Clinical examination of reliability/validity of scoring methods for Cube-Copying Test (CCT)

Shino Mori, MD,¹ Aiko Osawa, MD, PhD,¹,² Shinichiro Maeshima, MD, PhD,³
Kenichi Ozaki, MD, PhD,¹ Takashi Sakurai, MD, PhD,² Izumi Kondo, MD, PhD,¹
Eiichi Saito, MD, PhD⁴

¹Department of Rehabilitation Medicine, National Center for Geriatrics and Gerontology, Aichi, Japan
²Medical Center for Dementia, National Center for Geriatrics and Gerontology, Aichi, Japan
³Department of Rehabilitation Medicine II, School of Medicine, Fujita Health University, Aichi, Japan
⁴Department of Rehabilitation Medicine I, School of Medicine, Fujita Health University, Aichi, Japan

Introduction

The number of dementia patients is steadily increasing along with the growing population of older adults, which makes it particularly important to take measures to address the disease. It has been found that patients in the early stages of dementia are likely to develop not only memory impairment but also impaired spatial cognitive function – particularly visuo-spatial function – and constructional inability, as well as abnormal motor programming; therefore, evaluation of these functions is imperative in making a diagnosis of dementia [1–3]. Particularly in Alzheimer’s Disease (AD), which in the early stages is characterized by decreased blood flow in the parietal lobe region and the medial temporal lobe, impaired visuo-spatial cognitive function and constructional inability are often manifested in the early phase of the illness [3–7].

In the Cube-Copying Test (CCT), the examinee is asked to copy a perspective drawing of a cube, which is then evaluated. The test allows nonverbal assessment of visuo-spatial function and constructional ability and is widely used in ordinary clinical settings. Other methods for evaluating visuo-spatial function include the Wechsler Adult Intelligence Scale (WAIS) performance IQ test [8], Raven’s Colored Progressive Matrices (RCPM) [9], and Kohs Block Design Test [10]. However, these tests are difficult to conduct during an outpatient visit due to their relatively long administration time. In this respect, the CCT offers high clinical utility due to its advantage of fast application time. Nevertheless, a definitive method for evaluation or interpretation has not yet been established for the CCT; currently, scoring systems that can quantify the accuracy of figure reproduction and the
presence of a trend in reproduction error types by disease remain under examination and debate [11–13]. To date, a number of scoring systems have been reported including those developed by Kato et al. [12], Maeshima et al. [3, 13, 14], Shimada et al. [6], Takeda et al. [15], Otomo [16], and Yorimitsu et al. [17]. Although the validity of the CCT using each of these scoring methods has been examined and reported [14, 17], few studies have been conducted to assess both reliability and validity, and there have been no studies that assessed the reliability and validity of multiple scoring methods at the same time. In the present study, we examined the reliability of the CCT using two different methods of scoring by two raters multiple times. Additionally, the validity of the CCT was examined through a comparison with patient background and results of other neuropsychological tests.

**Subjects and Methods**

1. **Subjects**

The study included 33 patients (11 male and 22 female) who visited the Medical Center for Dementia at our hospital. Mean age was 76.5±8.3 (50–89 years), mean years of education was 10.2±2.5 (6–18 years), and mean duration since forgetfulness was first noted (duration of illness) was 27.8±23.0 months (5–84 months). Clinical diagnosis was as follows: 26 AD patients, 2 Vascular Dementia (VaD) patients, 3 combination of AD and VaD (AD+VaD) patients, and 2 normal patients. AD and VaD were diagnosed in accordance with the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) [18] (American Psychiatric Association 1994).

2. **Methods**

The results of the CCTs and the following neuropsychological tests administered at the memory clinic at our hospital for diagnosis of dementia were retrospectively collected from electronic medical records and the data were analyzed.

2.1. **CCT**

A perspective drawing of a cube printed on a sheet of paper was visually presented to the examinee, who was verbally asked to reproduce the figure on a sheet of blank A4 paper. No time limit was set for the task. The various scoring systems developed for the CCT include a qualitative method that measures the examinee’s ability to provide an accurate reproduction; a method that grades the examinee’s way of joining vertices and constructing lines; and a method that evaluates the resultant figure by pattern classification. The present study used the method by Maeshima et al. (hereinafter the M method) [3, 13, 14], in which assessment is made by scoring both the vertices and lines, and the method by Shimada et al. (hereinafter the S method) [6], in which the assessment is based on pattern classification. The scoring system of each evaluation method is described separately. Scoring was independently performed by two specialists who are members of the Japan Society for Dementia Research and who were blinded to the patients’ profiles (including diagnosis, symptoms, and severity), using both CCT scoring methods. Additionally, each scoring method was administered twice at an interval of one week to examine intra-rater reliability.

2.2. **Neuropsychological Tests**

The results of Mini-Mental State Examination (MMSE) [19], Frontal Assessment Battery (FAB) [20], and Raven’s Colored Progressive Matrices (RCPM) [9] were collected from medical records as the data for neuropsychological tests. The reason for adopting these tests was to examine the relationship between the CCT and tests that are considered to reflect constructional and visuospatial abilities, language comprehension and intellectual functioning, frontal lobe functioning, and memory – namely, to examine criterion-related validity of the CCT. MMSE is a screening test that measures cognitive function by assessing orientation, naming, attention and calculation, recall, and language, and the maximum score for assessment is 30 points. FAB consists of six items that are indicators of frontal lobe functioning, i.e., similarities, literal fluency, motor series, conflicting instructions, go/no-go task, and prehension behavior. Three points are allocated to each item, resulting in a maximum score of 18 points. In RCPM for assessing visual cognition, the examinee is asked to determine the element that fits the missing section of a large pattern from among six alternative patterns. No time limit is set and the number of correctly chosen patterns in 36 items is counted as the total points.

2.3. **Statistical Analysis**

All analyses were performed using SPSS Ver. 21.0.0.0 for Windows. Reliability of the CCT scoring methods was examined using Cronbach’s reliability coefficient (α) and Intraclass Correlation Coefficient (ICC). Models ICC (2,1) and (2,k) were used for inter-rater reliability and ICC (1,1) and (1,k) were used for intra-rater reliability [21]. Additionally, to examine criterion-related validity, the correlation between each CCT scoring method and the neuropsychological tests (MMSE, FAB, and RCPM), age, education yaers, and duration of illness was examined using Spearman’s rank correlation coefficient.

**Results**

Patient background and results of neuropsychological tests by clinical diagnosis are presented in Table 1.
Table 1. Patient background and neuropsychological tests by clinical diagnosis.

<table>
<thead>
<tr>
<th>Clinical Diagnosis</th>
<th>Total Cases</th>
<th>AD (n=26)</th>
<th>AD+VaD (n=3)</th>
<th>VaD (n=2)</th>
<th>Others (n=2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of patients (male/female)</td>
<td>33 (11/22)</td>
<td>26 (9/17)</td>
<td>3 (1/2)</td>
<td>2 (0/2)</td>
<td>2 (1/1)</td>
</tr>
<tr>
<td>Age (years)</td>
<td>76.5±8.3</td>
<td>77.0±7.4</td>
<td>81.0±5.6</td>
<td>76.5±3.5</td>
<td>63.0±18.4</td>
</tr>
<tr>
<td>Educational level (years)</td>
<td>10.2±2.5</td>
<td>10.0±2.2</td>
<td>10.0±2.0</td>
<td>11.0±2.8</td>
<td>13.5±6.4</td>
</tr>
<tr>
<td>Duration of illness (months)</td>
<td>27.8±23.0</td>
<td>30.9±23.9</td>
<td>14.0±9.2</td>
<td>8.5±3.5</td>
<td>27.0±29.7</td>
</tr>
<tr>
<td>MMSE (/30)</td>
<td>18.5±4.4</td>
<td>18.7±4.1</td>
<td>14.0±6.0</td>
<td>18.5±2.1</td>
<td>23.0±2.8</td>
</tr>
<tr>
<td>FAB (/18)</td>
<td>8.5±3.0</td>
<td>8.3±2.8</td>
<td>10.0±2.7</td>
<td>5.5±0.7</td>
<td>12.0±5.7</td>
</tr>
<tr>
<td>RCPM (/36)</td>
<td>21.9±6.4</td>
<td>21.6±6.4</td>
<td>26.5±3.5</td>
<td>16.0±2.8</td>
<td>27.0±8.5</td>
</tr>
</tbody>
</table>

AD, Alzheimer’s Disease. The values are mean±standard deviation.
VaD, Vascular Dementia.
MMSE, Mini-Mental State Examination.
FAB, Frontal Assessment Battery.
RCPM, Raven’s Colored Progressive Matrices.

Notes

Maeshima, et al. (1997)  
Number of Vertices: (Score 0–8) Number of joining points of 3 sides (vertical, horizontal, and diagonal). One point is awarded for every vertex where 3 sides are joined; since a correct cube has 8 vertices, it scores 8 points.
Number of Axial Errors: (Score 0–6) In each quartet of (vertical, horizontal, and diagonal) parallel lines, the presence of any side that is not parallel with the axis of each group of lines, omission of sides, inclusion of extra sides, and so forth are defined as axial errors. If there is an error or omission of an item, it scores 1 point. A correct cube scores 0 point.

Shimada, et al. (2006)  
* The items described in ( ) are optional.
Pattern 0: Lines only. Absence of rectangle.
Pattern 1: One quadrilateral (plus).
Pattern 2: Two or more quadrilaterals (plus lines) However, the drawing couldn’t be judged to be a three-dimensional (3D) figure.
Pattern 3: 3D but not a cube: Participants succeeded in constructing a 3D figure, but failed to make it a cube.
Pattern 4: Cube (plus lines): Participants succeeded in drawing a cube, but fell short of the Necker cube.
Pattern 5: Distorted model: Although the figure consisted of 12 or more lines and could be judged to be the Necker cube, the relationship between these line segments was different from that of the model, based on at least one of the following criteria: (i) each side of the figure could not be judged to be a quadrilateral, or the figure had more than six sides; and (ii) the two overlapping squares of the Necker cube were transposed from the left-lower-right upper pattern to the left-upper–right-lower pattern, or the two squares did not overlap each other.
Pattern 6: Almost the same as the model: Participants were able to copy a figure almost correctly, only some angles were incorrect.
Pattern 7: A perfect cube.
1. Reliability of the CCT

Table 2 shows the mean values and standard deviations of the scoring results and the ranges of scores given by rater (i) and rater (ii) in the first and second testing sessions. Regarding inter-rater reliability, as demonstrated in Table 3, Cronbach’s reliability coefficient (α) was higher than 0.9 for the number of vertices and axial errors in the M method and the S method (Number of vertices, 1st session α = 0.997 / 2nd session α = 0.934; Number of axial errors, 1st session α = 0.973 / 2nd session α = 0.936; and S method, 1st session α = 0.958 / 2nd session α = 0.902). In all cases, the ICC was higher than 0.81, thereby demonstrating extremely high (almost perfect) reliability. Also regarding intra-rater reliability, every Cronbach’s reliability coefficient (α) was over 0.9 for the number of vertices and axial errors and the S method (Number of vertices, rater (i) α = 0.993 / rater (ii) α = 0.942; Number of axial errors, rater (i) α = 0.990 / rater (ii) α = 0.956; and S method, rater (i) α = 0.977 / rater (ii) α = 0.925) and every ICC was also higher than 0.81.

2. Validity of the CCT

Since the reliability of the CCT was demonstrated by result 1, we then examined its validity using the resultant scores given by the rater (i) in the first testing sessions as representative values. The analysis results of the correlation between representative values and basic patient information (Table 4) showed a significant correlation with education years in all scoring methods for the number of vertices and axial errors of the M method and the S method (M method, r = 0.4521 for number of vertices and r = −0.4408 for number of axial errors; and S method, r = 0.4589), whereas no significant difference was exhibited for age or duration of illness. Additionally, the analysis results of the correlation between representative values and neuropsychological tests (Table 5) demonstrated a significant correlation with RCPM in both scoring methods, i.e., the number of vertices and axial errors of the M method and the S method (M method,

---

**Table 2.** Mean values, standard deviations, and ranges of scores in each evaluation method by raters/testing sessions.

<table>
<thead>
<tr>
<th>Evaluation Method</th>
<th>Rater (i)</th>
<th>Rater (ii)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1st Session</strong></td>
<td><strong>2nd Session</strong></td>
<td><strong>1st Session</strong></td>
</tr>
<tr>
<td>Maeshima Method (Number of Vertices)</td>
<td>3.6±2.5 (0.8)</td>
<td>3.6±2.5 (0.8)</td>
</tr>
<tr>
<td>Maeshima Method (Number of Axial Errors)</td>
<td>3.6±2.6 (0.9)</td>
<td>3.7±2.6 (0.9)</td>
</tr>
<tr>
<td>Shimada Method</td>
<td>4.2±1.9 (1.7)</td>
<td>4.5±2.0 (1.7)</td>
</tr>
</tbody>
</table>

Mean±Standard Deviation (SD) (Ranges of Scores)

**Table 3.** Inter-rater reliability and intra-rater reliability of each evaluation method.

<table>
<thead>
<tr>
<th>Inter-rater Reliability (Rater (i) vs. Rater (ii))</th>
<th>1st Session</th>
<th>2nd Session</th>
<th>1st Session</th>
<th>2nd Session</th>
<th>1st Session</th>
<th>2nd Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cronbach’s Reliability Coefficient (α)</td>
<td>0.997*</td>
<td>0.934*</td>
<td>0.973*</td>
<td>0.936*</td>
<td>0.958*</td>
<td>0.902*</td>
</tr>
<tr>
<td>ICC (2,1) (ρ)</td>
<td>0.993*</td>
<td>0.957*</td>
<td>0.943*</td>
<td>0.877*</td>
<td>0.921*</td>
<td>0.824*</td>
</tr>
<tr>
<td>ICC (2,k) (ρ)</td>
<td>0.996*</td>
<td>0.923*</td>
<td>0.971*</td>
<td>0.935*</td>
<td>0.959*</td>
<td>0.903*</td>
</tr>
</tbody>
</table>

Intra-rater Reliability (1st session vs. 2nd session)

<table>
<thead>
<tr>
<th>Intra-rater Reliability (1st session vs. 2nd session)</th>
<th>Rater (i)</th>
<th>Rater (ii)</th>
<th>Rater (i)</th>
<th>Rater (ii)</th>
<th>Rater (i)</th>
<th>Rater (ii)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cronbach’s Reliability Coefficient (α)</td>
<td>0.993*</td>
<td>0.942*</td>
<td>0.990*</td>
<td>0.956*</td>
<td>0.977*</td>
<td>0.925*</td>
</tr>
<tr>
<td>ICC (1,1) (ρ)</td>
<td>0.985*</td>
<td>0.873*</td>
<td>0.980*</td>
<td>0.918*</td>
<td>0.942*</td>
<td>0.856*</td>
</tr>
<tr>
<td>ICC (1,k) (ρ)</td>
<td>0.993*</td>
<td>0.932*</td>
<td>0.990*</td>
<td>0.957*</td>
<td>0.970*</td>
<td>0.922*</td>
</tr>
</tbody>
</table>

ICCs, Intra-class Correlation Coefficient.

*p (two-sided) <0.001
ρ = 0.7018 for number of vertices and ρ = −0.6594 for number of axial errors; and S method, ρ = 0.5248). Furthermore, a significant correlation was also found with FAB in the number of vertices and axial errors of the M method (ρ = 0.4467 for number of vertices and ρ = −0.4300 for number of axial errors). Conversely, there was no significant correlation with total MMSE score in either the scoring methods or outcome measures (M method, ρ = 0.2366 for number of vertices and ρ = −0.1727 for number of axial errors; and S method, ρ = −0.1767). Additionally, the S method showed no significant difference for FAB as well (ρ = 0.2715) and thus the only neuropsychological test that showed a significant correlation with the S method was RCPM.

### Discussion

#### 1. Reliability and Validity of the CCT

Both the M method and the S method with the two raters showed extremely high intra-rater reliability as well as very high inter-rater reliability, thereby demonstrating the reliability of these two scoring methods.

Additionally, both scoring methods showed a significant correlation with RCPM. As mentioned earlier, previous studies have reported that the CCT reflects visual cognitive functioning at a high rate and the present study has also shown a significant correlation between the CCT and RCPM. Thus, it was demonstrated that both the M method and the S method are likely to be valid scoring methods when the basic scale for assessment is visual cognitive functioning.

On the other hand, in the present study, the M method exhibited no significant correlation with MMSE although it showed a significant correlation with RCPM and FAB. This is likely related to the fact that MMSE includes less visual cognitive tasks in the test items. Additionally, the FAB results suggested a relationship between executive functions proposed by Lezak et al. [22] and constructional ability. That is, copying of a cube requires strategic procedures, which may make the task difficult for an examinee suffering from executive dysfunction. It has also been reported that accurate reproduction of a perspective drawing of a cube is affected by verbal IQ in addition to executive functions [23], and an association with language functioning and visual information processing ability
has also been reported [24]. In the present study, MMSE and FAB include verbal tasks in the test items. However, they are extremely easy screening tasks as measures of assessing verbal functioning, and thus further examination is necessary to discuss the relationship between the perspective drawing of a cube and language dysfunction in dementia patients. Additionally, with respect to the relationship with visual information processing, both the M method and the S method showed a significant correlation with RCPM, thereby suggesting a correlation between the perspective drawing of a cube and not only visual cognitive functioning but also performance IQ.

In addition, both scoring systems of the M method and the S method presented a significant correlation with education years, which is in agreement with previous reports on the relationship between constructional inability and years of education [6, 25]. Furthermore, it has also been reported that accurate reproduction of a perspective drawing of a cube requires six or more years of education [6]. The distribution of education level in the present study was 6–18 years, indicating that all patients had six or more years of education. This would suggest that in the present study, the short period of education was unlikely to be a confounding factor making the cube-copying task difficult for the examinee. Additionally, no significant correlation was found between the M and S scoring methods and age or duration of illness. This result suggests that age and duration of illness have less effect on cube reproduction compared with years of education. On the other hand, however, since AD is a progressive disease, it can clearly be predicted that constructional inability will be aggravated as the duration of illness becomes longer. When considering the factors behind the absence of significant correlation between both scoring methods and duration of illness, it is possible that the length of illness used in the present study is inaccurate. Determination was based on subjective observation by the examinee’s family that was written on the inquiry sheet at the time of the first consultation, or based on the information collected retrospectively from the medical records produced by the physician in charge. The effects of these information biases need to be examined in future studies.

2. Role of the CCT in Evaluation of Spatial Cognition and Constructional Inability

Constructional inability was first reported by Kleist in 1914 as a local cerebral symptom for which the responsible lesion is the parietal lobe of the dominant hemisphere. Since then, its pathogenic mechanism has been a constant topic of discussion. It is now considered that constructional inability develops as a result of an injury to either the right or left hemisphere, occurring mostly in the posterior foci while there has also been a report of the disorder caused by an injury to the frontal lobe [22]. Furthermore, it has been reported that constructional inability occurs at a high rate due not only to disorder of lesions in the parietal-occipital lobes of the right and left cerebral hemispheres but also to diffuse lesions including cerebral atrophy and cerebral ventriculomegaly [12], and that constructional inability is observed in approximately half the patients with Parkinson’s disease [13], while some researchers have reported that constructional inability can occur at a high rate with autoimmune diseases including Systemic Lupus Erythematosus (SLE) as part of the neurologic symptoms [26]; thus, constructional inability may be caused by a diverse range of pathologies.

As suggested by many previous studies and the present study, the CCT can measure visual cognitive functioning and constructional ability with validity, with the added advantage of being fast and easy to administer. In addition, the CCT can assess both aspects of cognition and behavior and may also reflect abnormalities in motor programming, thus making it a highly useful test. In the future, to optimize the utility of the CCT, effective combinations with other neuropsychological tests need to be explored and the pathological diagnostic significance should be examined.

Acknowledgments

We thank President Kenji Toba for the opportunity to write this paper.

References