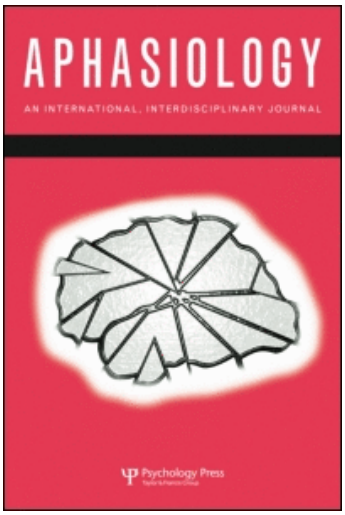

Aphasiology
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713393920

The inter-rater reliability of the story retell procedure

William Hula [a]; Malcolm McNeil [a]; Patrick Doyle [a]; Hillel Rubinsky [b]; Tepanta Fossett [a]
[a] VA Pittsburgh Healthcare System Geriatric Research Education & Clinical Center, and University of Pittsburgh, USA. [b] University of Pittsburgh, USA.

Online Publication Date: 01 May 2003

To cite this Article Hula, William, McNeil, Malcolm, Doyle, Patrick, Rubinsky, Hillel and Fossett, Tepanta(2003)'The inter-rater reliability of the story retell procedure',Aphasiology,17:5,523 — 528
To link to this Article: DOI: 10.1080/02687030344000139
URL: http://dx.doi.org/10.1080/02687030344000139

APHASIOLOGY, 2003, *17* (5), 523–528

The inter-rater reliability of the story retell procedure

William D. Hula, Malcolm R. McNeil, and Patrick J. Doyle

VA Pittsburgh Healthcare System Geriatric Research Education & Clinical Center, and University of Pittsburgh, USA

Hillel J. Rubinsky

University of Pittsburgh, USA

Tepanta R. D. Fossett

VA Pittsburgh Healthcare System Geriatric Research Education & Clinical Center, and University of Pittsburgh, USA

Background: McNeil, Doyle, Fossett, Park, and Goda (2001) have presented the story retell procedure (SRP) as an efficient means of assessing discourse in adults with aphasia, in part because it provides reliable, valid, and sensitive indices of performance without the need for time-consuming transcription of language samples.
Aims: The purpose of this study was to demonstrate that the SRP, when scored without transcription by judges with minimal training, produces a reliable measure of information transfer.
Methods & Procedures: Four judges who had not used the SRP previously scored audio-recorded language samples, produced by four subjects with aphasia and eleven normal subjects, for percent information units per minute (%IU/Min).
Outcomes & Results: The results demonstrate that the SRP has high inter-rater reliability. Reliability coefficients ranged from .89 to .995, and the standard error of measurement associated with inter-rater scoring error ranged from .59 to 1.42 %IU/Min. Point-to-point reliability in scoring individual information units ranged from 85–95% and averaged 91% for both subject groups.
Conclusions: The SRP is a potentially useful tool for quantifying connected language behaviour, and may be particularly valuable in clinical and research settings where economy of assessment procedures is essential.

In a series of recent publications, McNeil, Doyle, and colleagues have presented information on a story retell procedure (SRP) used to elicit language samples from persons with and without aphasia (Doyle et al., 2000; Doyle, McNeil, Spencer, Goda, Cottrell, & Lustig, 1998; McNeil et al., 2001; McNeil, Doyle, Park, Fossett, & Brodsky, 2002). The SRP consists of auditory presentation of stories derived from Brookshire and Nicholas's (1993) Discourse Comprehension Test to a subject or patient, followed by an immediate retell. The stories can be presented with or without picture support, and likewise, picture

Address correspondence to: William D. Hula MS, Doctoral Fellow, Audiology & Speech Pathology, VA Pittsburgh Healthcare System, 7180 Highland Drive, Pittsburgh, PA 15206, USA.
Email: william.hula@med.va.gov

The authors gratefully acknowledge the assistance of Stephanie Nixon and Joyce Poydence. This research was supported by VA Rehabilitation Research and Development Project # C894-2RA.

http://www.tandf.co.uk/journals/pp/02687038.html DOI:10.1080/02687030344000139

support can be provided for the retells, or not, depending on the patient. It has been argued that the SRP possesses some distinct advantages over other connected language sampling procedures described in the literature, including conversational observation (Oelschlaeger & Thorne, 1999), scripted interviews (Goodglass & Kaplan, 1983), on-line video narration (McNeil, Small, Masterson, & Fossett, 1995), fable generation and storytelling (Berndt, Wayland, Rochon, Saffran, & Schwartz, 2000; Ulatowska, Chapman, Highley, & Prince, 1998), picture description (Nicholas & Brookshire, 1993, 1995; Yorkston & Beukelman, 1980), and procedural description (Nicholas & Brookshire, 1993, 1995).

From a language sampling perspective, it has been suggested that the constrained nature of the SRP enables it to provide a well-standardised and replicable sample of language formulation and production. Specifically, data have been presented to support the internal validity of the SRP (Doyle et al., 1998) and the linguistic equivalence of language samples generated by four alternate forms of the procedure (Doyle et al., 2000).

In addition, a scoring metric was developed to quantify the information content and communicative efficiency of the samples generated by the SRP. This metric, labelled the information unit (IU), was derived from Nicholas and Brookshire's (1993, 1995) correct information unit, and was defined as ''an identified word, phrase, or acceptable alternative from the stimulus story that is intelligible and informative and conveys accurate and relevant information about the story'' (McNeil et al., 2001, p. 994). The primary virtue of the IU scoring metric used with the SRP is that all possible IUs are known a priori and can be printed on score sheets. This potentially allows scoring to be done directly from audio recordings, eliminating the need for time-consuming transcription of lengthy language samples. The IU scoring metric expressed as a percentage of total possible IUs (%IU) has been demonstrated to be reliable across forms of the SRP and to have good criterion validity (McNeil et al., 2001). Also, an efficiency measure obtained by dividing %IUs by the time taken to produce them (%IU/Min) has been demonstrated to be reliable across forms and to discriminate between normal and aphasic performance with reasonable accuracy (McNeil et al., 2002).

In addition to reporting on the validity and alternate form reliability of the %IU metric, McNeil and colleagues (2001) demonstrated that it has good inter-observer reliability. However, these data were obtained from scoring of printed transcripts that had themselves already been subjected to reliability checks. Furthermore, all of the data presented thus far on the SRP have been generated by scorers who were themselves involved in the development of the IU measure. If the SRP and its associated IU metrics are to be used to their fullest advantage, particularly in a clinical setting, they must be demonstrated to have acceptable reliability when scored directly from audio recordings by observers who have received training comparable to what a practising clinician could be expected to receive.

One final shortcoming of prior work done to demonstrate the psychometric strength of the SRP concerns the distinction between IUs that were directly stated in the stimulus stories (direct IUs) and IUs retold as synonyms (alternate IUs) of words and phrases contained in the stimulus stories. In a study investigating memory demands of the SRP (Brodsky, McNeil, Park, Fossett, Timm, & Doyle, 2000), a strong serial position effect was demonstrated for direct IUs, but not for alternate IUs. Thus far, no data have been presented to demonstrate that this distinction can be reliably scored.

The purpose of the current paper is to present additional information on the inter-rater reliability of the SRP using procedures and raters more representative of a clinical setting than have been used in the past. Inter-rater reliability coefficients and standard errors of

measurement (SEM) will be reported for the %IU/Min score. Also, point-to-point reliability for identification of individual IU's will be reported.

METHOD

Participants

Recordings of story retells by four persons with aphasia and eleven normal individuals were used. All recordings were randomly drawn from the sample of 15 subjects with aphasia and 31 normal subjects reported by McNeil et al. (2001). Descriptive statistics for the subjects with aphasia are presented in Table 1. Judges were four individuals with varying amounts of experience with aphasia and language transcription: a licensed psychologist, a master's student in speech-language pathology, and two doctoral students who are also certified speech-language pathologists. These judges were a convenience sample, as they were all new employees in the second author's laboratory and required training in scoring the SRP for their work. The two doctoral students both had 2–3 years of work experience that involved transcription of language samples from clinical populations. The psychologist had approximately 13 years of experience with neuropsychological testing of rehabilitation patients, including patients with aphasia, but little experience with language transcription *per se*. The master's student had no experience with language transcription for research or clinical purposes.

TABLE 1
Biographical and descriptive subject information for subjects with aphasia (N = 15)

Subject	*Age*	*MPO*	*RTT Percentile*	*ABCD Ratio*	*Raven's*	*PICA OA Percentile*	*PICA VRB Percentile*
1	62	11	73	84.62	34	92	78
2	77	44	19	118.18	24	59	63
3	47	11	4	100	24	65	54
4	51	77	53	133.33	29	87	60
5	79	13	77	233.33	20	75	77
6*	56	84	95	100	32	87	89
7*	74	71	96	85.7	27	94	97
8	55	30	63	100	32	75	71
9	66	33	80	100	27	89	76
10*	57	85	58	125	27	86	75
11	64	252	14	91	24	93	68
12*	71	94	4	100	22	43	37
13	52	17	92	100	36	87	91
14	73	23	66	116.66	21	76	70
15	74	11	54	100	18	63	54
M	63.87	57.07	56.53	112.52	26.47	78.07	70.67
SD	10.45	62.12	32.10	36.15	5.33	14.90	15.71

Subjects chosen for reliability analysis in this study are marked with an (*).

MPO = Months post onset; RTT = *Revised Token Test* (McNeil & Prescott, 1978), percentile compared to adults with left-hemisphere damage; ABCD ratio = *Arizona Battery for Communication Disorders of Dementia* (Bayles & Tomoeda, 1993) ratio, determined by number of delayed recall items/number of immediate recall items × 10; Raven's = *Raven's Coloured Progressive Matrices* (Raven, 1976), raw score out of a possible 36; PICA = *Porch Index of Communicative Ability* (Porch, 1981), percentile compared to adults with left-hemisphere damage, OA = overall percentile and VRB = verbal percentile.

Procedures

Prior to scoring any of the story retells, each of the four judges read the IU definition and examples published by McNeil et al. (2001), and practised scoring IUs on six to eight stories from printed transcripts. These language samples were drawn from the samples collected by Doyle et al. (2000) and McNeil et al. (2001). After training, each judge scored the same SRP form for each of the four persons with aphasia and eleven normal subjects. Each form consisted of three separate stories as reported by Doyle et al. (2000). All scoring was done from audio files using score sheets containing all possible direct and alternative IUs. Judges listened to each story as many times as they wanted to and placed a check on the score sheet wherever an IU was observed. Wherever an alternate IU (as opposed to a direct IU) was observed, they made an additional mark to denote which of the predetermined synonyms was produced. The %IU/Min for each story was calculated and averaged across the appropriate three-story form to give a total %IU/Min score for each subject. The total %IU/Min score was also broken down into %direct IU/Min and %alternate IU/Min to allow for assessment of inter-rater reliability on these more specific measures.

The judges all reported that it generally took 15–30 minutes for them to score a single form (three stories) of the SRP for a single subject. Data on the time spent scoring retells were kept for the least trained and experienced judge. Her average time to complete a single form was 23 minutes (range = 12–29; SD = 4).

RESULTS

Inter-rater reliability coefficients were calculated separately for subjects with aphasia and normal subjects using the %total, %direct, and %alternate IU/Min scores generated by each of the four judges for each of the subjects. To determine a reliability coefficient that would allow for generalisation to judges beyond those in this study, absolute-agreement intraclass correlation coefficients (ICCs) were calculated with both subjects and judges as random factors. The ICC has been argued to be a more conservative measure of reliability than the Pearson Product Moment Correlation (Denegar & Ball, 1993). The ICCs are presented in Table 2. They ranged from .94 to .995 for the subjects with aphasia and from .89 to .99 for the normal subjects. The SEM associated with inter-judge scoring error was also calculated for each metric. These results are presented in Table 3 and they ranged from .59 to .95 %IU/Min for the subjects with aphasia and from .99 to 1.42 %IU/Min for the normal subjects.

Point-to-point reliability between all six possible pairings of judges was calculated separately for the four subjects with aphasia and for four of the normal subjects. The

TABLE 2
Inter-rater reliability (intraclass) correlation coefficients for total, direct, and alternate %IU/Min

Subjects	*Total*	*Direct*	*Alternate*
Aphasic (n = 4)	0.995	0.986	0.944
Normal (n = 11)	0.993	0.979	0.885

All significant at $p < .001$.

TABLE 3
Inter-rater standard errors of measurement (SEM) for total, direct, and alternate %IU/Min

Subjects	*Total*	*Direct*	*Alternate*
Aphasic (n = 4)	0.69	0.95	0.59
Normal (n = 11)	0.99	1.42	1.04

formula used was [(agreements / disagreements + agreements) × 100]. Point-to-point reliability averaged 91% (range = 85–95%) for both subject groups.

DISCUSSION

The inter-rater reliability for the %IU/Min metric, when scored directly from audio recordings by newly and minimally trained judges, was high, with small differences in scoring reliability among judges with differences in professional experience. The SEMs were found to be much lower than the SEMs reported by McNeil et al. (2002) for the four alternate forms for subjects with aphasia (range = 4.8–5.6) and for the normal subjects (range = 3.2–4.7). The low SEMs suggest that measurement error attributable to differences between raters is small relative to the score variance due to the story forms themselves.

Furthermore, the present data, scored to include the direct vs alternate IU distinction, demonstrated point-to-point reliability that was high and comparable to previously reported values obtained from printed transcripts. Finally, the preliminary data presented regarding the time needed to score language samples elicited by the SRP suggest that it might be useful in clinical environments where economy of assessment procedures is essential.

REFERENCES

Bayles, K. A., & Tomoeda, C. K. (1993). *Arizona Battery for Communication Disorders of Dementia*. Tucson, AZ: Canyonlands Publishing, Inc.

Berndt, R. S., Wayland, S., Rochon, E., Saffran, E., & Schwartz, M. (2000). *Quantitative Production Analysis (QPA).* Philadelphia: Psychology Press.

Brodsky, M., McNeil, M., Park, G., Fossett, T., Timm, N., & Doyle, P. (2000). Auditory memory for story retelling in normal male, female, young, and old adult subjects in persons with aphasia. Poster presented to the Academy of Aphasia Conference, Montreal, CA.

Brookshire, R. H., & Nicholas, L. H. (1993). *Discourse Comprehension Test.* Tuscon, AZ: Communication Skill Builders.

Denegar, C. R., & Ball, D. W. (1993). Assessing the reliability and precision of measurement: An introduction to the intraclass correlation and standard error of measurement. *Journal of Sports Rehabilitation*, *2*, 35–42.

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

Doyle, P. J., McNeil, M. R., Spencer, K. A., Goda, A. J., Cottrell, K., & Lustig, A. P. (1998). The effects of concurrent picture presentations on retelling of orally presented stories by adults with aphasia. *Aphasiology*, *12*, 561–574.

Goodglass, H., & Kaplan, E. (1983). *The assessment of aphasia and related disorders.* Philadelphia: Lea & Febiger.

McNeil, M. R., Doyle, P., Fossett, T., Park, G., & Goda, A. (2001). Reliability and concurrent validity of an information unit scoring metric for the retelling procedure. *Aphasiology*, *15*, 991–1007.

McNeil, M. R., Doyle, P., Park, G., Fossett, T., & Brodsky, M. (2002). Increasing the sensitivity of the Story Retell Procedure for the discrimination of normal elderly subjects from persons with aphasia. *Aphasiology*, *16*, 815–822.

McNeil, M. R., & Prescott, T. E. (1978). *The Revised Token Test*. Austin, TX: Pro-Ed.

McNeil, M. R., Small, S. L., Masterson, R. J., & Fossett, T. R. D. (1995). Behavioral and pharmacological treatment of lexical-semantic deficits in a single patient with primary progressive aphasia. *American Journal of Speech-Language Pathology*, *4*, 76–87.

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.

Oelschlager, M. L., & Thorne, J. C. (1999). Application of the correct information unit analysis to the naturally occurring conversation of a person with aphasia. *Journal of Speech, Language, and Hearing Research*, *42*, 636–648.

Porch, B. E. (1981). *Porch Index of Communicative Ability*. Palo Alto, CA: Consulting Psychologists Press.

Raven, J. C. (1976). *Coloured Progressive Matrices*. Oxford: Oxford Psychologists Press, Ltd.

Ulatowska, H. K., Chapman, S. B., Highley, A. P., & Prince, J. (1998). Discourse in healthy old-elderly adults: A longitudinal study. *Aphasiology*, *15*, 619–633.

Yorkston, K. M., & Beukelman, D. R. (1980). An analysis of connected speech samples of aphasic and normal speakers. *Journal of Speech and Hearing Disorders*, *45*, 27–36.