Reliability of Rating Spontaneous Speech in the Western Aphasia Battery: Implications for Classification

Elizabeth Hillis Trupe
The Good Samaritan Hospital, Baltimore, Maryland

PROBLEM

"A test, to be a useful research and clinical took, must meet acceptable standards of reliability," stated Shewan and Kertesz (1980, p. 312) in their article on reliability and validity characteristics of the Western Aphasia Battery (Kertesz, 1982). They went on to say that reliability can be viewed as "a measure of the ratio of the true variance to the total variance in test scores" (p. 312). In the same article, the authors documented statistically significant correlations between test-retest scores and between scores of paired judges for the same test for all Western Aphasia Battery (WAB) subtests. However, both intrajudge and interjudge reliability measures were lowest for the Spontaneous Speech subtest content and fluency scores. It is these scores that have the greatest "weight" in calculating the Aphasia Quotient (AQ) and the Cortical Quotient (CQ). The AQ and the CQ allow quantification and simplicity in reporting level of function to other caretakers, but the usefulness of these scores depends on their reliability in measuring behavior. Furthermore, the fluency score is critical in the taxonomic approach to classifying aphasia proposed by Kertesz (1979). For example, a fluency score of 4 or 5 can differentiate between Broca's and Wernicke's aphasia. If one believes that there is a real difference between Broca's and Wernicke's aphasia, then there must be a real difference between fluency scores of 4 and 5 for this classification system to be meaningful. The impetus for this investigation was the observation that a single speech sample was assigned fluency scores of greater than 5 and less than 5 by separate examiners. This variation in test scores represented disagreement between scorers rather than a true difference in behavior. A desire for greater consistency in defining severity and classification of aphasic patients lead to the following studies.

METHOD

This investigation consisted of three major stages: a preliminary study, an experimental study, and two follow-up studies. Elements of each study are summarized in Table 1. The preliminary study involved assessment of interjudge agreement in scoring WAB Spontaneous Speech subtests. Results led to analysis of scoring criteria, clarification of terms, and revision of the scales to reflect clarifications. Subsequently, an experimental study was designed to determine whether clarifications result in improved reliability. Follow-up studies were conducted to replicate results with new samples and new judges.

Preliminary Study

Results. Variance between judges in rating one parameter of a single sample was measured by calculating the estimated variance as shown in Table 2. Average variance between judges approximated one for each score. Since a variation of one point on a fluency score can result in different diagnostic

Table 1. Summary of studies.

	PURPOSE	MATERIALS	JUDGES	PROCEDURE
Preliminary Study	Assess interjudge agreement in rating content and fluency of spontaneous speech, according to WAB scoring criteria	1) 20 Transcribed WAB spontaneous speech subtests (Set A) 2) Published scoring criteria	5 Speech-language pathologists on one hospital staff	1) Independent scoring of subtests by published criteria 2) Assessment of interjudge agreement 3) Analysis of data
Experimental Study	Assess interjudge agreement in scoring content and fluency using revised criteria as compared to published WAB criteria	1) 10 Transcribed subtests (subset of Set A) 2) Revised scoring criteria	Same as above	1) Independent scoring of subtests by revised criteria 2) Assessment of agreement 3) Comparison of agreement to that obtained using published criteria
Follow-Up Study	Address weaknesses of experimental study and confirm earlier conclusions that revised criteria result in improved agreement	1) 10 Taped and transcribed subtests (Set B) 2) Published and revised scoring criteria	Same as above	Same as above
Replication	Replicate experimental study, with different judges and different speech samples, to confirm conclusions	Same as above	5 Speech-language pathologists from community settings other than above hospital	1) Independent scoring by published criteria 2) Introduction of revised criteria 3) Re-scoring by revised criteria 4) Comparison of interjudge agreement using published and revised criteria

classifications, according to Kertesz's taxonomy, a routine variation of one point is unacceptable.

Table 2. Estimated variance within a patient.*

Score	Estimated Variance	Standard Deviation
Content	0.99	.995
Fluency	0.90	.95

*Estimated variance = $\frac{\text{Within sum of squares}}{\text{Degrees of freedom (80)}}$ S.D. = $\sqrt{\text{Variance}}$

Having determined that the attained level of consistency between judges was less than desirable, the data were analyzed to determine the points of disagreement between judges. It was hypothesized that: 1) one or more of the examiners may have a tendency to score either "high" or "low" relative to other examiners; 2) the criteria for one or more scores (on a scale of 0 - 10) may be weaker or "more nebulous" than for other scores.

Table 3 illustrates the test of the first hypothesis. The mean content score and the mean fluency score of each judge were calculated and compared to the overall mean scores. Table 3 indicates that one examiner had a tendency to rate both content and fluency lower than the other examiners, but this difference did not reach statistical significance. There was no significant difference between the means for judges for content ($F_{4,76}$ = 1.94, p = .11) or for fluency ($F_{4,74}$ = 2.23, p = .07).

Table 3. Comparison of judges' mean scores.

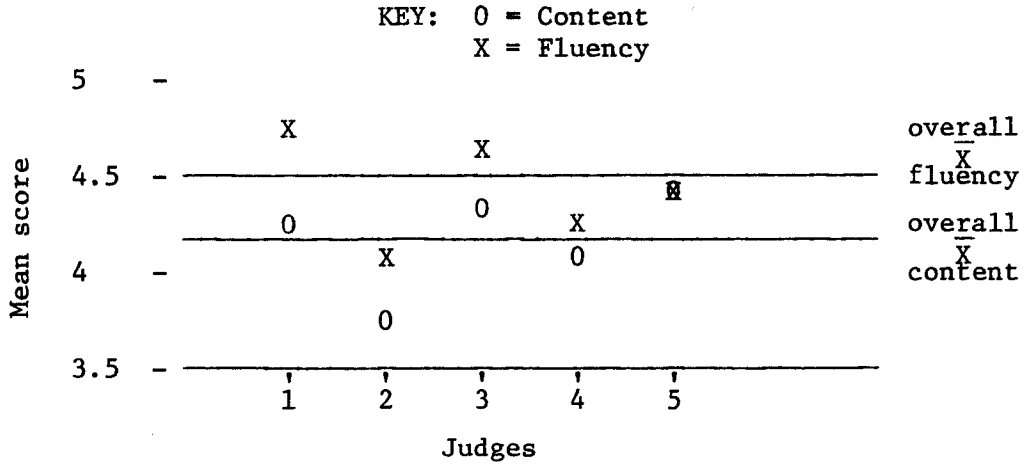


Table 4 and 5 were constructed to test the second hypothesis. The median content score and the median fluency score for each sample were calculated. Median scores were plotted against the range of scores for that sample, to determine which scores were associated with the greatest range or discrepancy. Each "X" on the tables represents a single sample, or Spontaneous Speech subtest. For example, four samples had a median content score of 0; three of these samples had a range of scores of 2.

That is, for these three samples, the middle score assigned by the five judges was 0, while the highest score assigned was 2. It was assumed that if specific median scores were associated with a large range of scores (wide discrepancies), then criteria for those particular scores would need to be clarified. It was generally found that only scores on the upper end of the scales were associated with low (acceptable) ranges, while scores in the middle of the scale were associated with ranges of up to 4 points. Content scores of 5 and 7 and fluency scores of 3, 4, and 6 showed the greatest discrepancies.

Table 4. Median content score.

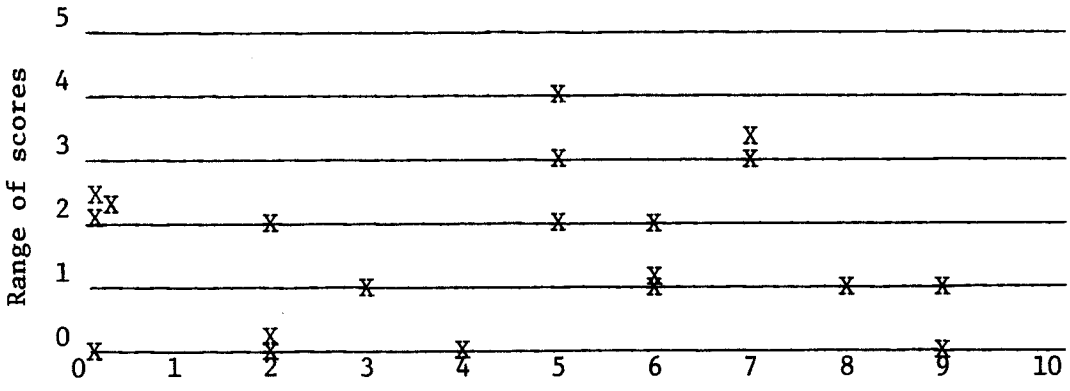
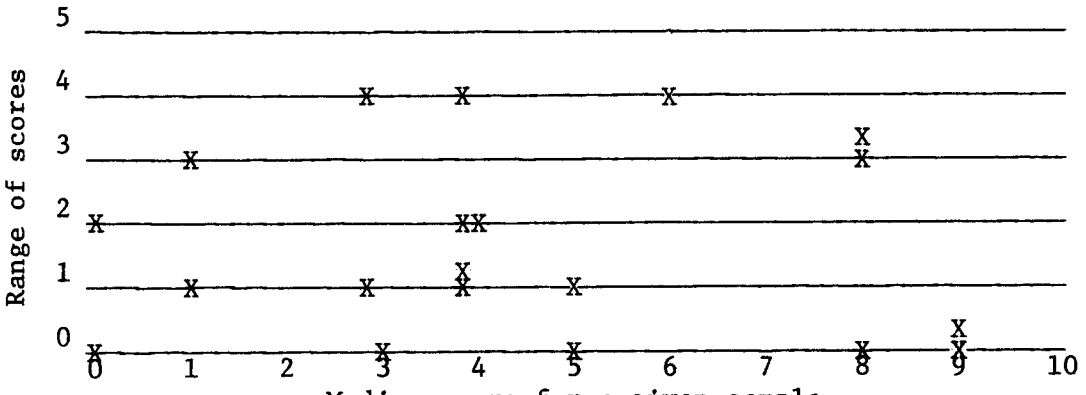


Table 5. Median fluency score.



Conclusions. 1) There was poor agreement among examiners in scoring the WAB Spontaneous Speech subtest using the published criteria. 2) Discrepancies were not associated with a single judge or with a single score but were widespread across judges and across the scales. 3) Clarification of the scoring criteria, particularly for those scores associated with the highest range across examiners, was necessary to achieve consistent diagnostic measurement and reporting.

KEY: 0 = uncharacteristic of pt.'s speech
+ = characteristic of pt.'s speech
blank = not specified by scoring criteria

TABLE 6 *script = inferred by criteria*

LARGE PRINT = STATED OR DIRECTLY IMPLIED BY CRITERIA

SCORE:	0	1	2	3	4	5	6	7	8	9	10
Telegraphic Speech	0		+ SINGLE WORDS *only*			OFTEN					0
Propositional sentences	0	*0*	*0*	*0*	0 AUTOMATIC SENTENCES ONLY	+ FEW	+ MORE		+	MOSTLY +	+
Hesitations	+		+		+			*0*	*0*	+ OCCASIONAL	0
Recurrent Utterances	+ *occasional* SHORT	+ STEREOTYPIC		+ FLUENT, *non-stereotypic*				*0*	*0*	*0*	*0*
Circumloc and/or Paraphasia	*0*	*0*	+		+	+ MAY BE PROMINANT	+ MAY BE PRESENT		+	+ OCCASIONAL	0
Fluent Phonemic Jargon	*0*	*0*	*0*	+ LOW VOLUME	*0*	*0*		+	*0*	*0*	*0*
Varied Phonemic Pattern		*0*		+ *low variability*		+	+	+ *high variability*	+	+	+
Fluent Semantic Jargon	*0*	*0*	*0*	*0*	*0*	*0*			+	*0*	*0*
Grammatical Competence					OCCASIONAL VERBS OR PREPOSITIONS	SOME GRAMMATICAL ORGANIZT'N	NORMAL SYNTATIC PATTERN MAY BE PRESENT	SEMBLANCE TO ENGLISH SYNTAX	SENTENCES OFTEN COMPLETE	MOSTLY COMPLETE SENTENCES	SENTENCES OF NORMAL LENGTH & COMPLEXITY
Other Characteristics	MEANINGLESS	VARIED INTONATION, CONVEYING SOME MEANING	EFFORTFULL					SEMBLANCE TO ENGLISH RHYTHM	SENTENCES MAY BE IRRELEVANT	MAY HAVE ARTICULATION ERRORS	WITHOUT SLOWING OR ARTICULATION ERRORS

Discussion. The findings of the preliminary study led to revision of the scoring criteria for the purpose of achieving improved consistency in diagnostic measurement. Procedures for revising the scales follow.

1) Published content and fluency scales were each analyzed as a whole, revealing a large number of variables included in each rating scale.
2) Five individual samples which were assigned a wide range of scores were discussed on a case-by-case basis.
3) An additional five samples were collaboratively rescored and justification for assigning specific scores was presented.
4) Criteria for each of the scores associated with the greatest disagreement were discussed.
5) Scoring criteria were revised to reflect clarifications and simplification of terminology. The outcome of each of these procedures is summarized below.

Analysis of published scales revealed that each of the two scales includes multiple variables. That is, each score involves rating several different speech behaviors. Table 6 represents a "distinctive feature" analysis of fluency scores only. Rating of speech fluency according to test criteria depends on presence, absence, or degree of telegraphic speech, proposition, meaningfulness, relevance, articulation, recurrent utterances, paraphasias, jargon, and grammatical competence. These characteristics are not necessarily correlated in aphasic patients. Hence, a single patient at one moment in time may exhibit features of several different fluency scores when all variables are assessed.

Case-by-case discussion of individual samples revealed differences in emphasis on certain variables among examiners. For example, a patient who correctly answered all biographical information questions but was unable to produce any meaningful picture description was assigned different content scores, depending on the examiner's focus. Similarly, a speech sample consisting of both telegraphic speech and islands of fluent, phonemic jargon was assigned different fluency scores, depending on the examiner's focus. Therefore, agreement on the "weight" of certain aspects and elimination of superfluous variables were targeted to improve consistency.

Collaborative scoring revealed that examiners tended to assign "half credit" for questionable or incomplete responses to biographical information questions and then rated content on a sum of full credit and half credit responses. However, there was wide variability in when and how often half credit was given, implying a need for consistent criteria for assigning and weighting partial credit.

Discussion of individual criteria revealed differences in interpretation of terminology. Relative terms, including "some," "more," "occasional," "few," "mostly," "often," "prominent," and "marked" were subject to varied interpretation, as were descriptive terms such as "stereotypic," "recurrent," "incomplete," and "recognizable" phonemic paraphasias. Clarification of terms, or substitution of more concrete terms, was considered potentially useful in reducing disagreement in this area.

The scoring revisions in Tables 7 and 8 were proposed to address problematic aspects of the published criteria as described above. The formats of the scales were modified for the sake of clarity. Rather than descriptions of each score or level of content, specific directions for scoring content were formulated. Since administration of the WAB requires assessment of the behaviors that the battery claims to assess, elimination of major variables from the fluency scale was ruled out. Instead, the scale was broken down into two separate scales which rate essentially different types of verbal output. Reliability of revised criteria was then assessed in the following study.

Table 7. WAB Scoring Clarification: Content

Instructions:

A. Score verbal responses only (if necessary, cue patient to report verbally)

B.

Give half credit for:		Give full credit for:
#3 first or last name only		
#4 Intelligible street name only (with or without city, state, zip code)		street name plus house number (with or without city, state, zip code)
#5 "retired" without ability to state previous occupation after cueing	OR	"retired" plus accurate report of previous occupation after cueing previous/current occupation without cueing
#6 incomplete message about diagnosis or rehab. goals		statement of diagnosis (e.g., "stroke") or rehab. goal (e.g., "for therapy")

C. Score intelligible (recognizable phonemic paraphasias) responses as correct (full credit).

D. Use the following table to assign content scores:

No. Points Correct on First 6 Items	Clarification of "Some" Response to Picture: No. of Items/Actions Identified	Content Score
0	N/A	0
½	N/A	1
1	N/A	2
1½-2	N/A	3
2½	N/A	4
3	<2	4
3	≥2	5
3½	<4	5
3½	≥4	6
4	<6	6
4	≥6	7
4½	N/A	7
5	N/A	8
5½	<10	8
5½	≥10	9
6	almost complete description	9
6	complete description (sentences of normal length & complexity)	10

Find the sum of full & half credit "points" in the left column, then move to the right of the table to locate the corresponding content score. Use the middle column only to differentiate between two scores corresponding to sum of points correct.

Table 8. WAB Scoring clarification: Fluency

High content relative to fluency:

0	2	4	5	6	9	10
No words	Single words (only) often paraphasias, effortful and hesitant	Telegraphic speech only No propositional sentences (some 2-3 word (combos)	Telegraphic speech with few propositional sentences (<25% of utterances)	Some telegraphic speech with more propositional sentences (≥25% of utterances)	Complete propositional sentences with occasional hesitations, paraphasias, or circumlocutions (No telegraphic speech)	WNL

Low content relative to fluency:

0	1	3	7	8	10
(Occasional) short meaningless utterances	1 or 2 recurrent stereotypic utterances (e.g., "wo, wo, wo", "no, no", "dididi" - varied intonation without varied phonemes)	Fluent, recurrent utterances (with some phonemic variation - not stereotypic) or Low volume mumbling (without semblance to English rhythm)	Phonemic jargon, (not recurrent) Voluable Semblance to English rhythm and syntax	Semantic jargon or Circumlatory, fluent speech with frequent verbal paraphasias	WNL

Experimental Study

Results. The estimated between-judges variation was significantly lower for scoring both content and fluency by revised criteria than by published criteria, as demonstrated in Table 9. A variance ratio test, or F test, was used to determine statistical significance.

Table 9. Estimated variance.

SCORE	PUBLISHED SCORING	REVISED SCORING	F RATIO	STATISTICAL SIGNIFICANCE
Content	0.37	0.19	1.947	$p < .05$
Fluency	0.54	0.09	6.0	$p < .001$

Conclusion. Revision or clarification of scoring criteria resulted in improved agreement between examiners in scoring content and fluency of a given speech sample elicited in the WAB Spontaneous Speech subtest.

Discussion. While statistical significance of these results was remarkable, the following limitations of the experimental study were determined.

1) Since scoring revisions were based on data that included scored samples of the study population, revisions could be specifically applicable to this group of patients.
2) Transcribed, rather than taped, speech samples were used, while taped samples would better replicate the real-life testing situation, particularly for judging fluency of speech.
3) Because the judges were involved in the development of scoring revisions, it was not known whether improved agreement was due to the actual revisions or due to extensive discussion among the five judges regarding clarification of terms and analysis of scales.

Therefore, two follow-up studies were conducted to address the above weaknesses. Both studies incorporated taped speech samples of a new population. The second study involved five new judges.

Follow-up Studies

Statistical analysis of the data from both studies is demonstrated in Table 10. A comparison of these results with results from previous studies is illustrated in Figure 2.

Table 10. Estimated variance.

	FOLLOW-UP STUDY 1:	FOLLOW-UP STUDY 2: NEW JUDGES			
Score	Original Judges Using Revised Criteria	Using Published Scales	Using Revised Scales	F Ratio (K/R)	Statistical Significance
Content	0.13	0.82	0.18	4.54	$p < .001$
Fluency	0.64	1.14	0.81	1.4	N.S.

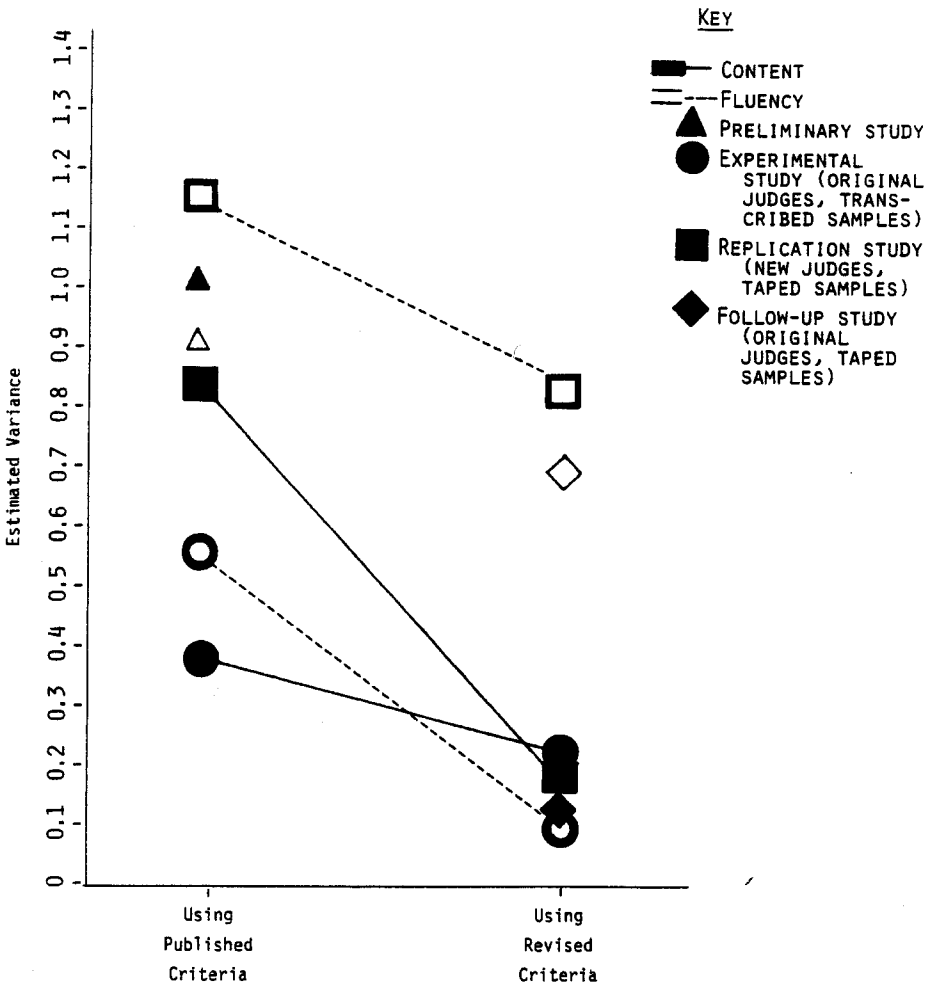


Figure 1. Results of experimental study.

Conclusions.

1) Low variance between original examiners was maintained in scoring content with the revised scales when taped speech samples were judged.
2) There was relatively high variance or poor agreement in judging fluency of these samples.
3) The general trends observed in the experimental study were replicated. That is, use of the revised scales resulted in less variation between observers in scoring both content and fluency, regardless of the population or use of taped speech samples. However, while reduced variation was highly significant for scoring content, it was not significant for scoring fluency.

Discussion. Both groups of judges exhibited difficulty in scoring fluency of this set of Spontaneous Speech subtests. The following example of the WAB picture description, which was assigned a variety of fluency scores, reveals the various possibilities in rating "fluency" according to test criteria:

"Well see picture. I see car and I see ren, I see ren, wen, ren, ren, one ren, ren--and that's a man. That's a man. 1, 2, 3, 4, 5 - What else? [unintelligible jargon] - This one's got a crate in it and I know what it - [QUESTION] I think he [unintelligible] - shrinking, swinking, that's it, shrinking. [QUESTION] He's [unintelligible jargon] - it's fishing boy -- it's fishing boy, but he must have done [unintelligible].

A diversity of speech characteristics associated with different WAB fluency scores can be identified in the above sample, as demonstrated in Table 11. Since so many discrete variables are included in a single scale, several fluency scores could be justified for such patients, precluding reliable rating. Use of revised scales improved interjudge agreement to a limited extent.

Table 11. Sample of aphasic patient's speech.

SPEECH CHARACTERISTIC	EXAMPLE	CORRESPONDING WAB FLUENCY SCORE
Fluent, recurrent utterances	"I see ren, I see ren, ren, ren"	3
Automatic sentences	"What else?"	4
Telegraphic speech	"Well, see picture"	4,5
Few propositional sentences	"That's a man"	5
Paraphasias	"Swinking"	4,5,6
Phonemic jargon	Unintelligible jargon	7
Circumlocutory, fluent speech, with marked word-finding difficulty	"This one's got a crate in it I think and I know what it"	8

CLINICAL IMPLICATIONS

Consistency among examiners in rating spontaneous speech with the WAB can be improved by clarification of scoring instructions and specification of terms, as was demonstrated by improved interjudge agreement achieved in each of the studies conducted. In addition, it was shown that discussion and collaborative scoring among examiners in a single setting results in further improvement in internal consistency. However, a multidimensional scale such as the WAB fluency scale, with or without revisions, cannot reliably be used to score spontaneous speech, because speech dimensions vary independently in aphasic patients. A single speech sample may simultaneously exhibit features of several different levels of fluency and thus result in a number of justifiable fluency scores. Therefore, independent dimensions of speech (e.g., jargon, grammatical competence) must be described or rated separately in order to have diagnostic value to other professionals. Such an approach is incorporated in the Boston Diagnostic Aphasia Examination (Goodglass and

Kaplan, 1972), in which melodic line, phrase length, and other variables are scored independently.

In the taxonomy proposed by Kertesz, a fluency score of greater than or less than 5 results in two different diagnostic classifications. Since the fluency scale lacks interjudge reliability, the classification system dependent on the fluency score is also unreliable. It therefore does not meet the requirement for a "useful research and clinical tool," given by Shewan and Kertesz (1980, p. 312). Inconsistent classification has negative implications for research, evaluation, reporting, and measuring progress of aphasic patients.

ACKNOWLEDGMENT

The author wishes to thank Patricia Linden, Marcia Roseman, Andrew Jinks, Ann Shaughnessy, Martha Hersey, Rhona Paul Cohen, Stuart Trippe, Denise Taneyhill, Kelly Whitman, and Carla Gress for their participation in rating speech samples for this study and Argye Hillis and Helen Abbey for their assistance in statistical analysis of the data. Presentation of this paper was financially assisted by the Division of Rehabilitation Medicine of The Johns Hopkins University School of Medicine.

REFERENCES

Goodglass, H. and Kaplan, E., The Boston Diagnostic Aphasia Examination. Philadelphia: Lea and Febiger, 1972.

Kertesz, A., Aphasia and Associated Disorders: Taxonomy, Localization, and Recovery. New York: Grune & Stratton, 1979.

Kertesz, A., Western Aphasia Battery. New York: Grune & Stratton, 1982.

Kertesz, A. and Poole, E., The aphasia quotient: The taxonomic approach to measurement of aphasic disability. Canadian Journal of Neurological Science, 1, 7-16, 1974.

Shewan, C.M. and Kertesz, A., Reliability and validity characteristics of the Western Aphasia Battery (WAB). Journal of Speech and Hearing Disorders, 45(3), 308-324, 1980.

APPENDIX

Published Scoring of Spontaneous Speech*

A. Information Content

(0) No information.
(1) Incomplete responses only, e.g., first name or last name only.
(2) Correct response to any 1 item.
(3) Correct responses to any 2 items.
(4) Correct responses to any 3 items.
(5) Correct responses to any 3 of the first 6 items plus some response to the picture.
(6) Correct responses to any 4 of the first 6 items plus some response to the picture.
(7) Correct responses to 4 of the first 6 items on page 2 and a mention of at least 6 of the items in the picture.
(8) Correct responses to 5 of the first 6 items, and in incomplete description of the picture. Recognizable phonemic paraphasias are to be counted as correct.

(9) Correct responses to all 6 items on page 2. An almost complete description of the picture: at least 10 people, objects, or actions should be named. Circumlocution may be present.
(10) Correct responses to all 6 items on page 2 and to the picture. Sentences of normal length and complexity, referring to most of the items and activities. A reasonably complete description of the picture.

B. Fluency, Grammatical Competence, and Paraphasias

(0) No words or short, meaningless utterances.
(1) Recurrent stereotypic utterances with varied intonation, conveying some meaning.
(2) Single words, often paraphasias, effortful and hesitant.
(3) Fluent recurrent utterances or mumbling, very low volume jargon.
(4) Halting, telegraphic speech. Mostly single words, often paraphasic but with occasional verbs or prepositional phrases. Automatic sentences only, e.g., "Oh I don't know."
(5) Often telegraphic but more fluent speech with some grammatical organization. Paraphasias may be prominent. Few propositional sentences.
(6) More complete propositional sentences. Normal syntactic pattern may be present. Paraphasias may be present.
(7) Phonemic jargon with semblance to English syntax and rhythm with varied phonemes and neologisms. May be voluble; must be fluent.
(8) Circumlocutory, fluent speech. Marked word finding difficulty. Verbal paraphasias. May have semantic jargon. The sentences are often complete but may be irrelevant.
(9) Mostly complete, relevant sentences; occasional hesitation and/or paraphasias. Some word finding difficulty. May have some articulatory errors.
(10) Sentences of normal length and complexity, without definite slowing, halting, or articulatory difficulty. No paraphasias.

*Acknowledgment: This is a portion of the Western Aphasia Battery (Grune & Stratton, 1982), reprinted by permission of the author, Dr. Andrew Kertesz, and the publishers.

DISCUSSION

Q: How much training did the judges have to have to learn to do the revised system?
A: None. I gave it to them in printed form.

Q: Do you think it is worth it to revise to get increased agreement?
A: I felt that it was worth it in our hospital since we frequently administer the WAB, and sometimes a patient may move from one clinician to another. Then, the retesting we do is not meaningful unless we have an idea of how the other person was trying to score. That's where the problem came up, and that's where it's been useful to us to have this revised scoring.

Q: Perhaps I should preface my remarks by saying that I'm Shewan. I wondered how much training your judges had had, if they had had any. Perhaps I should clarify that we did get very high reliability, both within and across our judges. Our judges did have some training prior to scoring the 10 videotape samples. So that may account for some of

the decreased variability which we found, both within and across our judges. We got extremely high reliability across 8 judges and extremely high reliability within judges. I agree with you--the fluency score was the one parameter on which we obtained the lowest amount of reliability, but it still was very high. I think that there are problems with the fluency scale, but I think perhaps one of the reasons that our reliability was so high is that we were all rating with the same kind of prior knowledge, in terms of having familiarity with the battery. That was the second question I was going to ask. How much familiarity with the WAB did you people have when you were scoring?

A: Over the past year or longer, each of us in that hospital has administered the WAB at least twice a month. I don't know what training each person had in the WAB before they worked there. They were familiar with it and used it often, but they didn't have Dr. Kertesz to tell them how it was supposed to be done. An additional reason for the higher variability I found was that I analyzed reliability in a different way. As I understand your article, you looked at correlations, or relationships between scores of paired judges rather than point-to-point agreement. Hypothetically, if one of your judges had consistently scored 2 points lower than the other, there would have been a perfect correlation. I was looking at the exact amount of agreement or disagreement between judges, rather than a relationship or tendancy to rate in the same direction.

C: When I give the WAB, I don't expect that the verbal behavior will be homogeneous--that it will necessarily fit into one scale exclusively. And as with any test, I have a concern that we become too score oriented--when we hear the speech, we think of score first and impression second. We should get an impression first, eliminate some of those choices on the 10-point scale, and then score on the predominant characteristics.

A: I feel strongly that each of those verbal behaviors or variables needs to be described or rated separately in order to have any value to other caretakers. I think that we need to be tuned in to each of those speech dimensions while the person's talking and not to a score. So I would agree with that. But I think that the score doesn't capture very much about the person's speech, because it doesn't tell you anything about those variables, specifically.

Q: You said that a major reason for doing the study is that just a difference of one point on the scale can make a difference in classification. Did you mention in what percentage of the patients you looked at that actually did make a difference in classification? Because the difference of one point is important only at one point in the scale.

A: Right. One point makes a difference only between scores of 4 and 5. I didn't mention those data--I don't have them with me. But it did make a difference in quite a number of patients. When I showed the range of scores associated with specific median scores, you could see that there was never exact agreement on a fluency score of 4. There was at least one patient who was given a median score of 4 and a score of 8. I don't remember in what percentage the variability made a difference in classification, but that would be useful information to document. (Author's note: Based on interjudge variability in scoring fluency, there would have been a discrepancy in classification in 7 of the original 20 patients and in 5 of the second group of 10 patients, using Kertesz's taxonomy.)

Q: I'd like to ask what your clinical impression was of the structure of the task that was used to elicit connected language in order to judge fluency. Do you think that looking at a picture with isolated events affects the fluency vs. another type of task that you would use to get a sample of connected language?

A: Yes. I think it affects fluency. There were patients, as I mentioned, who were able to say some meaningful things in response to questions or in conversation but looked at the picture and couldn't come up with anything meaningful to say. Or they would only respond when you pointed to specific items and asked, "What's this?" Then they'd say single words, like "dog." And yet in response to conversation, they spoke in sentences.

Q: On the new samples that you tested, did you use the published scale and then your new revised scale?

A: Yes.

Q: I wonder if familiarity with your patient influenced better scoring with the revised scale?

A: The judges were free to stop the tape and listen to it again and again when they were scoring as we do clinically. So people were as familiar as they wanted to be with the tape when they scored it by either method.

Q: But you didn't take that into consideration and try to control for that. I wonder if you controlled for order, too. Did they always use the old scale first and then the revised scale?

A: They always used the published scale first because I didn't want them to know what my revisions were yet, since my revisions were clarifications of the published scales. I wanted them to use the published scales as they were and then find out how I had tried to clarify them and use that. I didn't want my clarifications to bias the way they interpreted the published scales.

Q: But you don't feel that the fact that they were perhaps more familiar with the tape is why you got better reliability with your revisions?

A: I'm not sure how it would affect the reliability.

Q: But that's not something that you looked at?

A: No.

Q: Do you feel that validity is the real problem in the scoring system on the fluency part of the WAB; and if so, do you have any thoughts on the matter of how we might go about examining fluency with more validity?

A: I'm anxious to see what Dr. Shewan has to say on Wednesday. I think that it's impossible to give one fluency score that incorporates all variables of aphasic speech. I think that you really need to look at each of those speech dimensions separately. And I don't think that you can come up with a valid overall fluency measure that takes into account paraphasias, grammatical competence, effort, and propositionality, simultaneously.