A Comparison of Alternative Approaches to the Scoring of Clock Drawing

Holly Tuokko
University of Victoria

Thomas Hadjistavropoulos
University of Regina

Sandra Rae
Simon Fraser University

Norm O’Rourke
University of Ottawa

Although a number of scoring procedures for clock drawing have emerged in the literature, no systematic comparison of the psychometric properties of various approaches has yet been conducted on a large sample of persons over 64 years of age diagnosed with dementia or deemed cognitively intact. The present study examined the reliability and validity of five scoring approaches (Doyon, Bouchard, Morin, Bourgeois, & Cote, 1991; Shulman, Shedletsky, & Silver, 1986; Tuokko, Hadjistavropoulos, Miller, & Beattie, 1992; Watson, Arfken, & Birge, 1993; Wolf-Klein, Silverstone, Levy, Brod, & Breuer, 1989) among the 493 participants of the Canadian Study of Health and Aging who completed clock drawing and who had a final diagnosis...
The clock drawing task has been used traditionally in the assessment of constructional apraxia and has been thought to reflect parietal lobe function. Spreen and Strauss (1991) note that the procedure has been part of the brief mental status examination in neurology for many years (e.g., Battersby, Bender, Pollack, & Kahn, 1956; Critchley, 1953). Recently, interest in this procedure as a screening tool for the identification of dementia, and more specifically, Alzheimer’s disease, has emerged (e.g., Shulman, Shedletsky, & Silver, 1986; Tuokko, Hadjistavropoulos, Miller, & Beattie, 1992; Wolf-Klein, Silverstone, Levy, Brod, & Breuer, 1989). It appears that this quick and easily administered procedure may be sensitive to the early identification of dementia and prove to be a useful tool for screening large numbers of elderly persons for early signs of neurodegenerative dysfunction (Cahn, et al., 1996; Ferrucci et al., 1996; O’Rourke, Tuokko, Hayden, & Beattie, 1997). As compared to more traditional procedures (e.g., Teng & Chui, 1987), it has been suggested that clock drawing is less susceptible to confounds resulting from ethnicity, premorbid intelligence, and educational attainment (Shulman, Pushkar Gold, Cohen, & Zucchero, 1993).

Several different methods for scoring this task (see Tuokko & Hadjistavropoulos, 1998 for a brief review) have been developed in recent years (Goodglass & Kaplan, 1983; Kozora & Cullum, 1994; Shulman et al., 1986; Spreen & Strauss, 1998; Sunderland et al., 1989; Wolf-Klein et al., 1989). These scoring systems often involve the rating of clock drawings from best to worst (Sunderland et al., 1989) or the assignment of clock drawing performance to discrete error categories (Wolf-Klein et al., 1989). Watson, Arfken, and Birge (1993) award points for errors in number placement. A dual scoring system was developed by Tuokko, Hadjistavropoulos, Miller, Horton, and Beattie (1995), which assesses general level of efficiency of performance (i.e., quantitative appraisal) as well as the specific types of errors made (e.g., omissions, perseverations, rotations, misplacements, distortions, substitutions, additions).

In examining different approaches to clock drawing, Kozora and Cullum (1994) compared a 3-point scoring system (Goodglass & Kaplan, 1983), a 10-point scoring system (Spreen & Strauss, 1991) and a 16-point scoring system of their own for differentiating among healthy adults of different ages and between healthy adults (N = 49) and persons with Alzheimer’s disease (N = 49). All three rating systems were roughly equivalent in differentiating younger and older healthy participants and in discriminating between Alzheimer’s disease patients and controls. Kozora and Cullum (1994) concluded that the clock drawing task shows clinical utility and merits further qualitative and quantitative investigation.

To investigate further the clinical utility of clock drawing, we employed a sample of 493 elders and compared five scoring procedures commonly referenced (Spreen & Strauss, 1998) in the literature (i.e., Doyon, Bouchard, Morin, Bourgeois, & Cote, 1991; Shulman et al., 1986; Tuokko et al., 1992; Watson et al., 1993; Wolf-Klein et al., 1989) using a variety of comparison methods (e.g., inter-rater reliability, intra-rater reliability, sensitivity,
Scoring of Clock Drawing

specificity, receiver operating curves). These scoring procedures are reviewed by Spreen and Strauss (1998) in their Compendium of Neuropsychological Tests. The Tuokko et al. method was reviewed very favorably, largely because of the clarity of scoring and the “excellent standardization” data. However, this and several other widely cited methods (e.g., Shulman et al., 1986; Wolf-Klein et al., 1989) have never been systematically contrasted to one another on a common sample nor are we aware of any other relevant (i.e., related to clock drawing) psychometric comparison that utilized such a large sample. Hereafter, these procedures (to be described below) will be referred to as: Doyon, Shulman, Tuokko, Watson, and Wolf-Klein, respectively. These scoring procedures were chosen because they each have explicitly stated scoring criteria and the administration involves the use of a predrawn circle. The use of a predrawn circle focuses the clock drawing performance on number and hand placement, thereby circumventing some difficulties inherent in procedures in which the participant draws the circle as well. A poorly drawn circle confounds the remainder of the clock-drawing performance.

The Shulman, Tuokko, Watson, and Wolf-Klein scoring procedures were all applied to a single clock-drawing performance. The participant was presented a predrawn circle and asked to complete the clock face and place the hands on the clock to 10 past 11. The Doyon clock drawing task differed in that the predrawn circle contained the number 12 in the appropriate location and a dot in the centre of the circle.

Although many practitioners can identify grossly distorted clock drawing without using explicit scoring criteria, it is the subtleties of disturbance that may be most germane in terms of diagnostic sensitivity and specificity. The methods selected for comparison differ greatly in the complexity of the scoring procedure and this may be reflected in their relative stability and clinical utility. The analyses address issues specific to reliability, sensitivity and specificity, as well as differences with diagnosis for each scoring approach. This comparison of procedures is designed to provide clinicians with as much psychometric information as possible about these scoring methods.

METHOD

Participants

Participants for this study were 493 persons who completed the Clock Drawing component of the Clock Test (Tuokko et al., 1992, 1995) and had a final diagnosis assigned at the conclusion of a comprehensive clinical examination as part of the Canadian Study of Health and Aging (CSHA). The CSHA was a nationwide prevalence survey for dementia. As part of this study, a large representative sample of people aged 65 and over living in the community (N = 9008) were interviewed in 18 centres across Canada and administered the Modified Mini-Mental State (3MS; Teng & Chui, 1987) Examination. Those with a score of less than 78 (roughly equivalent to a score of 22 on the Mini-Mental State Examination; n = 1614) and a sample with a score of 78 or more (n = 503) were seen for clinical examinations that assessed the presence of cognitive impairment and provided a differential diagnosis of dementia.

The clinical component was designed to confirm the presence of cognitive impairment in those who screened positive (<78 on 3MS) and to allow for a further differential diagnosis if cognitive impairment was confirmed. It consisted of four parts: nurse’s evaluation, physical examination, laboratory blood work, and neuropsychological assessment. The nurse’s evaluation included a readministration of the 3MS, rudimentary measures of vision and hearing, recordings of vital signs, height, weight, and medication use. Information on the participant’s history, cognitive and functional status was obtained
from a collateral informant (usually a family member) using Section H of the Cambridge Examination for Mental Disorders (CAMDEX; Roth, Huppert, Tym, & Mountjoy, 1988). During the physical examination, the physician evaluated general appearance of the participant, examined the head, neck, limbs, chest, and cardiovascular system and evaluated the primitive and central reflexes, peripheral neuromuscular responses, and coordination. The physician made a preliminary diagnosis on the basis of the information collected during the physical examination and by the nurse. Laboratory blood work was done for participants suspected of having dementia or delirium. Those attaining 3MS scores of 50 or above during the nurse’s evaluation were administered a standardized neuropsychological battery by a trained psychometrician (i.e., technician trained in test administration). The psychologist made a neuropsychological diagnosis on the basis of information collected by the psychometrician and the nurse. In case conferences typically attended by physicians, psychologists, nurses, and/or psychometricians, a consensus diagnosis was derived taking into account all clinically relevant information. Participants were classified using a three-stage process: confirmation of cognitive impairment, diagnosis of dementia using criteria from the Diagnostic and Statistical Manual of Mental Disorders, third edition, revised (DSM-III-R; American Psychiatric Association, 1987), and differential diagnosis based on DSM-III-R, National Institute of Neurological and Communicative Disorders and Stroke-Alzheimer’s Disease and Related Disorders Association (NINCDS-ADRDA) (McKhann et al., 1984), and International Classification of Diseases-10th revision (World Health Organization, 1987) criteria for depression, Alzheimer’s disease, and other dementias, respectively. The clinical assessment resulted in classification of participants into the following categories: No cognitive loss; Cognitive loss but no dementia; Alzheimer type (possible, probable) dementia; Vascular dementia; Mixed vascular + Alzheimer dementia; Other specific dementia; Unclassifiable dementia.

At the end of the study, the consistency of diagnoses among study clinical teams was investigated. Computer algorithms were used to check consistency between clinical findings and the consensus diagnosis with the rate of agreement for the distinction between dementia and non-dementia being 97.6% (Canadian Study of Health and Aging Working Group, 1994). In addition, 210 clinical assessment forms were selected at random and sent for blind redetermination of the diagnosis. Classification into dementia or no dementia yielded a kappa index of 0.81 (95% CI 0.73–0.89). The methods of the study have been described in detail elsewhere (Canadian Study of Health and Aging Working Group, 1994; Tuokko, Kristjansson, & Miller, 1995).

There were 232 males and 261 females identified for this study. The participants ranged from 65 to 100 years of age ($M = 79.28$, $SD = 6.98$). Years of formal education ranged from 0 to 25 ($M = 9.18$, $SD = 4.01$). On the basis of a complete clinical work-up (see Canadian Study of Health and Aging Working Group, 1994), the participants were classified as: (a) exhibiting dementia as defined by DSM-III-R criteria ($n = 80$); (b) Normal ($n = 276$), as defined by showing no areas of cognitive loss; or (c) showing cognitive impairment but no dementia (CIND, $n = 137$). The age, education, gender, and language preference for these groups are shown in Table 1.

**Procedure**

The clock face from the file of each of the 493 selected participants who were administered the clock-drawing component of the Clock Test (Tuokko et al., 1995) was scored using the methods developed by Shulman, Tuokko, Watson, and Wolf-Klein. The participant was presented with a predrawn circle (7 cm in diameter) and asked to: “Imagine this is the face of a clock. Put the numbers on the clock.” “Now, put the hands on the
The scoring for the Shulman and Tuokko procedures takes into consideration that the participant was asked to set the hands of the clock to 10 past 11. The Shulman procedure ranks the clock drawings on a scale of 1 to 6, with 1 representing a “perfect” clock and thus a low level or no impairment and 6 representing the highest level of impairment. The published account of this scoring procedure (i.e., Shulman et al., 1993) provides examples of each clock rating, but no cut-off ranking for impairment. However, a previously published study used an equivalent rating greater than two as a threshold to indicate impairment (Shulman et al., 1986).

For the Tuokko procedure, clocks are scored using 25 different error types, which can be summed on seven subscales: omissions, perseverations, rotations, misplacements, distortions, substitutions, and additions (see Tuokko et al., 1995). The seven subscales can, in turn, be summed to obtain a total error score to assess the level of impairment. Low scores suggest minimal impairment with a score of 0 indicating error-free performance. There is no ceiling for error totals but within this sample a maximum score of 41 was obtained. A score of greater than two suggests impairment.

The Watson scoring procedure bases scores solely on number placement. Participants are not asked to indicate a time. Scorers are given a template that divides the clock face into quadrants and the score is based on number placement within each quadrant. Any three digits in a quadrant is considered correct. One point per quadrant is awarded for errors in the first three quadrants, while an error in the fourth quadrant receives 4 points. Scores for this procedure range from 0 (least impaired) to a maximum of 7 (most impaired). According to the authors, a score greater than 3 suggests clinically significant impairment.

The Wolf-Klein scoring method is similar to the Shulman method in its rating scheme (rating scale ranges from 1 [most impaired] to 10 [normal]). Like the Watson method, scores are based on number placement only. The published account of this method gives examples of each rating. The published cut-off for impairment is set at 7.

In addition to these clock drawing scores, another clock-drawing task (i.e., Doyon et al., 1991), which differed in administration from the four procedures already described, was administered to a subsample of the CSHA participants (n = 173; Normal = 112; CIND = 33; Dementia = 28) and these data were also examined. One centre administered the Tuokko version of the Clock Test as part of the screening component and Doyon’s version of Clock Drawing in the clinical component of the CSHA. In all 173 cases, the Doyon version was administered after the Tuokko version and, hence, the fact that this subsample was administered two clock drawing tasks would only affect scores on the Doyon clock, not the scores for the other scoring methods. The mean time between screening and clinical examination for the CSHA was 64 days (SD = 45.21).
For the Doyon clock drawing task, participants are presented with a predrawn circle containing the number 12 at the top of the clock face and a dot in the centre of the circle. Participants are then asked to write the remaining numbers and set the time to indicate 10 past 11. Scoring is based on the following criteria: numbers, number location, hands, hand location, and hand length. Scorers are provided with a template to aid scoring judgements. Scores ranged from 0 to 20 with a higher score indicating less impairment. The original paper used a score of 14.5 or greater to indicate normal performance (Doyon et al., 1991).

RESULTS

Intercorrelations

The scores derived from each procedure correlated significantly with all other scores. The correlations ranged from $-0.294$ (Doyon and Watson) to $-0.715$ (Tuokko and Wolf-Klein). Thus, although significantly related, the various methods are not so highly correlated as to be interchangeable.

Reliability

As with other clinical measures, it is important that scores derived from screening tools be consistent for the same judge who scores the same clock on repeated occasions (intra-rater reliability) or between two judges who score the same clock (inter-rater reliability).

Training of raters was conducted by the fourth author (N.O'R.) using published scoring materials and interaction with the authors of the various scoring methods. The random selection of subsets for rescoring was also performed by the third author (S.R.). The fourth author also served as the independent rater for determining the inter-rater reliability coefficient for the Tuokko method. Although the time taken for scoring by each method was not recorded, after training, scorers reported that no method took longer than 3 minutes, on average.

For each scoring method, one research assistant scored all of the clocks using that method. This person then rescored a randomly selected subset of 50 clocks to yield an intra-rater reliability calculation. A second rater scored another randomly selected subset of 50 clocks to yield an inter-rater reliability estimate. Pearson product-moment correlation coefficients were calculated. The results are presented in Table 2.

All correlations were statistically significant. Comparisons were calculated to determine if the reliability coefficients differed from each other using Fisher's $Z$ test (Kenny, 1987) for independent correlations. Bonferroni correction for multiple two-tailed comparisons provided a revised alpha level of .0025 (i.e., $\alpha = .025/10$).

The intra-rater reliability for the Tuokko method was significantly higher than the intra-rater reliability coefficients obtained for Doyon, Shulman, and Watson methods. Wolf-Klein intra-rater reliability was higher than that obtained for the Doyon and Watson methods. Inter-rater reliability coefficients for the Tuokko and Watson methods were significantly higher than those obtained for the Doyon, Shulman, and Wolf-Klein methods.

The percent agreement between raters was calculated for each scoring procedure to determine whether or not any systematic differences between the raters were apparent. This was done to determine both intra- and inter-rater agreement. Percent agreement
was determined as the number of pairs for which difference between scores (rater 1 score minus rater 2 score!) was no more than 1. The percent agreement when the same rater rescored 50 clock faces ranged from 76 to 100. The percent agreement when two independent raters scored the same clock faces ranged from 76 to 96%. Thus, it did not appear that systematic differences (i.e., one rater scoring consistently higher or lower than the other) between the raters were evident. Given that different types of scoring systems were being examined (i.e., ratings for the Shulman and Wolf-Klein methods and continuous scores for the Tuokko, Doyon, and Watson methods), it was not appropriate to investigate this further.

**Sensitivity and Specificity**

Sensitivity and specificity indices are used to assess how effectively a test discriminates among alternative states of health (Zweig & Campbell, 1993). In its simplest form, it involves distinguishing between those who are known to have a disease (i.e., disease positive) and those participants who are asymptomatic (i.e., disease negative) according to some gold standard. For this purpose, participants with a primary diagnosis of dementia formed the disease positive group and participants with a CSHA diagnosis of “normal” formed the disease negative group. Participants with a CIND diagnosis did not fall into either of these categories and were eliminated from this part of the analysis as this group appeared heterogeneous in terms of the types and reasons for cognitive impairment (i.e., these are not all people with early dementia). It may be expected that sensitivity values would decrease if they were to be included and the findings would be difficult to interpret.

Published impairment cut-off values for the clock-drawing scores were used to calculate the sensitivity and specificity for each scoring method. As shown in Table 3, the Tuokko and Shulman cut-off scores produced high sensitivities and relatively low specificities whereas the Wolf-Klein and Watson cut-offs produced relatively low sensitivities and specificities. The Doyon cut-off produced relatively low sensitivity and high specificity.

Receiver operating characteristic (ROC) curves, plotting the sensitivity versus one minus the specificity for all possible cut-off scores (Zweig & Campbell, 1993), were drawn for each scoring procedure to evaluate the overall sensitivity and specificity. While the data in Table 3 best illustrate the relations between sensitivity and specificities for each method, the comparison across scoring methods is best illustrated using ROC curves. An ROC curve provides an overall index for each scoring methods. Using a procedure developed by Hanley and McNeil (1983), the area under each curve was compared using a test statistic analogous to a Wilcoxon signed rank test. The closer the area is to one, the
more effectively that test discriminates between groups. The area under the ROC curve for the Watson procedure was significantly lower than the other procedures; no other significant differences were found (see Table 4; alpha level using a Bonferroni correction for multiple two-tailed comparisons = .025/10).

Differences by Diagnosis

One-way analysis of variance (ANOVA) was used to test for group (normal, CIND, dementia) differences in scores for each of the five clock drawing procedures. All analyses detected significant group differences [Doyon, F(2, 170) = 22.70, p < .000; Shulman, F(2, 490) = 52.82, p < .000; Tuokko, F(2, 490) = 46.08, p < .000; Watson, F(2, 490) = 15.35, p < .000; Wolf-Klein, F(2, 490) = 44.85, p < .000]. Using an alpha level of .05, the Shulman, Tuokko, and Wolf-Klein procedures distinguished scores for all groups as different from each other (normal different from CIND and dementia scores and CIND different from dementia scores) when Tukey’s test for multiple comparisons (Montgomery, 1991) was used. The Doyon and Watson procedures only distinguished normal scores from the scores obtained by the CIND and dementia groups. A tabulation of mean scores by final diagnostic categories is presented in Table 5. The magnitude of the differences between scores by group for each method is, at least in part, a reflection of the possible range of scores for the method. For example, the Tuokko scoring method

<table>
<thead>
<tr>
<th>Scoring Procedure/Method</th>
<th>Cut-Off-2</th>
<th>Cut-Off-1</th>
<th>Published Cut-Off</th>
<th>Cut-Off +1</th>
<th>Cut-Off +2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doyon score</td>
<td>16</td>
<td>15</td>
<td>14</td>
<td>13</td>
<td>12</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.61</td>
<td>0.57</td>
<td>0.54</td>
<td>0.46</td>
<td>0.39</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.83</td>
<td>0.97</td>
<td>0.91</td>
<td>0.93</td>
<td>0.96</td>
</tr>
<tr>
<td>Shulman score</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>1</td>
<td>0.93</td>
<td>0.66</td>
<td>0.71</td>
<td>0.91</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.12</td>
<td>0.48</td>
<td>0.48</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td>Tuokko score</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.99</td>
<td>0.95</td>
<td>0.91</td>
<td>0.81</td>
<td>0.66</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.15</td>
<td>0.35</td>
<td>0.5</td>
<td>0.6</td>
<td>0.69</td>
</tr>
<tr>
<td>Watson score</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.83</td>
<td>0.59</td>
<td>0.59</td>
<td>0.56</td>
<td>0.45</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.38</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
<td>0.81</td>
</tr>
<tr>
<td>Wolf-Klein score</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.95</td>
<td>0.94</td>
<td>0.74</td>
<td>0.55</td>
<td>0.40</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.34</td>
<td>0.34</td>
<td>0.72</td>
<td>0.87</td>
<td>0.91</td>
</tr>
</tbody>
</table>

**TABLE 4**

<table>
<thead>
<tr>
<th>Area under ROC Curves</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Scoring Procedure</td>
<td>Area</td>
</tr>
<tr>
<td>Doyon</td>
<td>0.8</td>
</tr>
<tr>
<td>Shulman</td>
<td>0.79</td>
</tr>
<tr>
<td>Tuokko</td>
<td>0.78</td>
</tr>
<tr>
<td>Watson</td>
<td>0.67</td>
</tr>
<tr>
<td>Wolf-Klein</td>
<td>0.79</td>
</tr>
</tbody>
</table>

ROC = receiver operating characteristic.
can result in scores exceeding 40 whereas scores for the Shulman method range from 0 through 6. Thus, the larger score differences between groups for the Tuokko method are a reflection of the scoring system and contribute to the high sensitivity of this approach noted earlier.

**DISCUSSION**

The purpose of this study was to examine the similarities and differences between clock-drawing scoring methods. Any judgement as to which method to use will ultimately be based on the specific needs and features of various settings. Only overall scores were examined in this context. However, some methods yielded additional information (e.g., types of errors) that may be of interest in a clinical context. This study validates all claims to the adequacy of the clock-drawing task for identifying dementia regardless of the scoring method applied. However, some important differences among the scoring methods were noted.

The Tuokko scoring procedure had higher intra- and inter-rater reliability than the other scoring procedures. The Doyon and Watson procedures yielded surprisingly low intra-rater reliability, since it was expected that these “objective” methods would do well in this regard. One potential source of error is how the scoring template is aligned with the numbers on the clock. If the method of lining up the template is not consistent, differences in scores may result on repeated scoring. Intra-rater reliability for the Shulman and Wolf-Klein scoring methods was high, but inter-reliability coefficients were low and most likely reflect the subjective nature of these scoring methods.

The Watson procedure yielded lower overall sensitivity and specificity (as indicated by the area under its ROC curve) as compared to the other clock drawing scoring procedures. This suggests that this procedure may not be as effective at distinguishing between dementia and normal participants as the other procedures. The relatively high sensitivities of the Shulman and Tuokko methods suggest they may be particularly useful as screening measures (Essex-Sorlie, 1995). In this context, most cases would be identified (i.e., high sensitivity) and further assessment would be needed to distinguish between the cases and controls (relatively low specificity). Alternatively, the Doyon method identified controls well (i.e., relatively high specificity), but missed many cases (i.e., relatively low sensitivity) and, hence, would be a poor screening measure.

Essex-Sorlie (1995) notes that in some cases, measures that are both highly sensitive

<table>
<thead>
<tr>
<th>Scoring Procedure</th>
<th>Normal</th>
<th>CIND</th>
<th>Dementia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doyon</td>
<td>18.04</td>
<td>15.02</td>
<td>13.52</td>
</tr>
<tr>
<td></td>
<td>(2.93)</td>
<td>(4.10)</td>
<td>(4.97)</td>
</tr>
<tr>
<td>Shulman</td>
<td>2.78</td>
<td>3.60</td>
<td>4.16</td>
</tr>
<tr>
<td></td>
<td>(1.19)</td>
<td>(1.13)</td>
<td>(1.16)</td>
</tr>
<tr>
<td>Tuokko</td>
<td>4.21</td>
<td>6.32</td>
<td>11.51</td>
</tr>
<tr>
<td></td>
<td>(4.66)</td>
<td>(6.00)</td>
<td>(9.31)</td>
</tr>
<tr>
<td>Watson</td>
<td>2.60</td>
<td>3.65</td>
<td>4.21</td>
</tr>
<tr>
<td></td>
<td>(2.60)</td>
<td>(2.58)</td>
<td>(2.64)</td>
</tr>
<tr>
<td>Wolf-Klein</td>
<td>8.00</td>
<td>7.27</td>
<td>5.54</td>
</tr>
<tr>
<td></td>
<td>(2.01)</td>
<td>(1.83)</td>
<td>(2.56)</td>
</tr>
</tbody>
</table>

CIND = cognitive impairment with no dementia.
and highly specific are most desirable. In these cases, the "cost" of both missing cases of and identifying false positive cases would be high. Follow-up testing to confirm the diagnosis may be expensive or traumatic for the participant. Certainly, this could be viewed as the case with dementia. Follow-up testing often requires a thorough neuropsychological evaluation and persons may be very concerned with the possibility that they are being investigated for dementia. None of the scoring procedures produced uniformly high sensitivities and specificities. Thus, clock drawing should not to be used as a definitive diagnostic tool for dementia but in conjunction with other tools. The choice of measure, and cut-off point to use, is a moral and economic decision rather than solely a statistical one.

It is possible that, in our sample, cut-off points other than the published ones would have resulted in better classification rates. Given that most of the cut-off scores were determined using clinic-based samples in comparison to normal controls, it is possible that different cut-off scores would be more appropriate for these participants who were, initially, randomly selected from the community. In fact, when different cut-off points were examined (Table 3), the published cut-off scores resulted in the best balance between sensitivity and specificity for the Watson and Wolf-Klein scoring methods. The best balance between sensitivity and specificity for the Shulman and Tuokko methods fell within one point of the published cut-off point. Of note, the Doyon scores may have been inflated as all participants who completed the Doyon task had been exposed previously to the Clock Drawing component of the Clock Test (Tuokko et al., 1995). Thus, the cut-off score for this measure in this sample which produced the best balance between sensitivity and specificity was 17.0 (sensitivity = .82; specificity = .78) in contrast to the published score of less than 14.5. Although it is not possible to determine with this data set whether or not a similar inflation in performance after previous exposure to the task would occur for the other scoring methods, the possibility exists and should not be ignored by clinicians. Scores within one point of the published cut-off scores for the other four scoring methods, that were not contaminated with previous exposure to a similar task, produced the optimal balance between sensitivity and specificity in this large sample. This, then, supports the use of these scores to guide clinical practice.

Finally, the ANOVA suggests that differences with respect to final diagnosis were detected by all of the measures. The Tuokko, Shulman, and Wolf-Klein procedures detected differences among the three diagnostic categories. The Doyon and Watson procedures only distinguished the normal participants from the impaired participants (i.e., CIND and dementia). Differences between CIND and dementia scores were not detected. These results confirm that clock drawing performance differs by level of cognitive impairment as defined by clinical diagnosis.

Some limitations of this study must be addressed. First, different scorers were used for different scoring methods whereas, ideally, raters would have been equally exposed to all methods. Although different raters were used for each scoring system, the high inter-rater reliability coefficients and percent agreement suggests that it is unlikely that personal characteristics of the raters are responsible for the identified differences. Second, the clocks were administered at eight different sites across Canada. However, given that the administration instructions were standardized and explicitly described, it is unlikely that regional differences in administration occurred. Finally, in applying the four scoring methods to the clock drawing obtained according to the procedures defined by Tuokko et al. (1995), an assumption is made that different administrations have no effect on the outcome. It may be that differences in administration, such as being asked to place hands to a specified time (e.g., Shulman, Tuokko), result in correction of errors that would not occur if this were not requested (e.g., Watson). It is not possible to address this possibility in this data set, but may be an interesting line of investigation for future studies.
The primary finding of this study is that different scoring procedures may vary in terms of reliability estimates, sensitivity, specificity as well as identification of diagnostic groups. Many versions of clock drawing are available and are not necessarily comparable or equal in utility. Clinicians require as much information as possible about the various procedures if these instruments are to be used appropriately and most effectively. This article is one source of information on some commonly referenced scoring methods for a clock drawing task.

REFERENCES


