Introduction

Reliability in aphasia testing is a crucial issue. Specifically, intra- and inter-rater reliability in scoring is an important concern for tests that employ multidimensional scoring. The Revised Token Test (McNeil & Prescott, 1978) uses a multidimensional scoring system, wherein a single score describes a response in terms of several dimensions: accuracy, responsiveness, promptness, completeness, and efficiency (McNeil & Prescott, 1978; Porch, 1967). Such a system helps examine the nature of responses in detail, which in turn aids in planning treatment and predicting recovery (McNeil & Prescott, 1978; Porch, 1967). Although inter-rater reliability in scoring for this test has been investigated in the past (McNeil & Prescott, 1978), the nature and pattern of scoring discrepancies that arise between raters has not been documented.

The purpose of this study was to explore inter-rater agreement in RTT scoring by providing descriptive analyses of the scoring disparities between raters, with the aim of delineating specific factors that result in inter-rater scoring disagreements. Implications of this exploratory analysis are important for a) enhancing the effectiveness and efficiency of training to improve RTT scoring reliability, and b) developing effective clinical observation skills.
The research questions addressed were:
1. How does inter-rater agreement vary across subtests for participants with aphasia and language-normal adults?
2. How does inter-rater agreement vary across the 15 possible response types for patients with aphasia and language-normal adults?
3. How does inter-rater agreement vary with auditory comprehension scores and aphasia severity?
4. How does inter-rater agreement vary with the number of error types exhibited by individuals with aphasia?

Method

Participants

Ten adults over 20 years of age (ranging from 42 to 72 years, with a mean age of 54.4 years) with aphasia constituted the patient group. The number of years post-onset ranged from one year to 16 years with an average of 6.17 years. All patients were premorbid right-handers. The etiology of aphasia in seven of the 10 patients was a left-sided cerebrovascular accident, one had aphasia secondary to a traumatic brain injury with left-sided focal lesions, and two had aphasia subsequent to brain tumor and follow-up medical intervention The history of neurological etiology, site of lesion and presence of aphasia were confirmed through case histories, reports from neuro-radiological investigations and neurological evaluation, and results on the Western Aphasia Battery (Kertesz, 1982).

Every patient passed a vision screening comprised of observation of eye symmetry, lesions, eye swelling, drainage, and screening tests for visual acuity, visual field and visual attention deficits, central and peripheral visual fields, color vision, and nystagmus. All patients had hearing thresholds of at least 25 dBHL or less for the hearing screening performed at pure tone frequencies of 500Hz, 1000Hz, and 2000Hz. As language-normal individuals tend not to get perfect scores on the RTT (Hallowell, Wertz, & Kruse, 2002), a total of 30 normal adults with no reported history of brain damage, 10 each from the age ranges 41 to 50, 51 to 60, and 61 to 70 (mean age = 52.16 years), were also included as participants.

Procedure

The testing for individuals in the patient group was performed over two sessions. Every subject was tested individually. In the first session, a detailed case history was taken and the Western Aphasia Battery was administered. During the second session, a vision and hearing screening and the RTT were administered. The RTT version used in this test used five items in each of the 10 subtests as opposed to all the 10 items used for subtest 9 in the standardized shortened form of the RTT (Arvedson & McNeil, 1985; Park, McNeil, & Tompkins, 1999). RTT scores range from 1 to 15, with 15 representing a response that is appropriate in terms of all five dimensions. Every element of every command is assigned a score.

Scoring

Two raters scored each individual's RTT performance. Both scorers were graduate students in speech-language pathology and had substantial training to score the RTT prior to scoring the performance of the participants of this study. Training included the use of two RTT training videotapes prepared by Hageman (2001a & 2000b). They had also practiced scoring participants' videotaped performances for which reliable scores had already been determined by experienced scorers. For this study, the first rater (examiner) administered and scored the RTT live, and the second rater (coder) independently scored the videotaped administration of the RTT. Inter-rater agreement was calculated by comparing scores assigned by the examiner and the coder (Park, McNeil, & Tompkins, 1999). For every participant, the total of the number of elements for which same scores were assigned by the two scorers was divided by the total number of elements in the entire test to obtain the scoring agreement for each participant. These values were averaged across participants in the language-normal group and aphasia group separately to obtain the overall percent agreement values for the two groups of participants.

Results

At the subtest level, the percent agreement was least for subtest 6 for both groups of participants (Table 1). Additionally, a linear regression analysis indicated a significant decrease in percent agreement with increase in the number of elements per command, $t(148) = -3.015$, $p = 0.003$ (Figure 1). This relationship was not significant for the aphasia group (Figure 2). Low overall percent agreement for the language-normal group (Mean: 82.75%, SD: 10.25) may have been due to the fact that the types of responses resulting in disagreements (as explained below) have a bearing on the scores of all the elements within the command and thus remarkably reduce the overall percent agreement values.

The scores on which the raters diverged the most represented immediacy and delay for the language-normal group, most of the discrepancies resulting either from a failure to note "change in direction" (McNeil & Prescott, 1978) or mistaking a change in direction for self-correction during video scoring (Figure 3). For the aphasia group, responses representing delay, repetition, error, and perseveration were not noted by one of the scorers, mostly in cases where multiple error responses were observed within a single command (Figure 4). A greater variety of error responses in some participants with aphasia did correspond to lower values of inter-rater agreement. No trends in inter-rater agreement values with an increase in aphasia severity or auditory comprehension deficits were found.

Conclusion

The results provide insights into some factors that contribute to inter-rater disagreements. The extent and nature of disagreements may depend on subtest attributes within the test, patient-

related variables, and the differences between video and live scoring contexts. It is likely that the probability of poor inter-rater agreement could be minimized and accuracy and reliability of scoring enhanced by specifically attending to and emphasizing these factors during training of scorers and RTT administration.

Research Directions and Clinical Implications

Multidimensional scoring aids clinicians in appreciating subtle deficiencies in individuals' performance. Training clinicians to reliably identify these discreet responses is important not only for RTT administration but also for developing good clinical observation skills that help in diagnosis, gauging stimulability and planning treatment for individuals with aphasia. Further empirical studies in this area conducted with a greater number of patient participants and standardized training protocols for scorers could help provide deeper insights into the role of each of the above factors in improving or diminishing scoring reliability.

References

Arvedson, J.C., McNeil, M.R., & West, T. L. (1985). Prediction of Revised Token Test overall, subtest and linguistic unit scores by two shortened versions. Clinical Aphasiology, 15, 57-63.

Hageman, C. (2001a). RTT training tapes. Videotape. #1. Available from School of Hearing, Speech and Language Sciences, Ohio University, Athens, Ohio 45701.

Hageman, C. (2001b). RTT training tapes. Videotape. #2. Available from School of Hearing, Speech and Language Sciences, Ohio University, Athens, Ohio 45701.

Hallowell, B., Wertz, R.T., & Kruse, H. (2002). Using eye movement responses to index comprehension: An adaptation of the Revised Token Test. *Aphasiology, 16,* 587-594.

Kertesz, A. (1982). *Western Aphasia Battery.* USA: The Psychological Corporation.

McNeil, M.R., & Prescott, T.E. (1978). *Revised token test.* Austin, TX: PRO-ED, Inc.

Park, G., McNeil, M.R., & Tompkins, C. Reliability of the five-item revised Token Test for individuals with aphasia. Presented to the Clinical Aphasiology Conference, Key West, FL, June, 1999.

Porch, B.E. (1967). *Porch index of communicative ability*. California: Consulting Psychological Press.

Table 1

Percent Agreement Values by Subtest for Language-Normal and Aphasia Group

Subtest	Percent Agreement Language-Normal Group (N=30)		Percent Agreement Aphasia Group (N=10)	
	Mean	SD	Mean	SD
1	90.44	18.66	88.67	19.12
2	90.83	16.29	92.50	11.36
3	84.67	20.48	82.33	12.86
4	83.58	24.51	74.55	22.21
5	82.67	21.18	82.67	22.26
6	76.21	24.88	74.45	29.81
7	76.22	28.88	80.00	23.77
8	76.00	34.53	89.00	20.62
9	82.50	20.70	85.00	13.54
10	87.20	16.65	87.20	20.81

Figure 1. Percent agreement values with increase in number of elements per command for language-normal group.

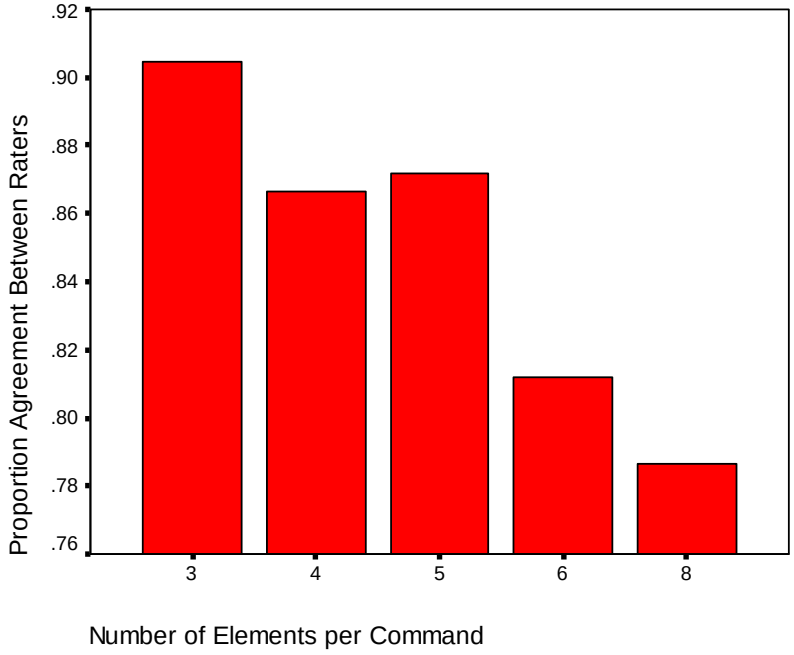

Figure 2. Percent agreement values with increase in number of elements per command for aphasia group.

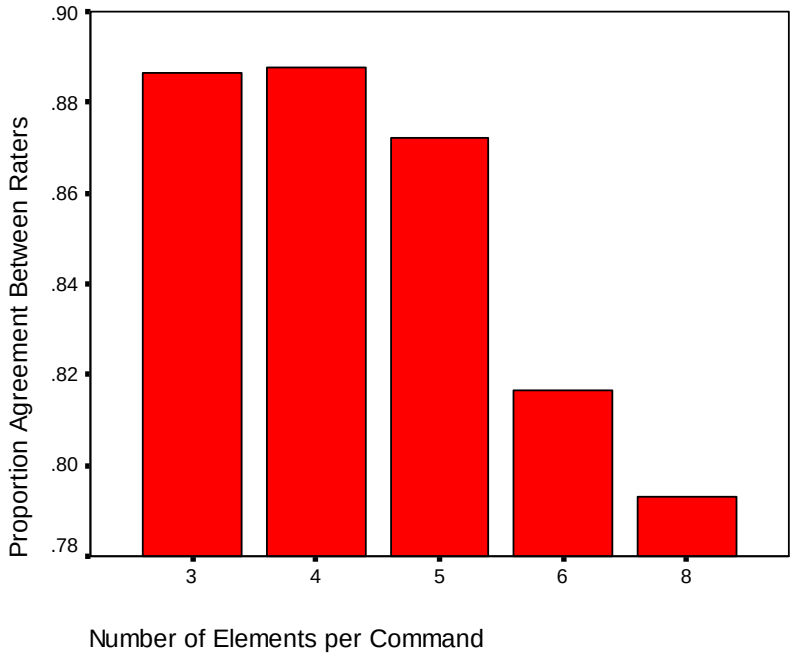

Figure 3. Inter-rater agreement across response types for language-normal group.

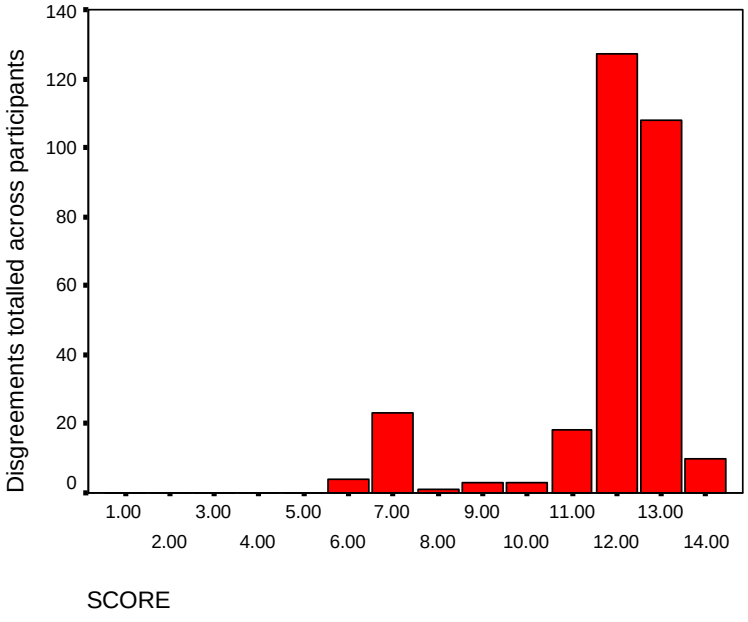

Figure 4. Inter-rater agreement across response types for aphasia group.