Clinical Aphasiology, Vol. 22, 1994, pp. 119–133

Test-Retest Stability of Measures of Connected Speech in Aphasia

Robert H. Brookshire and Linda E. Nicholas

Almost all aphasic adults have impairments in connected speech. Clinicians often are called on to describe, measure, and treat connected speech impairments in adults with aphasia, and they may categorize persons in terms of the severity and type of their aphasia based in large part on their connected speech. Clinicians may also use samples of connected speech to establish baselines against which to assess the effects of neurologic recovery or treatment. Such uses of connected speech require two assumptions: (a) that the speech samples are representative of the speaker's general connected speech abilities and (b) that differences in connected speech among patients, or differences in a given patient's connected speech on different test occasions, are reliable differences. Although these assumptions are crucial, their legitimacy has not been established, in large part because we know little about the test-retest stability of measures typically used to assess connected speech in adults with aphasia.

Numerous studies in the area of child language have shown that the size of a speech sample often has potent effects on the stability of the measures used to quantify it (Hess, Haug, & Landry, 1989; Hess, Sefton, & Landry, 1986). Consequently, it is generally recognized that children's speech samples should meet certain minimum length requirements, although no generally accepted minimum length exists (Miller, 1981).

We know very little, however, about the effects of sample size on measures of connected speech for adults with aphasia. The issue of speech sample size seems particularly germane to aphasiology, because speech samples collected in clinical and research activities with aphasic adults tend to be quite short—often fewer than 100 words. Such samples often come from an aphasic person's description of a single picture, which is frequently the so-called Cookie Theft picture from the *Boston Diagnostic Aphasia Examination* (BDAE) (Goodglass & Kaplan, 1983).

We became concerned with the issue of sample size as we were working to validate measures of connected speech of adults with aphasia and to measure how different elicitation stimuli affected the speech samples obtained. On the average, our stimuli each elicited fewer than 100 words from aphasic speakers, and it seemed to us that basing our measures on such short speech samples might compromise their stability. Consequently, we decided to evaluate whether measures based on short speech samples are unstable and, if so, whether combining short speech samples to make longer ones increases test-retest stability.

METHOD

Subjects

Subjects were 20 non-brain-damaged (NBD) adults and 20 aphasic (APH) adults. All were native speakers of English who had hearing and vision adequate for the tasks. Aphasic subjects were at least 3 months post onset of a single left-hemisphere cerebrovascular brain injury. Six exhibited nonfluent (essentially Broca's) aphasia and 14 exhibited fluent aphasia (fluent speech with literal paraphasias and word retrieval difficulty). Their aphasia severity, as estimated by their overall percentile on a four-subtest shortened version (SPICA) (Disimoni, Keith, & Darley, 1980) of the *Porch Index of Communicative Ability* (PICA) (Porch, 1971), ranged from the 40th to the 85th percentile. Aphasic subjects ranged in age from 51 to 77 years (*M* = 64.9, *SD* = 6.8) and in education from 10 to 16 years (*M* = 13.1, *SD* = 1.7). Non-brain-damaged subjects were nonhospitalized and noninstitutionalized adults who were similar to the aphasic adults in age (*M* = 64.2 years, *SD* = 7.0; range = 50–73) and education (*M* = 12.8; *SD* = 2.2; range = 8–16).

Stimuli

Connected speech was elicited from aphasic and non-brain-damaged speakers with a set of 10 stimuli. The stimuli included two aphasia test pictures, the Cookie Theft picture from the BDAE and the Picnic picture from the *Western Aphasia Battery* (WAB) (Kertesz, 1982). The stimuli also included the following:

- two single pictures drawn for the study (*Cat in Tree* and *Birthday Party*, see Figure 1);
- two picture sequences drawn for the study (*The Argument* and *Directions*, see Figure 2);

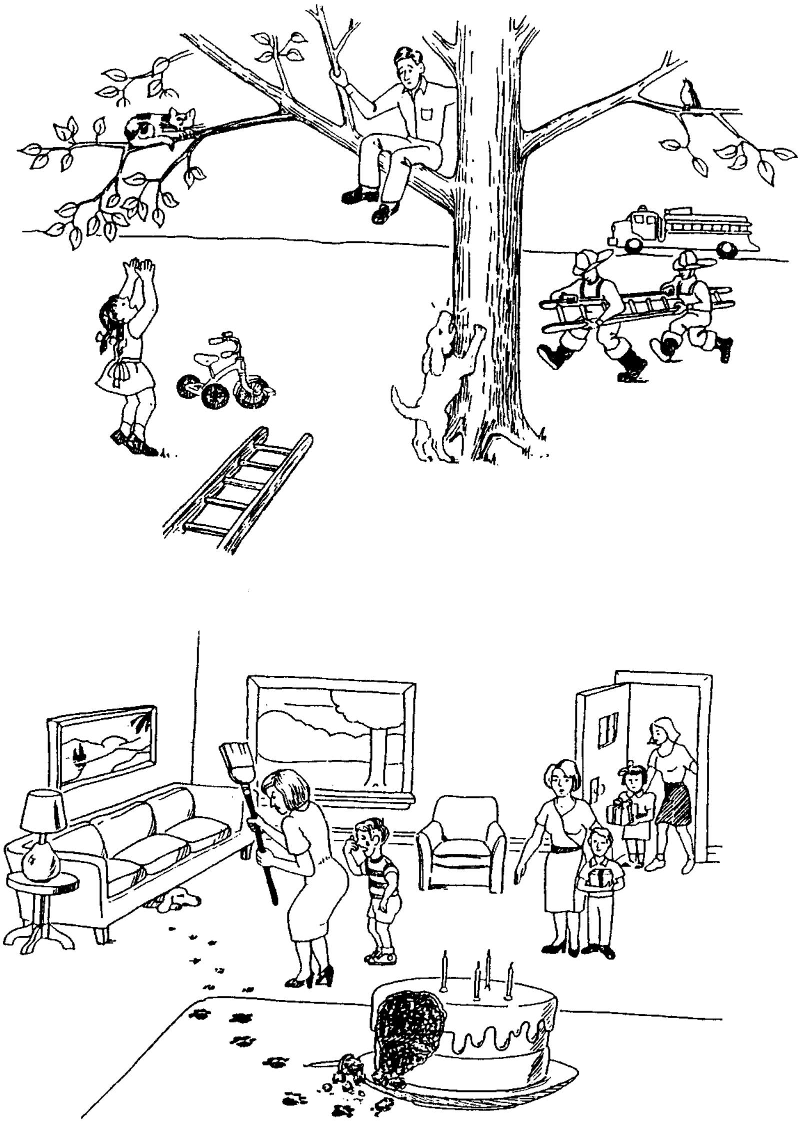

Figure 1. Two single-picture elicitation stimuli drawn for the study. (Copyright 1992, Robert H. Brookshire and Linda E. Nicholas. Reproduced with permission.)

- two requests for personal information:
  "Tell me what you usually do on Sundays," and
  "Tell me where you live and describe it to me"; and
- two requests for procedural information:
  "Tell me how you would go about doing dishes by hand," and
  "Tell me how you would go about writing and sending a letter."

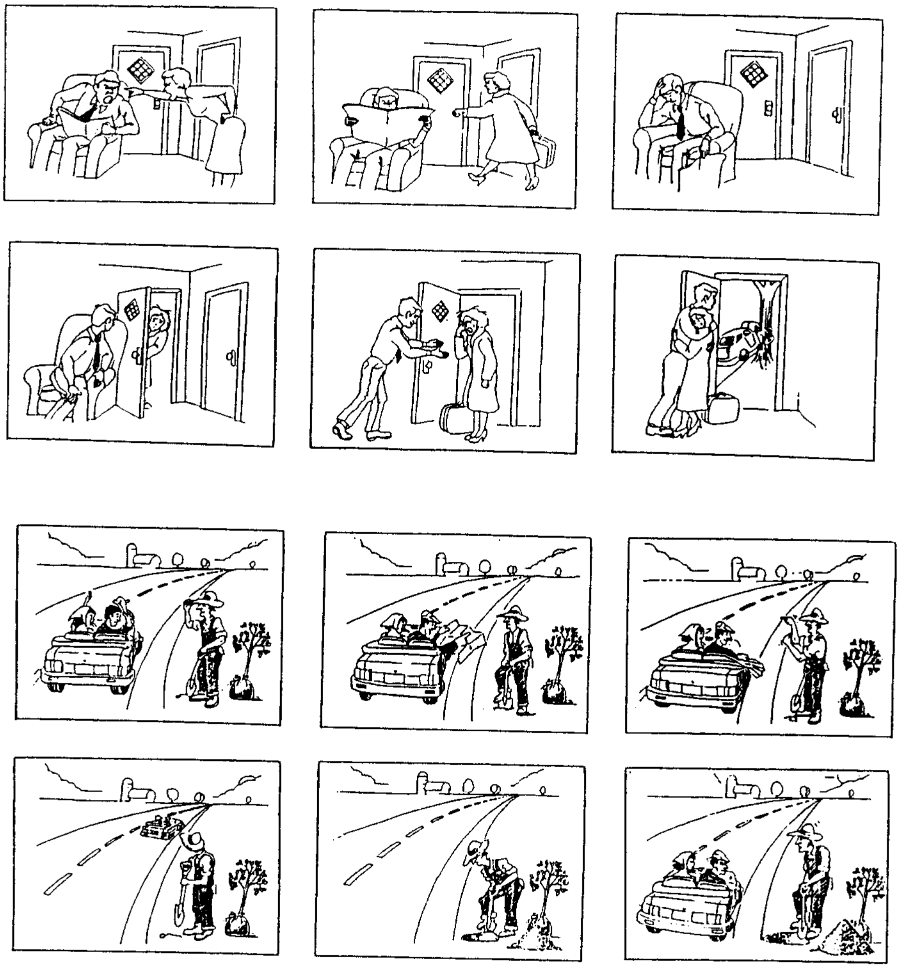

Figure 2. Two sequential picture elicitation stimuli drawn for the study. (Copyright 1992, Robert H. Brookshire and Linda E. Nicholas. Reproduced with permission.)

Measures of Connected Speech

This study evaluated the test-retest stability of three measures of connected speech, *words per minute, correct information units per minute,* and *percent correct information units.*[1] Speech samples were elicited from the non-brain-damaged and the aphasic subjects with the 10 elicitation stimuli

[1] We developed correct information unit (CIU) analysis to quantify the *informativeness* of connected speech (Nicholas & Brookshire, 1993). Correct information units are words that are intelligible in context and accurately convey information relevant to the eliciting stimulus. Percent correct information units is calculated by dividing the number of CIUs in a speech sample by the number of words in the sample.

on two occasions. The second session followed the first after 7 to 10 days. A different random order of stimuli was established for each subject, and each subject responded to the 10 stimuli in the same random order in both sessions. Each subject's responses to the stimuli were recorded on audiotape and orthographically transcribed. Two judges independently verified the accuracy of the transcripts and made corrections as needed. The corrected transcripts were then timed and scored for words and correct information units (CIUs).

To assess scoring reliability, two scorers independently scored the transcripts of six non-brain-damaged and six aphasic subjects. Point-to-point percentage agreement on words ranged from 97 to 100 percent; on CIUs, it ranged from 90 to 99 percent.

Procedures

To evaluate the effects of sample size on the stability of the measures, we divided the 10 elicitation stimuli into two sets of 5 stimuli each, which we labeled Sets A and B; Table 1 shows their contents. We then evaluated the test-retest stability of the three measures for all 10 stimuli combined (hereafter called Set AB), for Sets A and B (which contained 5 stimuli each), and for one single stimulus—the *Cookie Theft* picture from the BDAE. We chose the *Cookie Theft* picture as the single stimulus because it has been a popular vehicle for eliciting connected speech from aphasic adults, not because we believed that it would yield samples with either greater or less stability than any other single elicitation stimulus.

RESULTS

To determine an overall measure of test-retest stability, we first calculated correlation coefficients between subjects' Session 1 and Session 2 scores

Table 1: Elicitation Stimuli Contained in Set A and Set B

Category of Stimulus	*Set A*	*Set B*
Aphasia test picture	BDAE Cookie Theft	WAB Picnic
Single picture	Birthday Party	Cat in Tree
Picture sequence	Argument	Directions
Personal information	Sunday	Live
Procedures	Dishes	Letter

Note: BDAE = *Boston Diagnostic Aphasia Examination;* WAB = *Western Aphasia Battery.*

for Sets AB, A, and B and for the "Cookie Theft" picture. The results are shown in Table 2. The correlations were always higher for scores based on sets of stimuli than for scores based on the single stimulus, suggesting greater test-retest stability for scores based on larger speech samples.

To determine the magnitude of changes in scores from Session 1 to Session 2, we calculated the absolute difference between the scores from both sessions for each subject on each measure and averaged them for each of the two groups. We ignored the sign of the differences so that negative differences would not cancel out positive differences when we calculated group statistics. Also, we were interested in *how much* scores changed, not the direction in which they changed.

Figure 3 shows the average absolute change in words per minute from Session 1 (T_1) to Session 2 (T_2) for the set of 10 stimuli, for the two 5-stimulus subsets, and for each stimulus. Change scores for both groups were considerably smaller when they were based on sets of either 5 or 10 stimuli than when they were based on single stimuli. Aphasic subjects' average change scores for sets of stimuli ranged from about six to eight words per minute. When the scores were based on individual stimuli, change scores often doubled or tripled. A similar pattern can be seen for the non-brain-damaged subjects, although their scores were more unstable overall.

Figure 4 shows test-retest stability for correct information units per minute. The pattern is similar to that for words per minute. Both groups' change scores for sets of stimuli generally were considerably smaller than those for individual stimuli. Once again, non-brain-damaged subjects' scores were somewhat more unstable overall than aphasic subjects' scores.

Table 2: Correlations Between Session 1 and Session 2 Scores of Non-Brain-Damaged and Aphasic Subjects

Measure	*Set AB*[a]	*Set A*[a]	*Set B*[a]	*BDAE*[b]
		Non-Brain-Damaged Subjects		
Words per minute	.90	.90	.82	.79
CIUs[c] per minute	.90	.89	.83	.77
% CIUs	.96	.84	.92	.62
		Aphasic Subjects		
Words per minute	.98	.98	.97	.86
CIUs[c] per minute	.97	.94	.97	.75
% CIUs	.98	.93	.94	.71

[a]Set AB contained 10 stimuli; Set A and B contained 5 each. [b]*Boston Diagnostic Aphasia Examination* (Goodglass & Kaplan, 1983) Cookie Theft picture. [c]CIU = correct information unit.

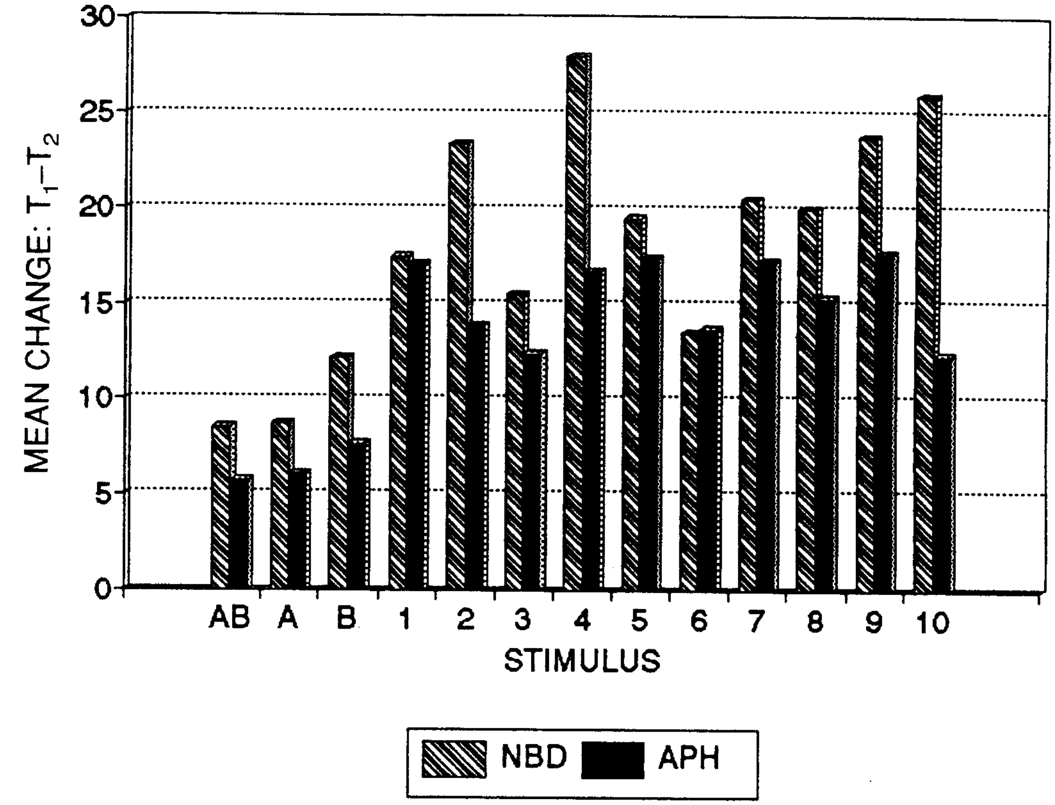

Figure 3. Mean change in words per minute from Test 1 (T_1) to Test 2 (T_2) for non-brain-damaged (NBD) and aphasic (APH) subjects when speech was elicited with Set AB (10 stimuli), Sets A or B (5 stimuli each), or each of 10 single elicitation stimuli (numbers 1–10). (1 = BDAE Cookie Theft picture; 2 = WAB picnic picture; 3, 4 = single pictures; 5, 6 = sequence pictures; 7, 8 = personal information; 9, 10 = procedures.)

Figure 5 shows the results for percent correct information units. The percent CIUs measure was more stable overall than the other two measures. However, as was true for the other measures, the average change from Session 1 to Session 2 generally was greater for individual stimuli than for sets of either 5 or 10 stimuli, especially for aphasic subjects. Their average change scores usually more than doubled when scores were based on individual stimuli rather than a 5- or 10-stimulus set. Non-brain-damaged subjects' scores show a similar relationship; however, their percent CIU scores tended to be more stable overall than the percent CIU scores for aphasic subjects.

To determine whether the five categories of elicitation stimuli differed in test-retest stability, we calculated an average change score for each category. (These average change scores represent change scores that could be expected from testing and retesting with a single stimulus in each

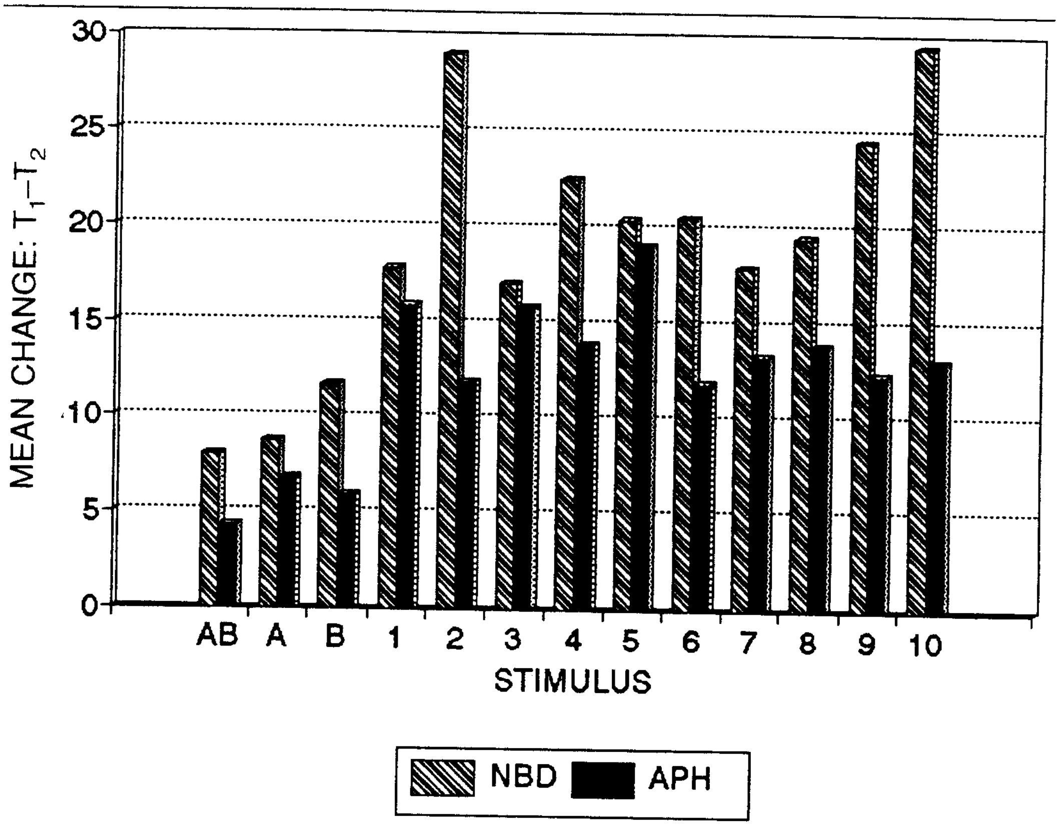

Figure 4. Mean change in correct information units per minute from Test 1 (T_1) to Test 2 (T_2) for non-brain-damaged (NBD) and aphasic (APH) subjects when speech was elicited with Set AB (10 stimuli), Sets A or B (5 stimuli each), or each of 10 single elicitation stimuli (numbers 1–10).

category. They do not represent what would be expected if the speech samples obtained for the two stimuli within each category were *combined.*) The results for words per minute are shown in Figure 6. For aphasic subjects, no category of elicitation stimuli was appreciably more or less stable than any other; they all averaged about a 15-words-per-minute change from Session 1 to Session 2. Non-brain-damaged subjects showed greater variability, with change scores ranging from 18 to 27 words per minute across categories.

Figure 7 shows the results for correct information units per minute. For aphasic subjects the differences among stimulus categories for CIUs per minute were somewhat greater than they had been for words per minute. However, the actual difference between procedures (which yielded the most stable scores) and sequence pictures (which yielded the least stable scores) was less than three CIUs per minute—probably not a clinically significant difference. Again, non-brain-damaged subjects showed greater variability in change scores across stimulus categories than aphasic subjects did.

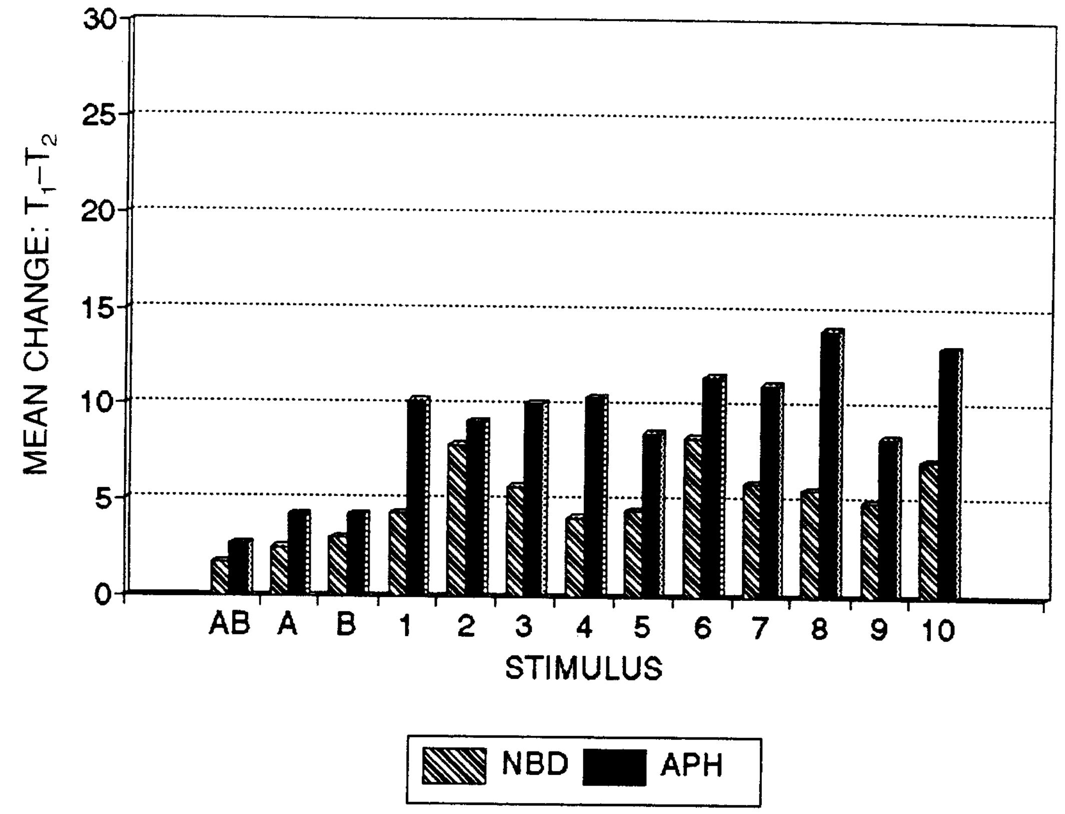

Figure 5. Mean change in percent correct information units from Test 1 (T_1) to Test 2 (T_2) for non-brain-damaged (NBD) and aphasic (APH) subjects when speech was elicited with Set AB (10 stimuli), Sets A or B (5 stimuli each), or each of 10 single elicitation stimuli (numbers 1–10).

Figure 8 shows the results for percent correct information units. Differences in the categories of elicitation stimuli had little effect on the stability of aphasic and non-brain-damaged subjects' percent CIU scores. The largest difference (between aphasia test pictures and requests for personal information for aphasic subjects) was less than 3%.

To go beyond group averages, and to assess the test-retest stability of the measures on an individual-subject basis, we compared each of the 20 aphasic subjects' change scores for samples elicited with the Cookie Theft picture with their change scores for samples elicited with Set A. Only the results for Set A are presented, because the results for Set B were similar, and although differences between Set AB (with 10 stimuli) and the Cookie Theft picture would be greater yet, a 5-stimulus set seemed clinically more practical.

Figure 9 shows the subject-by-subject results for words per minute. In most cases, scores for the *Cookie Theft* picture were more unstable than

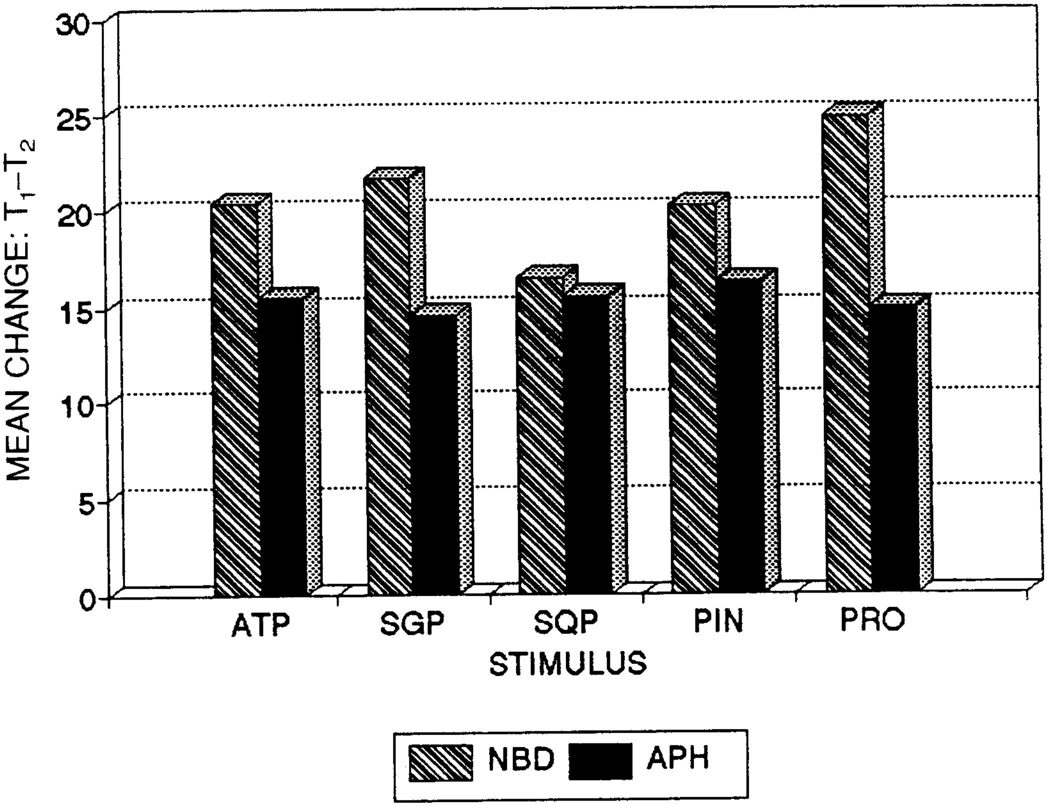

Figure 6. Mean change in words per minute from Test 1 (T_1) to Test 2 (T_2) for non-brain-damaged (NBD) and aphasic (APH) subjects when speech was elicited with aphasia test pictures (ATP), single pictures (SGP), sequence pictures (SQP), requests for personal information (PIN), or requests for descriptions of procedures (PRO).

scores for Set A, with some subjects exhibiting dramatic instability on the *Cookie Theft* picture. Only 3 of 20 aphasic subjects exceeded a 10-words-per-minute difference between Session 1 and Session 2 on Set A, whereas 12 of 20 exceeded this difference on the BDAE picture, including 4 whose change scores exceeded 30 words per minute.

The results were similar for correct information units per minute (Figure 10). Again, scores for the Cookie Theft picture were more unstable than scores for the five-stimulus set, with 4 subjects showing changes of more than 30 CIUs per minute.

Figure 11 shows the results for percent correct information units. Overall, percent CIU scores were more stable than either words per minute or CIUs per minute scores, but scores on the *Cookie Theft* picture again were less stable than scores on the five-stimulus set. Only 1 of the 20 aphasic subjects exhibited a greater than 10% change in percent CIUs on Set A, whereas 11 exhibited changes of this magnitude or more on the Cookie Theft picture.

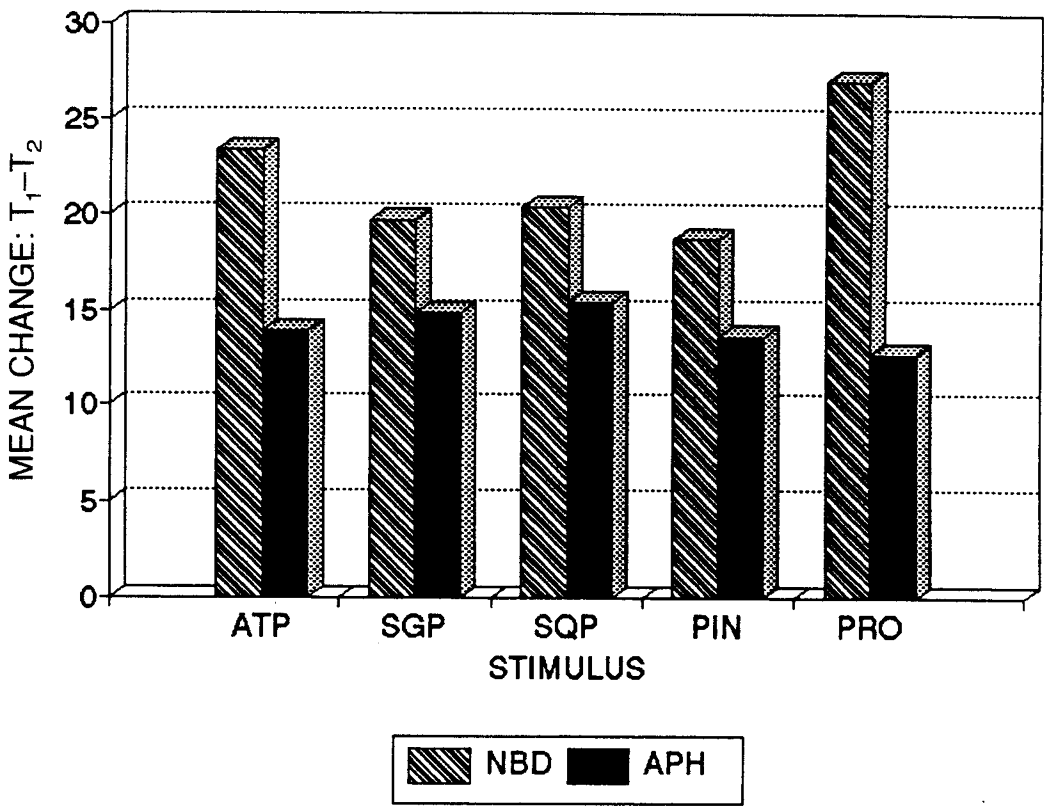

Figure 7. Mean change in correct information units per minute from Test 1 (T_1) to Test 2 (T_2) for non-brain-damaged (NBD) and aphasic (APH) subjects when speech was elicited with aphasia test pictures (ATP), single pictures (SGP), sequence pictures (SQP), requests for personal information (PIN), or requests for descriptions of procedures (PRO).

DISCUSSION

Our results clearly show that one should not make decisions about the connected speech of an adult with aphasia based on measures obtained from one short speech sample, because such measures can be highly unstable from test to test. Because of this instability, a patient's type or severity of aphasia might appear to have changed, even though no actual change has occurred. Likewise, if such short speech samples are used to establish baselines against which the effects of treatment or neurologic recovery are to be measured, actual differences may be obscured by test-retest instability, or spurious differences generated by test-retest instability may be misconstrued as the effects of treatment or neurologic recovery.

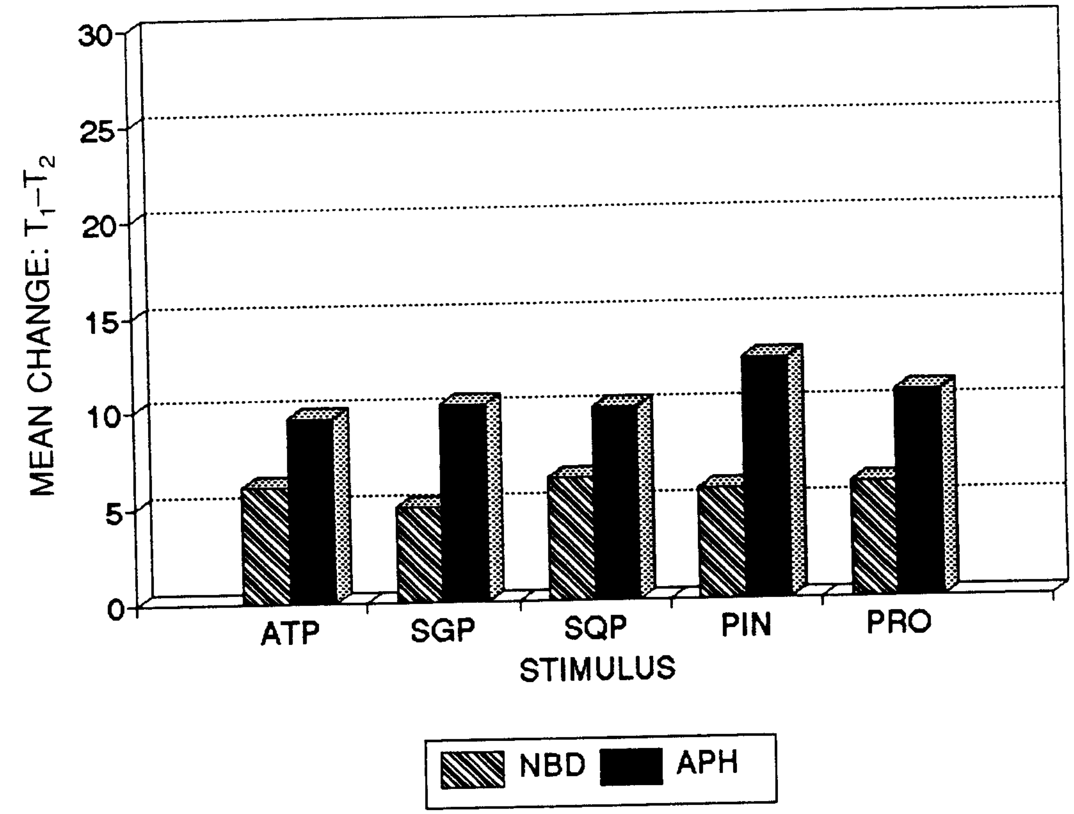

Figure 8. Mean change in percent correct information units from Test 1 (T_1) to Test 2 (T_2) for non-brain-damaged (NBD) and aphasic (APH) subjects when speech was elicited with aphasia test pictures (ATP), single pictures (SGP), sequence pictures (SQP), requests for personal information (PIN), or requests for descriptions of procedures (PRO).

On the other hand, our results show that when measures such as those described herein are based on longer speech samples, they are likely to have adequate test-retest stability. We do not know, at this time, how much one could reduce the number of elicitation stimuli below five and still maintain acceptable stability. Because transcribing and scoring connected speech samples can be time consuming, it seems clinically important to analyze only a large enough sample to ensure representativeness and stability. Our results suggest that there is substantial gain in stability by moving from 1 to 5 stimuli and that going from 5 to 10 stimuli yields substantially smaller gains. When the time needed to transcribe and score the extra 5 stimuli is considered, that small gain may not be worthwhile. At this time we do not know whether samples based on responses to 2, 3, or 4 stimuli would yield acceptably stable measures, but we plan to explore that possibility.

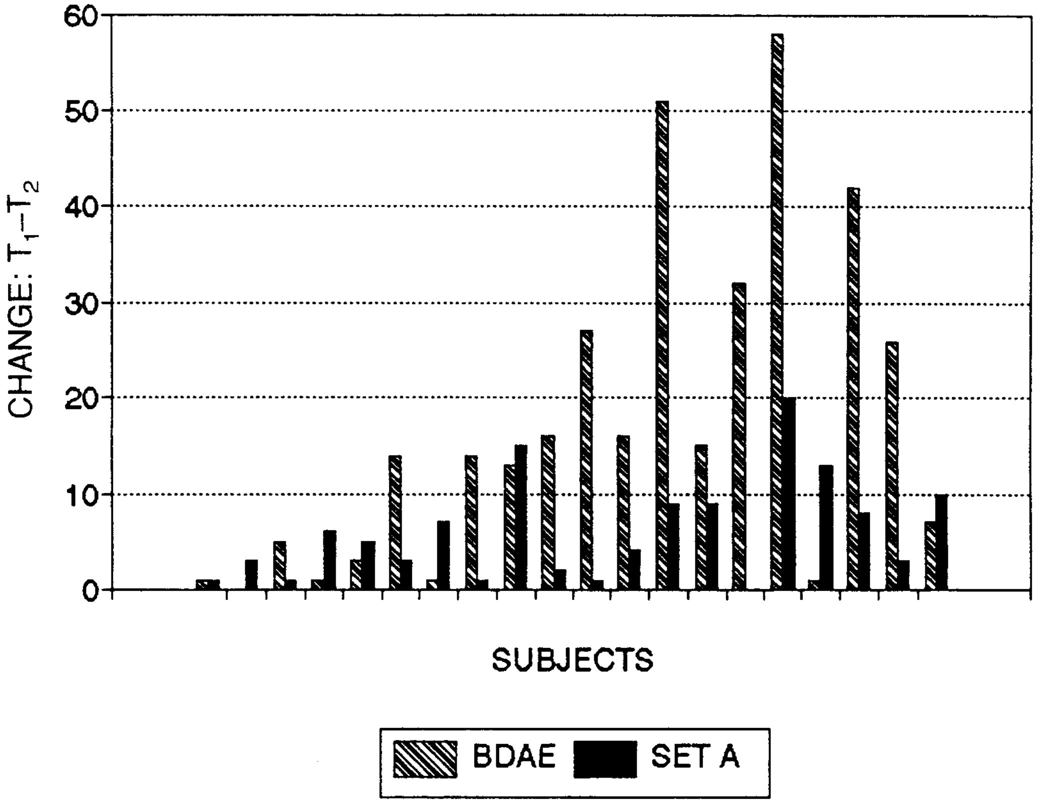

Figure 9. Subject-by-subject absolute change scores for words per minute when speech was elicited with the *Boston Diagnostic Aphasia Examination* Cookie Theft picture (BDAE) or with Set A (5 elicitation stimuli).

Although the results of this study do not speak directly to the issue of representativeness, we feel that speech samples composed of responses to a variety of elicitation stimuli are likely to be more representative of an individual's everyday connected speech than are samples based on a single type of stimulus. At this point, we believe that a combined sample representing an aphasic speaker's responses to the five types of elicitation stimuli described herein should, in most cases, satisfy the need for both stability and representativeness.

ACKNOWLEDGMENTS

This research was supported by the Department of Veterans' Affairs Medical Research Service and by the Medical Research Service of the Minneapolis VA Medical Center.

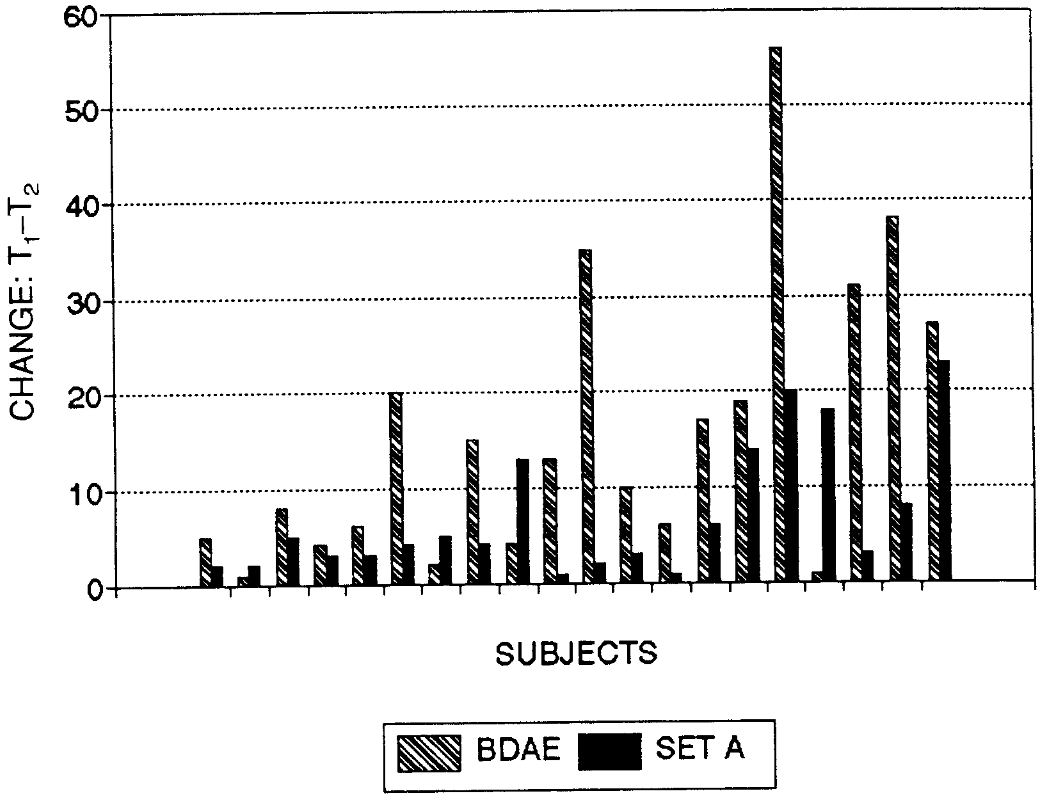

Figure 10. Subject-by-subject absolute change scores for correct information units per minute when speech was elicited with the *Boston Diagnostic Aphasia Examination* Cookie Theft picture (BDAE) or with Set A (5 elicitation stimuli).

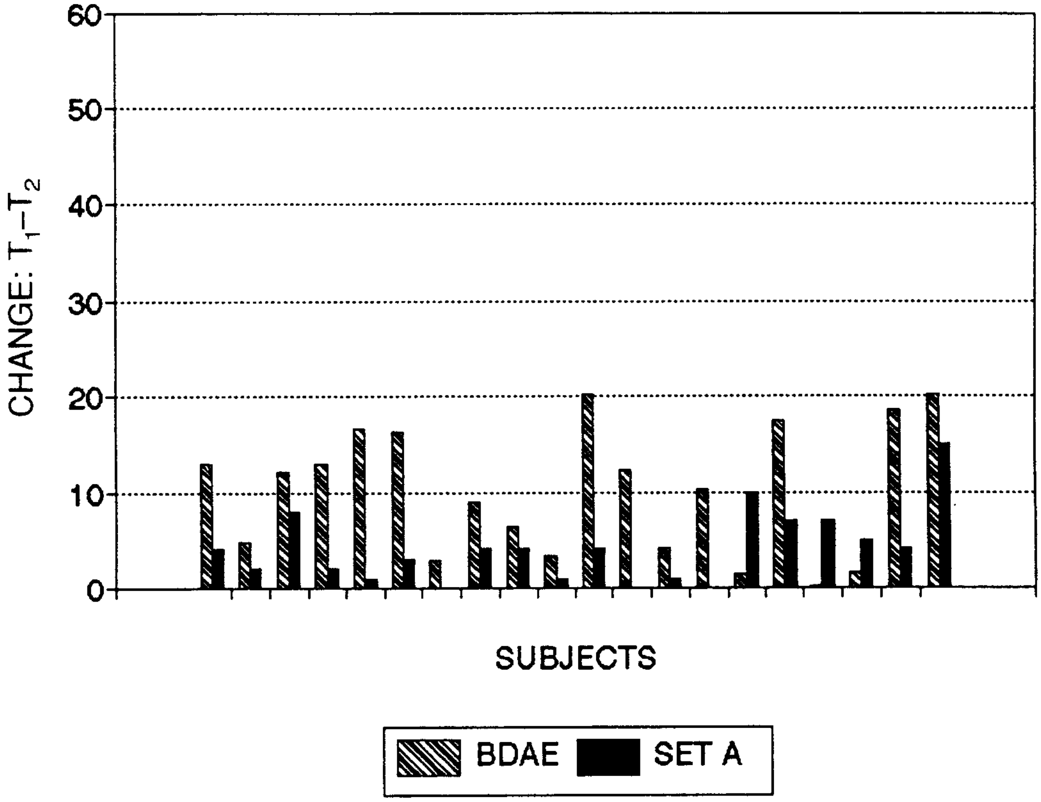

Figure 11. Subject-by-subject absolute change scores for precent correct information units when speech was elicited with the *Boston Diagnostic Aphasia Examination* Cookie Theft picture (BDAE) or with Set A (5 elicitation stimuli).

REFERENCES

Disimoni, F. G., Keith, R., & Darley, F. L. (1980). Prediction of PICA overall score by short versions of the test. *Journal of Speech and Hearing Research, 23,* 511–516.

Goodglass, H., & Kaplan, E. (1983). *The Boston Diagnostic Aphasia Examination.* Boston: Lea & Febiger.

Hess, C. W., Haug, H. T., & Landry, R. G. (1989). The reliability of type-token ratios for the oral language of school age children. *Journal of Speech and Hearing Research, 32,* 536–540.

Hess, C. W., Sefton, K. M., & Landry, R. G. (1986). Sample size and type-token ratios for oral language of preschool children. *Journal of Speech and Hearing Research, 29,* 129–134.

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

Miller, J. F. (1981). *Assessing language production in children: Experimental procedures.* Baltimore: University Park.

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

Porch, B. E. (1971). *The Porch Index of Communicative Ability,* Palo Alto, CA: Consulting Psychologists Press.