate, complete, inaccurate, or absent main concepts) on the shorter, five-stimulus sets. The current investigation aims to replicate the results of Brookshire and Nicholas (1994) for the test–retest stability of wpm, CIUs/min, and %CIUs on the five stimulus items they designated as Set A and to extend their results by ascertaining the test–retest stability of the number of words, the number of CIUs, and the main concept analyses in discourses elicited from PWA using the Set A stimuli.

There is test–retest stability information available about the Test of Word Finding in Discourse (German, 1991) for a group of 30 children with typically developing language (a subset of the 856 children in the norming sample). This test examines manifestations of word-finding difficulty, which it calls word-finding behaviors (WFBs), and yields indices of word-finding difficulty in discourse. The global index is the percentage of T-units, which consist of at least a noun phrase plus a verb phrase and can stand alone to represent a complete thought, that contains any word-finding behaviors (%TWFB), with similar indices for each specific kind of word-finding difficulty (e.g., delays, substitutions). The manual for the test reports that the 30 children were tested on two occasions 14 days apart, resulting in a correlation coefficient of 0.84. This suggests strong test–retest stability when the measure is used with non-language-impaired children. However, because the test was not developed for or normed on adults with aphasia, there is no information available about its test–retest stability when it is applied to discourses from this population. This investigation aims to assess the test–retest stability of the indices of word-finding behavior when the test was adapted for use with adults who have aphasia.

Mayer and Murray (2003) did not provide information about the test–retest stability of their functional measure of word retrieval in discourse (%WR). Likewise, investigators who have suggested *D* as a proxy measure for word retrieval (Fergadiotis & Wright, 2011; MacWhinney et al., 2011; Rider et al., 2008; Wright & Capilouto, 2009; Wright et al., 2003) did not provide information about its test–retest stability when applied to the discourses of people with aphasia. This investigation aims to ascertain the test–retest stability of these two measures when they are applied to the discourses of people with aphasia.

To summarize, the purpose of this study was to examine the test–retest stability of select measures that have been used in published studies as direct or indirect measures of word retrieval in the discourses of PWA elicited using a five-item set of stimuli.

Method

The methodology used in this investigation was approved by the institutional review boards of the Winifred Masterson Burke Rehabilitation Hospital (White Plains, NY) and Montclair State University (Bloomfield, NJ). All participants provided signed informed consent.

Participants

The participants were 12 right-handed native-English-speaking PWA who demonstrated anomia as a prominent characteristic in connected speech. They were recruited from local hospital and university speech-language pathology clinics. Aside from sustaining a single left-hemisphere cerebrovascular accident or, in one case, a traumatic brain injury, there was no other history of neurologic impairment. None of the participants received concomitant speech-language treatment while these data were collected. 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. Participants ranged from 38 to 87 years of age ($M = 62$, $SD = 13.5$) and had been living with aphasia from 7 to 72 months ($M = 36$, $SD = 23.6$), placing them in the chronic stage of recovery from aphasia. All participants lived in their communities with a family member. Three of the participants were African American (1 woman and 2 men), and nine were Caucasian (2 women and 7 men). All participants had completed high school; four also had completed college, and two of the college graduates had 1 or 2 years of post-baccalaureate education (mean years of education = 14, $SD = 2.7$). Aphasia severity, assessed by the Aphasia Quotient from the Western Aphasia Battery (Kertesz, 1982), ranged from mild to moderate. Four participants had Broca's aphasia, four had anomic aphasia, two had conduction aphasia, and two had Wernicke's aphasia. No participant had dysarthria, and none had more than a very mild apraxia of speech as assessed by methods described by Duffy (2013). Results of the Test of Adolescent/Adult Word Finding (German, 1990) revealed that all participants had fairly significant word retrieval impairments, with none achieving better than the 23rd percentile rank.

Procedure

Stimuli and Elicitation Tasks

The five stimuli designated as Set A by Brookshire and Nicholas (1994) served as discourse elicitation stimuli. These consist of two complex pictures, one picture sequence, one request for a biographical narrative concerning one's typical

Table 1. Participants' demographic information and test results.

	Participant											
Characteristic	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**P11**	**P12**
Age	57	67	61	61	70	87	65	38	52	50	80	51
Gender	M	F	M	M	M	M	M	F	F	M	M	M
Race	B	W	W	W	W	W	B	W	B	W	W	W
Education (years)	12	12	12	12	18	12	16	12	12	12	19	16
MPO	65	15	59	14	15	36	37	38	64	7	14	72
WAB AQ (100)	72	72	67	54.5	90.6	72.2	86.6	67.4	70.2	82	61.2	46.3
Fluency (10)	4	4	4	4	9	9.0	9	9	5	8	8	8
Comp. (10)	10	10	9	5.45	9.4	9.8	9.3	8.2	9	9	6.7	4.95
Repetition (10)	7.9	7.8	6	5.8	9.7	8.9	7	8.2	4.2	5	3.8	1
Naming (10)	8	8	7.5	4	8.2	8.4	9	8.3	7.9	8	4.1	6.2
Aphasia type	BA	BA	BA	BA	AA	AA	AA	AA	CA	CA	WA	WA
TAWF (107)	67	82	68	28	84	63.0	53	56	57	84	28	8
Standard score	63	88	76	< 70	90	72.0	< 70	< 52	< 58	78	< 70	< 58
Percentile rank	0.2	19	4	< 1	23	2	< 1	< 0.1	< 0.1	6	< 1	< 0.1
Etiology	L CVA	L CVA	L CVA	L CVA	L CVA	L CVA	L CVA	TBI	L CVA	L CVA	L CVA	L CVA

Note. M = male; F = female; B = Black; W = White; MPO = months post-onset; WAB AQ = Western Aphasia Battery (Kertesz, 1982) Aphasia Quotient; Comp. = Comprehension BA = Broca's aphasia; AA = anomic aphasia; CA = conduction aphasia; WA = Wernicke's aphasia; TAWF = Test of Adolescent/Adult Word Finding (German, 1990); L = left; CVA = cerebrovascular accident; TBI = traumatic brain injury.

Sunday routine, and one procedural request concerning how to do dishes by hand. These stimuli were chosen for this study because they are readily available (Brookshire & Nicholas, 1994; Nicholas & Brookshire, 1993, 1995) and have been used to measure outcomes in a number of aphasia treatment studies (e.g., Antonucci, 2009; Boyle, 2004; Wambaugh & Ferguson, 2007). Using the procedures set forth by Nicholas and Brookshire (1993, 1995), the examiner asked the participant to tell a story about the pictures and picture sequences and asked participants to describe, "from the beginning to the end," their Sunday routines and how to do dishes by hand.

The five elicitation tasks were randomized in each of three sessions that occurred 2 to 7 days apart without intervening treatment. The sessions were audio-recorded and later orthographically transcribed by a trained research assistant. The author independently checked the transcriptions. Disagreements were resolved prior to scoring and analysis.

Data Analysis

Informativeness, efficiency, and main concepts. The CIU analysis, a standardized, rule-based scoring system described by Nicholas and Brookshire (1993), was used to evaluate the informativeness and efficiency of the discourse samples contained in the transcripts. This scoring system first requires a count of all words that are intelligible in context without regard to their accuracy, relevance, or informativeness. From those words, CIUs (i.e., words that are accurate, relevant, or informative relative to the eliciting stimulus) are identified and counted. The results from each elicitation task are added to yield a single score for all five discourses on each measure.

To evaluate the presence, accuracy, and completeness of main concepts in the discourse samples, the author used the standardized, rule-based scoring system described by Nicholas and Brookshire (1995). Main concepts are the main information, or gist, about a topic. According to the scoring system, a score of *accurate complete* (AC) denotes that all information associated with the concept is present, accurate, and complete. A score of *accurate incomplete* (AI) indicates that part of the essential information associated with the concept is accurate, but one or more essential parts are missing. A score of *inaccurate* (IN) denotes that one or more parts of the essential information are inaccurate. A score of *absent* (AB) indicates that none of the essential information associated with that concept is present. As Nicholas and Brookshire noted, it is not possible to determine main concepts for the personal discourse about one's usual Sunday routine given its idiosyncratic nature; therefore, the personal discourse was not included in the main concept analysis. The main concept lists and sets of scoring examples developed by Nicholas and Brookshire for the remaining stimuli were used to score the presence, accuracy, and completeness of main concepts from the transcripts of the participants' discourses. The results from these four elicitation tasks were combined to yield a single score for each category.

Functional measure of word retrieval. In their investigation, Mayer and Murray (2003) elicited composite descriptions of author-created picture sequences that depicted a series of events, each of which included multiple characters and activities. The present investigation differs by using the Nicholas and Brookshire (1993, 1995) stimuli to elicit narrative discourses. Narrative discourses rather than composite descriptions were elicited because story retelling seems to be a more common activity than providing composite descriptions, making it more ecologically valid in terms of adult communication activities. In addition, the picture stimuli developed by Nicholas and Brookshire are included in their published studies, making them available for this and subsequent investigations, whereas the pictures developed by Mayer and Murray (2003) were not included in their publication. To make the task as similar as possible to the Mayer and Murray task, only narrative discourses (stories about the complex pictures and the picture sequence) were analyzed for this part of the study. Nicholas and Brookshire's (1993) procedures were used to count words. As in Mayer and Murray's (2003) study, the discourses from each task were combined, and the first 300 words were scored; for those participants whose total output was fewer than 300 words, the entire narrative transcripts were scored. Mayer and Murray's procedures to score accurate and error noun and verb productions and to compute the percentage of noun and verb retrieval attempts that were successful (%WR Nouns and %WR Verbs) were used.

WFB analysis. The Test of Word Finding in Discourse (German, 1991) was developed for children, so some changes were made to analyze aphasic language of adults. The pictures published with the test were replaced by the Set A stimuli developed by Brookshire and Nicholas (1994) to elicit the discourses. In the test manual, German (1991) discussed using the test with stimuli other than those created for it, and she recommended elicitation and analysis procedures for use with such stimuli. Those recommendations were followed in this study. Because the test was developed for children, some categories that are typically used to identify word-retrieval difficulty in aphasia, such as paraphasias or neologisms, were not included in the test's analysis system. Boyle (2004) modified the test's categories to capture typical manifestations of lexical retrieval difficulty caused by aphasia. Specifically, German's general category, *substitutions,* was replaced with the following three categories: *verbal paraphasia, phonemic paraphasia,* and *neologism.* The operational definition for *verbal paraphasia* was an unintended substitution of one word for another whether or not the substitution was semantically related to the target. The operational definition for *phonemic paraphasia* was a fluently produced nonword obviously related in sound to the target. If the phonemic substitution resulted in a real word, it was classified as a verbal paraphasia. The operational definition for *neologism* was a nonword with no, or only a remote (fewer than 50% of phonemes in common), relation to the target. Two additional modifications of German's categories were made. An *initial sound* category was added. The operational definition of initial sounds was partial production of the target or of some attempt at the target (e.g., "wa" for water, "wom" for wife). German's category *insertion* was renamed *comment* to limit confusion with the category that she calls *time fillers* (e.g., "uh," "um"). *Comments* were statements made by the participants about the task or the language process (e.g., "I can't think what you call that"). Operational definitions of the remaining categories, *repetition, reformulation, empty/indefinite words, time fillers,* and *delays* were those used by German (1991, pp. 36–41). The procedures described by German (1991) for segmenting discourses, calculating total T-units, and counting word-finding behaviors were followed. Results from all five elicitation tasks were combined to yield a single score for each of the WFB measures.

A global measure of word-finding impairment was obtained by calculating the percentage of T-units that contained evidence of any category of WFB (%TWFB; German, 1991). In addition, the percentage of T-units containing each specific category of WFB was calculated to compare the different categories across a speaker's sessions.

Lexical diversity (D). For the lexical diversity analysis, trained research assistants converted the narrative discourse transcriptions to the Codes for the Human Analysis of Transcripts format and coded them according to the Computerized Language Analysis (CLAN) programs described by MacWhinney (2000). Only narrative discourses (i.e., stories about complex pictures and picture sequences) were submitted to this analysis because Fergadiotis and Wright (2011) reported that results for *D* varied by genre type. The three narrative discourses were combined to yield a single diversity score. The coded transcripts were checked by the author. Interrater agreement for word-by-word transcription and coding was above 93%. Disagreements were resolved by consensus before the transcripts were submitted to the CHECK program that is built into the CLAN software editor. The CHECK program ensures that transcriptions and codes adhere to the system requirements. The MOR command was used to tag parts of speech and was written to include only the participants' speech, so that the examiner's speech was excluded from further analysis. The POST command was applied to the output of the MOR command in order to assign parts of speech to ambiguous cases. *D* was calculated with the VOCD program in CLAN using a command code developed by MacWhinney et al. (2011, p. 1298). This command excludes false starts, neologisms, and unintelligible words, then examines lemmas (i.e., it treats inflected forms of the same base, like *cry, crying,* and *cried,* or like *man* and *men,* as the same lexical item) to obtain a measure of lexical diversity.

Reliability. After training and guided practice using transcripts from nonparticipant PWA with the analysis systems for informativeness, efficiency, main concepts, functional measures of word retrieval, and the WFB analysis, a research assistant scored all participant transcripts separately for each kind of analysis. The author independently scored one third of all transcripts for each system, randomly selecting one of the three discourse sets from each of the 12 participants. The computerization of the VOCD program rendered reliability checks of its analysis unnecessary once the reliability of the transcript and coding was established.

Assessment of session-to-session stability. Mean difference scores for the group of participants between Sessions 1 and 2, Sessions 2 and 3, and Sessions 1 and 3 were calculated. The mean scores and standard deviations for each measure in each session were calculated for use in the session-to-session stability measures. To assess the extent to which scores in the first session were related to scores in subsequent sessions, Pearson product–moment correlation coefficients were calculated. Using recommendations by Fitzpatrick and colleagues (1998), a 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*s) were calculated with the formula $\mathrm{SEM} = \mathrm{SD}\sqrt{1-r}$, 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 MDC was calculated with the formula $\mathrm{MDC}_{90} = \mathrm{SEM} \times \sqrt{2} \times 1.65$ (Stratford, 2004).

Results

Scoring Reliability

Point-to-point interrater agreement exceeded 88% for T-units, WFBs, number of words, and number of CIUs; exceeded 85% for %WR; and exceeded 80% for each of the four main concept scoring categories. Because the lexical diversity scoring was computerized, scoring reliability was not a concern.

Session-to-Session Stability of Measures

Absolute difference scores were calculated for each participant and used to calculate the mean difference score for the group on each measure. These mean difference scores, along with the range of difference scores produced by the participants, are in Table 2. Changes in scores on many of the measures were reasonably small, but changes on others (e.g., the number of words, several of the individual WFB categories, and the functional word-retrieval measures) were relatively large. When considering individual participants' difference scores as reflected in the range of the difference scores for the group, it is apparent that at least some of the participants produced fairly substantial session-to-session changes on many of the measures. The stability of each category of word-retrieval measurement will be considered separately.

Stability of Informativeness, Efficiency, and Main Concept Measures

The mean scores, standard deviations, ranges of scores, Pearson product–moment correlations, standard errors of measurement, and MDCs for informativeness, efficiency, and main concept measures are in Table 3. The number of words, CIUs, ACs, AIs, and ABs had correlation coefficients greater than .70, suggesting their suitability for use in group research studies (Fitzpatrick et al., 1998). Words per minute and CIUs/min had correlation coefficients greater than 0.90, indicating that they are sufficiently stable to use for clinical decision making about individuals (Fitzpatrick et al., 1998) as well as for group research studies. The MDC_{90} values for words per minute and CIUs/min indicate that for individual clinical decisions, changes of at least 9 wpm and 12 CIUs/min are probably unrelated to measurement error.

Correlation coefficients for the remaining measures (% CIUs, INs, and AI + IN) were below .70. The correlations for % CIUs ranged from moderate to very strong (*r*s = .61 to .95). The source of the weaker correlations for % CIUs was the variability of a single participant, P4, who nearly doubled the number of words he produced from the first to the second session without a concomitant increase in the number of CIUs. Reanalysis of the data without P4's scores yielded strong to very strong correlations for the % CIUs (*r*s = .70 to .93) Although this reanalysis suggests that this measure might be sufficiently reliable for use in group research studies, the notable variability of P4 on the measure suggests that it should be used cautiously. The range of IN responses was extremely limited in this sample of participants (0 to 6) in contrast to the ranges for the other main concept categories, and this probably accounted for the weak correlations for the categories that included this measure.

Stability of Functional Measures of Word Retrieval

The mean scores, standard deviations, ranges of scores, Pearson product–moment correlations, standard errors of measurement, and MDCs for the functional measures of word retrieval in discourse, %WR Nouns, and %WR Verbs are in Table 4. None of the correlation coefficients were above .70 for all three comparisons (i.e., Sessions 1 and 2, Sessions 2 and 3, and Sessions 1 and 3). A second analysis of the data that combined noun and verb productions (%WR Nouns + %WR Verbs) was done to assess whether this would improve the stability of the scores across sessions. This yielded one value above .70 (Session 1 to 2, *r* = .79) but did not increase the correlation coefficients for the other two comparisons (*r*s = .63, .32, respectively). These results suggest that these measures are probably not sufficiently stable for use in group research studies unless multiple pretreatment baselines are obtained to assess the variability of the measure.

Stability of WFB Analysis Measures

German's (1991) procedures for WFB analysis in discourse specify that a participant must produce more T-units than fragments in the discourse sample. T-units consist of at least a noun phrase plus a verb phrase, and they can stand alone to represent a complete thought. Fragments may consist of a noun phrase or a verb phrase, but fragments do not contain both a noun phrase and a verb phrase. Participant 3 produced more fragments than T-units, so his data were excluded from these analyses.

The mean scores, standard deviations, ranges of scores, Pearson product–moment correlations, standard errors of measurement, and MDCs for the various measures of word-finding difficulty in discourse are in Table 5. The correlation coefficients for %TWFB were greater than .70, indicating that this measure is sufficiently stable for use in group research studies. For two of the three correlation tests (Sessions 1 and 2; Sessions 2 and 3), the correlation coefficients were greater than .90, indicating that this measure might be stable

Table 2. Means and ranges of absolute difference scores for participants' performance on measures of informativeness, efficiency, main concepts, word finding, and lexical diversity between sessions.

	Sessions 1 to 2		Sessions 2 to 3		Sessions 1 to 3	
Measure	**Mean difference**	**Range of difference scores**	**Mean difference**	**Range of difference scores**	**Mean difference**	**Range of difference scores**
Number of words	18	2–35	18	0–57	19	0–49
Number of CIUs	9	1–39	8	0–27	7	0–27
Words per minute	4	1–13	6	0–15	8	1–29
CIUs per minute	5	1–17	3	0–16	7	0–33
%CIUs	10	2–28	4	0–13	11	1–29
AC	3	0–11	3	0–8	3	0–8
AI	2	0–4	2	1–5	2	1–4
IN	1	0–4	1	0–2	1	0–4
AB	3	0–7	3	1–5	3	1–7
AI + IN	2	0–6	2	0–6	2	0–5
%TWFB	7	0–14	5	0–9	9	2–21
%TVP	8	0–16	6	0–14	6	0–15
%TIS	13	1–46	9	0–29	12	0–38
%TPP	6	0–22	5	0–18	5	0–13
%TN	4	0–13	3	0–7	4	0–16
%TRep	15	1–36	9	3–17	14	4–29
%TRef	10	1–33	14	2–23	12	0–22
%TE	8	0–15	6	0–20	10	0–24
%TTF	9	0–29	5	0–19	8	0–29
%TD	5	0–13	5	0–29	3	0–16
%TC	3	0–13	3	0–14	4	0–14
%WR Nouns	8	1–22	12	0–23	14	1–24
%WR Verbs	15	3–37	10	0–19	15	2–37
%WR Nouns + Verbs	8	0–26	8	1–17	12	2–31
D	9	2–18	11	2–37	9	1–35

Note. CIUs = correct information units; %CIUs = percentage of all words that were CIUs; AC = accurate and complete; AI = accurate but incomplete; IN = inaccurate; AB = absent; %TWFB = percentage of T-units with one or more word-finding behaviors of any type; %TVP = percentage of T-units with one or more verbal paraphasias; %TIS = percentage of T-units with one or more initial sounds; %TPP = percentage of T-units with one or more phonemic paraphasias; %TN = percentage of T-units with one or more neologisms; %TRep = percentage of T-units with one or more repetitions; %TRef = percentage of T-units with one or more reformulations; %TE = percentage of T-units with one or more empty words; %TTF = percentage of T-units with one or more time fillers; %TD = percentage of T-units with one or more delays; %TC = percentage of T-units with one or more comments; %WR Nouns = percentage of words that were nouns that were retrieved accurately; %WR Verbs = percentage of verbs that were retrieved accurately; %WR Nouns + Verbs = the percentage of nouns plus verbs that were retrieved accurately; *D* = lexical diversity.

enough to use for individual clinical decision making. However, because the correlation between Sessions 1 and 3 did not reach .90, caution should be exercised, and a change of at least 19% (the MDC_{90} value for Sessions 1 and 3) should be used to estimate change that is not related to measurement error. For the measures of each kind of WFB, correlation coefficients for the percentage of T-units that contained empty or indefinite words (%TE) indicated sufficient stability for use in group research studies, and the percentage of T-units that contained time fillers (%TF) and delays (%TD) indicated sufficient stability for individual clinical decision making as well as in group research studies. A change of at least 21% on time fillers and a change of at least 7% on delays would indicate changes that are not related to measurement error. None of the other measures (the percentage of T-units that contained verbal paraphasias [%TVP], initial sounds [%TIS], phonemic paraphasias [%TPP], neologisms [%TN], repetitions [%TRep], reformulations [%TRef], and comments [%TC]) consistently produced correlation coefficients greater than .70, indicating that these measures are probably not sufficiently stable to use in group research studies or for clinical decision making about individuals.

Stability of Lexical Diversity

The mean scores, standard deviations, ranges of scores, Pearson product–moment correlations, standard errors of measurement, and MDCs for the measure of lexical diversity (*D*) in discourse are in Table 6. Correlations between sessions ranged from .77 to .88, indicating that this measure is sufficiently stable to use in group research studies but that multiple pretreatment baselines should be obtained if the measure is used for individual clinical decision making.

Discussion

The purpose of this study was to examine the test–retest stability of select measures of word retrieval in the discourses of PWA. The results suggest that some but not all

Table 3. Mean scores of informativeness, efficiency, and main concepts and Pearson product–moment correlation, *SEM,* and MDC_{90} values for differences in participants' performance between sessions on measures of informativeness, efficiency, and the accuracy and completeness of main concepts.

Session	Number of words	Number of CIUs	Words per minute	CIUs per minute	% CIUs	AC	AI	IN	AB	AI + IN
Session 1										
M	76	39	54	28	52	10	5	2	11	7
SD	31.17	16.65	32.06	18.37	8.47	6.60	2.68	1.45	5.26	2.39
Range	18–111	9–60	16–101	8–75	37–64	0–19	1–10	0–6	5–20	3–11
Session 2										
M	78	40	56	30	50	11	4	1	10	6
SD	39.39	25.46	32.50	23.54	16.29	8.00	2.77	0.90	6.29	2.83
Range	27–145	8–99	18–115	6–92	23–80	0–22	1–9	0–3	0–21	2–10
Session 3										
M	80	40	61	32	50	11	6	1	9	7
SD	36.44	22.32	37.93	27.44	17.77	6.90	3.45	0.75	4.92	3.23
Range	21–130	7–55	16–129	7–108	23–84	0–18	1–13	0–2	3–19	2–14
Sessions 1 to 2										
r	.84	.85	.99	.97	.61	.85	.70	.34	.83	.42
SEM	13	6	4	3	5	3	2	1	2	2
MDC_{90}	30	14	9	7	12	7	5	2	5	5
Sessions 2 to 3										
r	.78	.89	.99	.99	.95	.86	.86	.23	.92	.75
SEM	17	8	2	3	4	3	1	1	1	2
MDC_{90}	40	19	5	7	9	7	2	2	2	5
Sessions 1 to 3										
r	.74	.88	.99	.96	.64	.88	.80	.41	.87	.62
SEM	20	8	4	5	10	3	1	1	2	2
MDC_{90}	47	19	9	12	23	7	2	2	5	5

Note. MDC_{90} = minimal detectable change.

of the measures were stable across three sessions, with no intervening treatment, on the set of discourses used in this study. The number of words, the number of CIUs, AC, AI, AB, %TWFB, %TE, and *D* were stable enough to use in group research studies. In addition, some of the measures (wpm, CIUs/min, %TF, and %TD) yielded correlation coefficients greater than .90, indicating that they are also sufficiently stable to use for clinical decision making about individuals. These results are summarized in Table 7.

Informativeness, Efficiency, and Main Concept Measures

The results of this investigation yielded very strong correlations across sessions for the words per minute and CIUs/min (*r*s = .96 to .99), which replicates the results that Brookshire and Nicholas (1994) reported for PWA with these stimuli (*r*s = .98, .97, respectively). In contrast, the correlations for %CIUs (*r*s = .61 to .95) were lower than those reported by Brookshire and Nicholas (1994; *r* = .98). When the scores of the participant (P4) who appeared to be skewing the group scores were removed, the correlations for %CIUs improved, ranging from strong to very strong (*r*s = .70 to .93). Nevertheless, the marked variability that this participant demonstrated suggests that %CIUs should be used for clinical decision making only when an individual's variability on this measure has been assessed with multiple pretreatment baselines. This practice will increase the likelihood that decisions about change on the measure will

Table 4. Means scores on functional word-retrieval measures and Pearson product–moment correlation, *SEM,* and MDC_{90} values for participants' performance between sessions on %WR.

Session	% WR Nouns	%WR Verbs	%WR Nouns + Verbs
Session 1			
M	74	75	74
SD	12.15	19.71	14.36
Range	51–87	23–97	44–89
Session 2			
M	80	79	80
SD	8.68	16.40	10.22
Range	63–91	50–97	70–91
Session 3			
M	76	77	77
SD	13.73	15.85	10.99
Range	47–100	43–100	57–93
Sessions 1 to 2			
r	.68	.50	.79
SEM	7	14	7
MDC_{90}	16	33	16
Sessions 2 to 3			
r	.38	.72	.63
SEM	7	9	6
MDC_{90}	16	21	14
Sessions 1 to 3			
r	.20	.46	.33
SEM	11	15	12
MDC_{90}	26	35	28

Table 5. Mean scores of word retrieval measures, Pearson product–moment correlation, *SEM,* and MDC_{90} values for differences in participants' performance between sessions.

Session	%TWFB	%TVP	%TIS	%TPP	%TN	%TRep	%TRef	%TE	%TTF	%TD	%TC
Session 1											
M	76	13.9	27.3	8.6	3.5	40.5	35.6	28.2	17.9	9.2	4.1
SD	16.78	9.40	19.05	11.23	5.41	15.89	13.1	19.57	29.50	15.69	6.85
Range	53–100	0–25	3–69	0–30	0–19	8–69	8–57	0–67	0–100	0–54	0–23
Session 2											
M	72	13.7	28.7	10.4	3.7	41.2	41.7	22.5	17.1	10.1	2.5
SD	16.52	8.67	26.59	11.23	4.73	19.88	16.23	18.22	26.55	19.73	3.11
Range	44–100	0–29	0–70	0–33	0–13	14–87	17–70	0–63	0–93	0–67	0–10
Session 3											
M	74	14.7	25.5	7.2	1.5	38.6	40.8	22.0	14.7	6.5	4.6
SD	13.53	8.52	23.89	8.30	2.54	19.65	17.06	17.11	25.39	11.61	6.55
Range	50–94	0–25	0–69	0–21	0–8	21–91	14–72	0–46	0–88	0–38	0–18
Sessions 1 to 2											
r	.90	.38	.69	.68	.30	.46	.60	.91	.90	.95	.73
SEM	5	7	11	6	5	12	8	6	9	3	4
MDC_{90}	12	16	26	14	12	28	19	14	21	7	7
Sessions 2 to 3											
r	.94	.55	.87	.75	.76	.85	.53	.89	.97	.94	.73
SEM	4	6	9	4	1	8	12	6	5	3	3
MDC_{90}	9	14	21	9	2	18	28	14	12	7	7
Sessions 1 to 3											
r	.76	.63	.71	.78	.04	.55	.66	.83	.93	.97	.56
SEM	8	5	14	5	5	13	9	8	7	3	2
MDC_{90}	19	12	33	12	12	31	21	19	16	7	5

take into account day-to-day variability rather than erroneously judging change to be due to treatment or recovery.

In their 1994 study that aimed to assess the stability of a set of five discourses, Brookshire and Nicholas did not report correlation values for all of the informativeness and efficiency measures that they included in the original study of a 10-discourse set (Nicholas & Brookshire, 1993). Among the measures that were not reported for the five-discourse set were the number of words and the number of CIUs. This investigation extends their results by suggesting that these two measures, obtained with the five-discourse stimuli that Brookshire and Nicholas (1994) designated as Set A, are sufficiently stable to be used in group research studies. For individual clinical decisions, multiple baselines on these

Table 6. Mean scores on the lexical diversity measure (*D*) and Pearson product–moment correlation, *SEM,* and MDC_{90} values for participants' performance between sessions.

Session	*D*
Session 1	
M	48
SD	19.32
Range	22–96
Session 2	
M	50
SD	22.42
Range	15–98
Session 3	
M	44
SD	14.34
Range	23–72
Sessions 1 to 2	
r	.88
SEM	7
MDC_{90}	16
Sessions 2 to 3	
r	.79
SEM	10
MDC_{90}	23
Sessions 1 to 3	
r	.77
SEM	9
MDC_{90}	21

Table 7. Summary of results for research and clinical purposes.

Measures stable enough for group research studies	Measures stable enough for clinical decision making about individuals (with MDC_{90} values)
Number of words	Words per minute (9 wpm)
Number of correct information units	Correct information units/minute (12 CIUs/min)
Accurate complete main concepts	Percentage of T-Units with time-fillers (21%TTF)
Accurate incomplete main concepts	Percentage of T-Units with delays (7%TD)
Absent main concepts	
Percentage of T-Units with word-finding behaviors of any type	
Percentage of T-Units with empty/indefinite words	
D	

Note. The MDC_{90} value is the score on an instrument necessary to reflect real change versus measurement error. It is used in making clinical decisions about individuals. PWA = people with aphasia.

measures are necessary to ascertain the day-to-day variability of individuals.

Nicholas and Brookshire (1995) reported correlation coefficients for the main concept analyses for the entire set of 10 discourse elicitation tasks, but they did not report the stability of these analyses for the smaller set of stimuli that comprise Set A (Brookshire & Nicholas, 1994). Results of the present investigation resulted in strong to very strong correlation coefficients for the AC, AI, and AB main concept analyses (*r*s = .70 to .92), which are similar to those reported by Nicholas and Brookshire (1995) for the entire set of discourse elicitation stimuli (*r*s = .71 to .96). This result suggests that these main concept analyses were sufficiently stable using the smaller set of stimuli that they can be used in group research studies. In contrast, this investigation resulted in far lower correlations for the IN and the AB + IN categories (*r*s = .23 to .75) than those reported by Nicholas and Brookshire for the entire stimulus set (*r*s = .71 to .86). The PWA in this investigation produced an extremely limited range of IN responses (0 to 6) compared with the PWA in the Nicholas and Brookshire study, who produced 0 to 17 IN responses. This restricted range of inaccurate responses accounts for the reduced correlations.

Functional Measures of Word Retrieval

The correlations for the functional measures of word retrieval (%WR Nouns and %WR Verbs) were relatively low, suggesting that their session-to-session stability is not ideal. Combining nouns and verbs into a single measure (%WR Nouns + Verbs) did not improve the stability sufficiently to support its use in group studies. Participants' scores on these measures varied between sessions by as much as 37% (Table 2). Given the instability of this measure between sessions without intervening treatment, it is probably not an ideal choice to measure treatment-related changes, at least with the Set A stimuli. In addition, this instability suggests that it may not provide an accurate picture of an individual's word-retrieval abilities in discourse, so clinicians should be cautious in using it for such diagnostic purposes unless they obtain data from more than one session in order to account for its variability.

In accordance with the procedures used by Mayer and Murray (2003), only the narrative discourses (two complex-picture narratives and one picture-sequence narrative) were examined, and only the first 300 words of the narratives were subjected to analysis. The functional measures of word retrieval proved to be unstable with these discourse samples. It is possible that better test–retest stability could be obtained with larger discourse samples than those used in this study.

WFB Analysis Measures

The results for the analysis of behaviors that characterize word-finding difficulties indicated that a global measure of word-finding behaviors (%TWFB), which indicates the percentage of utterances in a discourse that contain evidence of word-finding difficulty, is sufficiently stable for use in group research studies. The results for %TE, the percentage of T-units containing empty words, were also at or above the level of stability recommended for use in group research designs (Fitzpatrick et al., 1998). Two of the WFB measures, %TF and %TD (the percentage of T-units containing time fillers and delays, respectively), yielded evidence of stability at or above the level recommended for individual clinical decision making (Fitzpatrick et al., 1998).

In contrast, there was poor session-to-session stability on the other measures of specific categories of word-finding problems. This was a surprising finding. Why is there such variability of specific WFBs in discourse? Several investigations have demonstrated that the naming ability of PWA differs in confrontation naming and discourse (Mayer & Murray, 2003; Pashek & Tompkins, 2002; Williams & Canter, 1982, 1987). Pashek and Tompkins (2002) postulated that discourse provides semantic facilitation because the semantic features of the target are often contained elsewhere in the discourse. For example, some of the verbs produced in the discourse will be the action or function feature of a noun that must be retrieved, so the production of the associated verb facilitates production of the noun. Syntactic facilitation occurs in discourse because grammatical role assignments that are made during discourse production may constrain the number of viable candidates that are activated during word retrieval, thereby reducing competition. For example, Edmonds, Nadeau, and Kiran (2009) have demonstrated that verb retrieval influences successful retrieval of associated noun arguments at the sentence level.

Because lexical selection and syntactic structures can vary from day to day without compromising the accuracy of the information conveyed, the potential for various kinds of errors may also change from one day to the next. For example, the words and syntactic forms that a person chooses one day might not be exactly the same as the ones chosen the next day to tell the same story, and that may influence the errors that are produced. Research on serial picture naming in adults without neurological impairment has shown that, although naming an item from a semantic category primes that same item for later naming, it simultaneously interferes with subsequent naming of other items from the same category (Belke, 2008; Howard, Nickels, Coltheart, & Cole-Virtue, 2006; Oppenheim, Dell, & Schwartz, 2010; Vigliocco, Vinson, Damian, & Levelt, 2002). Oppenheim and colleagues (2010) referred to this as *cumulative semantic interference*. It seems plausible that if cumulative semantic interference occurs during serial naming tasks, it may also occur in discourse production. If so, then if a person retrieves more items from the same semantic category during discourse production on one day than on another day, it might cause more semantic errors on that category, because greater cumulative semantic interference would be present. Similarly, it might be that using more phonologically complex words in a discourse on one day than another might make that discourse more susceptible to phonemic paraphasias. These explanations for the instability of particular error types across sessions are speculative. Further investigation of these influences could improve understanding of word retrieval in the discourse of PWA.

Lexical Diversity

The correlations for the measure of lexical diversity, *D,* were strong (*r*s = .77 to .88) and met the test–retest criterion for use in group research studies but not for use in clinical decision making about PWA (Fitzpatrick et al., 1998). As with the other measures that yielded these results, investigators and clinicians who use *D* to make decisions about the word-retrieval impairment of an individual or about changes in word-retrieval ability over time should obtain measures across several days to assess variability before drawing conclusions. In this investigation, *D* scores changed by as much as 98 units over a relatively short period of time (2 to 7 days) with no intervention. It is possible that using larger discourse samples than the two complex-picture narratives and the single sequential-picture narrative used in this study could result in session-to-session stability that is sufficient for individual clinical decision making. Additional research into this possibility would be useful.

Severity of Aphasia, Age, and Time Post-Onset

The participants with aphasia in this investigation exhibited a variety of aphasia types and ranged in severity from mild to moderate. None of the participants exhibited severe aphasia. It is not clear that the word-retrieval measures that were examined in this study would be stable for participants with severe aphasia. Similarly, because participants only in the chronic stage of recovery from aphasia were included in this study, the results may not apply to individuals in the acute recovery period.

Although this investigation was not designed to assess the relationship of variables such as aphasia severity, age, and time post-onset to the test–retest stability of the word-retrieval measures, it is possible that these variables might contribute to response variability across sessions. Therefore, post hoc analyses of the relationships of aphasia severity (as measured by the WAB Aphasia Quotient), months post-onset, and age to the absolute difference scores between Sessions 1 and 2 on each measure were completed to provide preliminary information about potential relationships. The results are in Table 8.

There were no strong or very strong correlations between aphasia severity, time post-onset, or age with difference scores on any of the word-retrieval measures. There were moderate correlations between these variables and difference scores on some of the word-retrieval measures.

Severity of Aphasia

There were moderate positive correlations between aphasia severity and the number of words (*r* = .40), AC (*r* = .50), and %TVP (*r* = .56). This suggests that there was a moderate tendency for participants with milder aphasia (and hence larger WAB AQ scores) to be more variable between Sessions 1 and 2 on these measures than participants with moderate aphasia. There were moderate negative correlations between aphasia severity and words per minute (*r* = –.69), suggesting that there was a modest tendency for participants with moderate aphasia (i.e., with smaller WAB AQ values) to be more variable in the number of words produced per minute in Sessions 1 and 2 than participants with mild aphasia.

Table 8. Pearson product–moment correlation coefficients for absolute difference scores of participants' performance between Sessions 1 and 2 on word-retrieval measures and participants' aphasia severity (measured by the Western Aphasia Battery Aphasia Quotient), months post-onset, and age.

	Correlation coefficient		
Word-retrieval measure	Aphasia severity	Months post-onset	Age
Number of words	.40	–.15	.03
Number of CIUs	.24	–.36	–.06
Words per minute	–.69	.07	–.26
CIUs per minute	.02	–.06	.07
%CIUs	.05	–.35	.05
AC	.50	–.53	.19
AI	.07	–.15	.21
IN	–.02	.31	.39
AB	.36	–.67	.09
AB + IN	.10	–.09	.47
%WR Nouns	–.30	.36	–.59
%WR Verbs	.17	.37	–.12
%WR Nouns + Verbs	–.04	.29	–.16
%TWFB	.04	–.59	.26
%TVP	.56	–.37	.26
%TIS	.02	–.07	–.20
%TPP	–.35	–.40	.35
%TN	–.12	–.42	–.09
%TRep	.28	–.25	–.12
%TRef	–.29	.42	–.20
%TE	.20	–.13	–.19
%TTF	–.08	–.19	–.19
%TD	.27	.25	–.51
%TC	.30	.14	–.27
D	.10	.00	.23

Time Post-Onset

All of the participants in this investigation were in the chronic stage of recovery, more than 6 months post-onset of aphasia. Within this chronic stage, there was a moderate positive correlation between time post-onset and %TRef (*r* = .42), suggesting that there was a moderate tendency for participants whose aphasia was more chronic to be more variable in producing reformulations in Sessions 1 and 2. There were moderate negative correlations between time post-onset and AC (*r* = –.53), AB (*r* = –.67), %TWFB (*r* = –.59), %TPP (*r* = –.40), and %TN (*r* = –.42). This suggests that there was a moderate tendency for participants who were earlier in the chronic stage of recovery to be more variable on these measures between Sessions 1 and 2.

Age

There was a moderate positive correlation between age and AB + IN (*r* = .47), suggesting that there was a modest tendency for older participants to be more variable on this measure between Sessions 1 and 2. There were moderate negative correlations between age and %WR Nouns ($r = -.59$) and between age and %TD ($r = -.51$). This suggests that there was a modest tendency for younger participants to be more variable on these measures between Sessions 1 and 2.

Summary

It is evident that there is no simple, clear relationship between session-to-session variability of word retrieval in discourse and aphasia severity, time post-onset, or age. This is not altogether surprising. It is likely that a variety of factors, including those just discussed, contribute to the variable performance demonstrated in this investigation. Bennett and Miller (2010) noted that tasks involving higher cognition had lower test–retest reliability than motor and sensory tasks in functional MRI studies of participants without brain injury. They reported that test–retest reliability in participants with clinical disorders was typically even lower than that of participants without neurological impairment. They attributed variable session-to-session performance to the many physiological and cognitive changes that may take place within a participant between the testing sessions and declared that "test–retest methodology involving human beings is akin to hitting a moving target" (Bennett & Miller, 2010, p. 137). In light of this description, it is impressive that several of the word-retrieval measures in this study were stable across three sessions. It suggests that these measures, summarized in Table 7, are sufficiently robust to overcome the physiological and cognitive sources of variability hypothesized by Bennett and Miller. It is also important to note that there are no published data regarding the day-to-day variability of adults without aphasia on these word-retrieval measures. Without such normative data, it is difficult to ascertain how much of the variable performance is due to aphasia and how much reflects normal day-to-day fluctuations in performance experienced by adults without neurological impairment. Despite this limitation, these measures have been used to assess word-retrieval performance of PWA, and so it is important to be aware of their test–retest reliability, regardless of the source of variable performance. Research of the test–retest performance of individuals without aphasia on these measures would improve our understanding of aphasia's role in variable performance on them. However, it is likely that many factors influence session-to-session variability in word retrieval during discourse production. Research designs such as multiple regressions or factor analysis, requiring relatively large samples of participants, will probably be necessary to begin to understand what the factors are and how they interact.

Discourse-Elicitation Stimuli

This investigation used the five discourse-elicitation tasks and stimuli that Brookshire and Nicholas (1994) designated as Set A in their study. It is not clear that the measures of informativeness, efficiency, main concepts, and word-finding behavior based on samples smaller than the five-task samples would yield acceptable test–retest stability. As previously discussed, it is possible that analyzing only the first 300 words from three of the five tasks for the functional word retrieval measures might have contributed to the poor session-to-session stability that was found for these measures. Therefore, investigators who wish to use the word-retrieval assessments that were examined in this study on smaller samples of discourses should first assess how reducing the length of the discourse samples might affect the session-to-session stability. Similarly, as it is not clear how the pictures used to elicit the discourses might affect their production, it cannot be assumed that the word-retrieval measures that were examined in this study would have similar stability if different stimuli were used to elicit the discourses.

Conclusion

Several word-retrieval measures met Fitzpatrick and colleagues' (1998) test–retest stability criterion for use in group studies. Fewer met the criterion for use in clinical decision making. Researchers who use single-subject designs can account for the variability of these measures by collecting enough pretreatment baselines to establish the variability of the measure and then consider that variability when assessing change in performance. Clinicians rarely have the luxury of using such a practice. Therefore, one important goal for future research is to develop discourse-level word-retrieval measures that are stable enough to be applied in clinical decision making about individuals.

Acknowledgments

This work was supported by the Winifred Masterson Burke Rehabilitation Hospital and by a Separately Budgeted Research Award from Montclair State University. Thanks go to the people with aphasia who participated and to the students who assisted in various stages of this project—namely, Catherine Walsh, Jacqueline Frank, Erika Hough, Maria Guth, Gretchen Beideman Szabo, Christine Baker, Jacqueline Karpman, Benjamin Brophy, Lindsay Iser, Elizabeth Kline, and Keli Meyer.

References

Antonucci, S. M. (2009). Use of semantic feature analysis in group aphasia treatment. *Aphasiology, 23,* 854–866.

Belke, E. (2008). Effects of working memory load on lexical-semantic encoding in language production. *Psychonomic Bulletin & Review, 15,* 357–363.

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. (2004). Semantic feature analysis treatment for anomia in two fluent aphasia syndromes. *American Journal of Speech Language Pathology, 13,* 236–249.

Boyle, M. (2011). Discourse treatment for word retrieval impairment in aphasia: The story so far. *Aphasiology, 25,* 1308–1326.

Boyle, M., & Coelho, C. A. (1995). Application of semantic feature analysis as a treatment for aphasic dysnomia. *American Journal of Speech-Language Pathology, 4,* 94–98.

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

Cameron, R. M., Wambaugh, J. L., Wright, S. M., & Nessler, C. L. (2006). Effects of a combined semantic/phonologic cueing treatment on word retrieval in discourse. *Aphasiology, 20,* 269–285.

Coelho, C. A., McHugh, R. E., & Boyle, M. (2000). Semantic feature analysis as a treatment for aphasic dysnomia: A replication. *Aphasiology, 14,* 133–142.

Donoghue, D., & Stokes, E. K. (2009). How much change is true change? The minimum detectable change of the Berg Balance Scale in elderly people. *Journal of Rehabilitation Medicine, 41,* 343–346.

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

Edmonds, L. A., Nadeau, S. E., & Kiran, S. (2009). Effect of Verb Network Strengthening Treatment (VNeST) on lexical retrieval of content words in sentences in persons with aphasia. *Aphasiology, 23,* 402–424.

Falconer, C., & Antonucci, S. M. (2012). Use of semantic feature analysis in group discourse treatment for aphasia: Extension and expansion. *Aphasiology, 26,* 64–82.

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

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), i–iv, 1–74.

German, D. J. (1990). *Test of Adolescent/Adult Word Finding*. Austin, TX: Pro-Ed.

German, D. J. (1991). *Test of Word Finding in Discourse (TWFD): Administration, scoring, interpretation, and technical manual*. Austin, TX: Pro-Ed.

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.

Howard, D., Nickels, L., Coltheart, M., & Cole-Virtue, J. (2006). Cumulative semantic inhibition in picture naming: Experimental and computational studies. *Cognition, 100,* 464–482.

Kaplan, E., Goodglass, H., & Weintraub, S. (2001). *Boston Naming Test* (2nd ed.). Philadelphia, PA: Lippincott, Williams, & Wilkins.

Kertesz, A. (1982). *Western Aphasia Battery*. New York, NY: The Psychological Corporation.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Erlbaum.

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

Mayer, J. F., & Murray, L. L. (2003). Functional measures of naming in aphasia: Word retrieval in confrontation naming versus connected speech. *Aphasiology, 17,* 481–497.

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. (2010). The dark side of incremental learning: A model of cumulative semantic interference during lexical access in speech production. *Cognition, 114,* 227–252.

Pashek, G. V., & Tompkins, C. A. (2002). Context and word class influences on lexical retrieval in aphasia. *Aphasiology, 16,* 261–286.

Peach, R. K., & Reuter, K. A. (2010). A discourse-based approach to semantic feature analysis for the treatment of aphasic word retrieval failure. *Aphasiology, 24,* 971–990.

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. *American Journal of Speech-Language Pathology, 17,* 161–172.

Schiavetti, N., Metz, D. E., & Orlikoff, R. F. (2011). *Evaluating research in communicative disorders* (6th ed.). Upper Saddle River, NJ: Pearson Education.

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

Vigliocco, G., Vinson, D. P., Damian, M. F., & Levelt, W. (2002). Semantic distance effects on object and action naming. *Cognition, 85,* B61–B69.

Wambaugh, J. L., & Ferguson, M. (2007). Application of semantic feature analysis to retrieval of action names in aphasia. *Journal of Rehabilitation Research & Development, 44,* 381–394.

William, S. E., & Canter, G. J. (1982). The influence of situational context on naming performance in aphasic syndromes. *Brain and Language, 17,* 92–106.

Williams, S. E., & Canter, G. J. (1987). Action-naming performance in four syndromes of aphasia. *Brain and Language, 32,* 124–136.

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.