BACKGROUND

The multidimensional scoring system first introduced in the Porch Index of Communicative Ability (1981, 2001), or PICA, has been considered a conceptual breakthrough in the assessment of aphasia. The multidimensional scoring system is based on the assumption that each response on any test item involves a delicate brain circuitry whose behavior is only understandable and thus inferable through a fine-grained analysis of that response (Porch, 2008). The rather coarse-grained nature of the binary system can be imprecise or even misleading. The advantage of the multidimensional scoring lies in its better ability to make prognosis of a patient's recovery (Porch & Callaghan, 1981; Porch et al., 1981). On the other hand, the simultaneous inspection of the five dimensions for a given response can be daunting, even for a trained clinician (Odekar & Hallowell, 2005). Further, it has always been a challenging issue that the inter-rater reliability of any test employing such a complicated scoring method may be susceptible. For instance, Odekar and Hallowell (2005) found that the inter-rater agreement for the aphasic performance on the Revised Token Test varied from .74 to .93 across the ten subtests. The purpose of this presentation is to report the effect of scoring training on the rating performance by speech clinicians on the Concise Chinese Aphasia Test (or CCAT; Chung, Lee, & Chang, 2002), which is similar in format and structure to PICA and is widely used in Taiwan for aphasia assessment. Like PICA, the CCAT was designed to elicit verbal, graphic, and gestural responses from an aphasic patient with the aid of ten everyday objects. It consists of nine subtests and each is comprised of ten items. The test adopts a multidimensional scoring system based on accuracy, responsiveness, completeness, promptness, and efficiency with a maximum score of 12, instead of 16 adopted in PICA.

METHOD

Nine speech clinicians and three graduate students volunteered to participate as trainees in the study. In a six-hour group session, the principles and procedures of test administration and scoring were explained and demonstrated to the participants. Then, each participant watched a training video and practiced to score at home. Responses on a total of one hundred items elicited from eight aphasic patients with mild, moderate and

severe levels of severity were viewed. The trainees then gave score for each item at this point (i.e., Time 1). Each participant's scoring performance was judged against the "gold standard", that is, the scores given for those items by one of the test developers. The next stage of training was optional, only applied to those who failed to pass. If a participant's scoring accuracy (as compared to the gold standard) on a subtest was lower than 80%, she then received the follow-up training on that subtest. At the end of the training program (i.e., Time 2), the scoring performance by the trainees was evaluated again. The participants watched test responses by ten aphasic patients on various test items from videotapes. A total of 900 items were rated by each participant. The whole training program lasted roughly for about eight months.

RESULTS

The results of the training program are summarized as follows. First, as shown in Table 1, the overall rating error (i.e., the score difference between a trainee's and the gold standard) was smaller than one point on each subtest at both Time 1 and Time 2. However, the rating error did not appear to decline over the training course except on one subtest. Second, the standard deviation of rating error, which reflects the inter-rater variation on the same test item, appeared to decrease from a mean of .35 to that of .16. This seems to mean that the twelve trainees, though unable to perfectly match the gold standard, were able to reach consensus among themselves. Second, as shown in Table 2, the rating accuracy (i.e., the percentage of items which were identically rated by a trainee and by the gold standard) did not appear to increase significantly from Time 1 to Time 2 except for two subtests. Again, the inter-rater variability decreased over the period. Finally, a coarse-grained analysis of the rater's performance was done so that accuracy was now redefined according to a loose within-the-ballpark standard; that is, the rating was judged to be adequate if it fell within any of the three categories (accurate and prompt, accurate but repeated, or anything below). Not surprisingly, as indicated in Table 3, the coarse-grained accuracy was considerably high at 94% at Time 1 and it remained there even at Time 2. As a matter of fact, it even dropped substantially on three subtests. Again, the inter-rater variability decreased from .05 to .03. In short, the scoring

difference between the trainees and the gold standard did not diminish as a result of training although the trainees were becoming more and more like one another at the end of the program.

DISCUSSION

The participants reached a fair performance level after a six-hour training course which was basically a lecture plus scoring demonstrations. This finding indicates that clear instructions on the scoring principles and some amount of demonstrations are adequate enough to achieve a fair outcome. Further training in the format of videotaped demonstrations and home practice did not diminish the gap between the trainees and the gold standard. As the gold standard had been provided by a senior test developer, who has had at least twenty-year experience in multidimensional scoring, the expert knowledge involved might be so subtle, so implicit, and so subconsciously embedded that it can only be acquired from long-term practice. However, the good news is that a proper training did reach a very satisfactory level of inter-rater agreement.

REFERENCES

Chung, Y. M., Lee, S. E., & Chang, M. H. (2001). The concise Chinese aphasia test.
Taipei: Psychological Corporation.
Odekar, A., & Hallowell, B. (2005). Comparison of alternatives to multidimensional
scoring in the assessment of language comprehension in aphasia. American
Speech-Language Pathology, 14, 337-345.
Porch, B. E. (1981). The Proch index of communicative ability. Palo Alto, CA:
Consulting Psychologists Press.
Porch, B. E. (1981). The Proch index of communicative ability. Albuquerque: Pica
Programs.
Porch, B. E. (2008). Treatment of aphasia subsequent to the Proch index of
communicative ability (PICA). In R. Chapey (Ed.), Language intervention
strategies in aphasia and related neurogenic communication

disorders (pp. 800
813). Philadelphia: Woters Kluwer.
Porch, B. E. , & Callaghan, S. (1981). Making predictions about recovery: Is there
HOAP? In R. H. Brookshire (Ed.), Clinical Aphasiology Conference proceedings. Minneappolis, MN: BRK.

Table 1
Rating Errors (Mean and SD) at Time1 and Time 2 for Each Subtest of CCAT

Rating error Time 1	Mean	SD	Time 2 Mean	SD	t statistic
Conversation	.58	.27	.50	.14	.91
Picture description	.63	.28	1.09	.17	-4.87*
Word-picture matching	.54	.46	.43	.09	.81
Auditory comprehension	.38	.27	.39	.13	-.12
Confrontation naming	.63	.46	.56	.20	.48
Reading comprehension	.46	.34	.22	.07	2.40*
Repetition	.57	.41	.47	.17	.78
Copying	.73	.40	.85	.29	-.84
Spontaneous writing	.57	.29	.60	.20	-.30
Group mean	.57	.35	.57	.16	

**p < .05

Table 2
Rating Accuracy (Mean and SD) at Time1 and Time 2 for Each Subtest of CCAT

Rating accuracy Time 1	Mean	SD	Time 2 Mean	SD	t statistic
Conversation	.69	.11	.67	.09	-.46
Picture description	.60	.18	.53	.06	-1.28
Word-picture					

matching .65 .15 .81 .04 3.54**
Auditory comprehension .77 .14 .84 .05 1.60
Confrontation naming .74 .15 .70 .09 -.70
Reading comprehension .65 .18 .84 .05 3.44**
Repetition .64 .18 .72 .16 1.25
Copying .55 .18 .55 .15 .12
Spontaneous writing .58 .17 .68 .07 1.93
Group mean .65 .16 .70 .08
* * p < .01

Table 3
Coarse-Grained Rating Accuracy (Mean and SD) at Time1 and Time 2 for Each Subtest of CCAT
Coarse-grained accuracy Time 1
Mean
SD Time 2
Mean
SD t statistic
Conversation .96 .05 .99 .01 1.31
Picture
description .91 .04 .82 .06 -5.81**
Word-picture
matching .95 .09 .99 .01 1.22
Auditory comprehension 1.00 .00 .98 .01 -5.50**
Confrontation naming .94 .08 .95 .02 .62
Reading comprehension 1.00 .00 .99 .01 -3.19**
Repetition .92 .08 .91 .04 -.75
Copying .95 .07 .95 .05 -.18
Spontaneous writing .88 .08 .91 .04 .86
Group mean .94 .05 .94 .03
* * p < .01