Reliability and Construct Validity of the Clinical Observations of Motor and Postural Skills

Brenda Wilson, Nancy Pollock, Bonnie J. Kaplan, Mary Law, Peter Faris

Key Words: assessment process, occupational therapy • motor skills • test construction • tests, by title, Clinical Observations of Motor and Postural Skills

Children with subtle motor coordination problems present assessment and intervention challenges to occupational therapists. Although motor incoordination may present as a minor problem in athletic ability, it is often associated with far more devastating problems, including difficulty in social situations, loss of self-confidence and motivation, communication impairments, difficulties in activities of daily living, and academic problems (Levine & Satz, 1984; McKinney, 1989; Regehr & Kaplan, 1988; Reuben & Bakwin, 1968; Schaffer, Law, Polatajko, & Miller, 1989). Several different labels have been applied to this disability, including developmental dyspraxia (Ayres, 1985; Denckla, 1984), clumsiness (Hall, 1988), and developmental coordination disorder (DCD) (American Psychiatric Association [APA], 1987). The prevalence of DCD is estimated at 6% of children between the ages of 5 and 11 years (APA, 1987). A 10-year follow-up study of children with clumsy child syndrome showed that well into their adolescence these children continued to have motor difficulties as well as educational, social, and emotional problems (Losse et al., 1991). Ehrhardt, McKinlay, and Bradley (1987) found that children with mild delay tended to catch up with their peers as they grew older, but those with a moderate delay still exhibited motor problems at the time they left school.

In the assessment of children who present with DCD, occupational therapists have relied both on standardized tests and on informal observations made during testing. The Southern California Sensory Integration Tests (Ayres, 1972b), the Sensory Integration and Praxis Tests (Ayres, 1989), the Bruininks-Oseretsky Test of Motor Proficiency (Bruininks, 1978), the Peabody Developmental Motor Scales (Folio & Fewell, 1983), and the Developmental Test of Visual-Motor Integration (Beery, 1982) are frequently used by occupational therapists to determine if children have a DCD (Reid, 1987).

Ayres (1972a, 1976) proposed the use of clinical observations to supplement the information received from
these standardized measures. These 11 clinical observations include items testing postural stability, coordination, rate and quality of movement, and ability to inhibit primitive reflexes. The clinical observations are closely related to tests of cerebellar-vestibular function, or soft neurological signs. The protocol of 19 items developed by Ayres’ colleague, Johnson (1977), has not been standardized in its administration or instructions, and it relies on a 3-point rating scale that reflects the therapist’s judgment. Despite these obvious limitations, a survey of occupational therapists working with children with DCD showed that 80% use the clinical observations as part of their routine assessment (Yack, 1989).

Several attempts have been made to improve the objectivity of the clinical observations, but all have limitations. For example, Dunn (1981) developed objective administration and scoring procedures for use with kindergarten children. However, reliability and validity were not assessed and the application of these procedures is limited to 5-year-olds. Another example is the Balcones Sensory Integration Screening tool, which included some clinical observation items (Jones & Monkhouse, 1981). The scope of this measure is broad, however, and it is meant to identify the need for further testing rather than to provide a diagnosis. Additionally, the scoring system is not sensitive to differing degrees of ability. The Quick Neurological Screening Test, developed by Motti, Sterling, and Spalding (1978), is a brief screening procedure to classify children according to one of three categories: (a) children with no problems, (b) children with distinct neurological signs, and (c) children without frank neurological signs who do not demonstrate age-appropriate skills. Cermak, Ward, and Ward (1986) found the Quick Neurological Screening Test difficult to use with children under 7 years of age, because the test tends to erroneously identify problems in 5- and 6-year-old children. This deficiency results in limitations in the widespread use of the Quick Neurological Screening Test; only 17.5% of pediatric occupational therapists who were surveyed reported using this test (Yack, 1989).

Several authors have studied some individual clinical observation items in terms of their developmental characteristics (Bowman & Katz, 1984; Bundy & Fisher, 1981; Gregory-Flock & Yerxa, 1984; Harris, 1981; Parmenter, 1981; Zemke, 1983). Even though these studies have shown that there are differences in performance on the clinical observations across different ages, therapists continue to use the same scoring criteria for all children. So far, the research in this area has not resulted in a formal assessment. As an informal assessment, the clinical observation does not have demonstrated reliability or validity but is being used to make decisions about which children exhibit dysfunction and which children may benefit from intervention.

In summary, children with coordination dysfunctions have been a challenge for researchers and clinicians. Assessment of some skills has been carried out with non-standardized tools for which reliability and validity have not been demonstrated. As Henderson (1987) stated, “There remains a dearth of instruments of demonstrated reliability and validity from which the ordinary clinician can make an informed selection” (p. 518).

Test Development

To assist occupational therapists in performing valid and reliable assessments of children with suspected DCD, we developed the Clinical Observations of Motor and Postural Skills (COMPS). The COMPS is based primarily on the original clinical observations protocol developed by Ayres (1976) and Johnson (1977); it is not a new test, but rather a revision and elaboration of a commonly used test. Important objectives in developing the COMPS included the standardization of administration procedures and the development of objective criteria for scoring responses. The COMPS is intended to be used mainly as a screening tool for identification of motor problems (descriptive) and possibly as an evaluative instrument to measure change over time.

Of the 19 items in the original protocol (Ayres, 1976; Johnson, 1977), 9 were eliminated because of difficulties in standardizing the scoring criteria. Three items were eliminated because they did not discriminate across age groups. One item, ocular pursuits, was eliminated for two reasons: first, because Gilligan, Mayberry, Stewart, Kenyon, and Gaeber (1981) had developed a formal test of oculomotor function, thus its inclusion in the COMPS seemed redundant, and second, oculomotor testing by occupational therapists is controversial within the field of pediatrics. The remaining six items were revised and a seventh item was added (finger-nose touching). The COMPS was then used by therapists in two cities (Calgary, Alberta, and Hamilton, Ontario), and revisions were made based on their feedback. The version evaluated in the present study (referred to as the research edition) included items chosen on the basis of clinical utility and objectivity. The test was not designed to measure one domain of motor control; rather, the COMPS appears to be related to cerebellar function, postural control, and motor coordination. Once the research edition of the COMPS was completed and a manual had been developed and revised, we received funding to evaluate test-retest reliability, interrater reliability, internal consistency, and construct validity.

Test Description

The research edition of the COMPS consisted of 7 items, as are described below.

**Slow motion.** Developed from the Dunn (1981) protocol, this item measures the ability to move slowly and symmetrically (ramp measurement). The COMPS item is...
performed within 6 sec and provides a 3-point scoring system in three categories: symmetry, quality of performance, and speed. Control of slow, smooth movements is associated with cerebellar function.

**Finger-nose touching.** Developed from the Jones and Monkhouse (1981) protocol, this item measures cerebellar coordination. The criteria for scoring have been clarified. The administration was changed to ensure that the action is done continuously between the eyes open and eyes closed sequences so that the proprioceptive mechanisms of motor control are facilitated. By 7 years of age, a child should not miss contact with either the nose or the finger more than once (Touwen, 1979).

**Rapid forearm rotation.** This test of diadochokinesis was developed from the protocols of Dunn (1981) and Levine, Brooks, and Shonkoff (1980) and is commonly used as a test of cerebellar integrity. The skill matures at age 7 (Levine et al., 1980) to 8 (Touwen, 1979) years. Scoring was based on the number of forearm rotations accurately completed in 10 sec.

**Schilder's arm extension.** The administration and scoring of this item were slightly adapted from the Dunn (1981) protocol. This item measures the degree of body rotation accompanying passive head turning in the standing position to allow for assessment of the degree to which a child's arm movements are disassociated from the head movements. It also evaluates evidence of choreiform or athetoid movements of the fingers. These signs are considered to be indicative of basal ganglia involvement.

**Prone extension posture.** This item was further developed from the works of Bundy and Fisher (1981), Harris (1981), and Bowman and Katz (1984). It measures the ability to assume and maintain an arched back position against gravity, which is theorized to be related to vestibular-propioreceptive processing dysfunction (Fisher, Murray, & Bundy, 1991).

**Asymmetrical tonic neck reflex.** The rating scale was taken from Parmenter (1981) and measures the degree of inhibition of the asymmetric tonic neck reflex in the quadrupedal position. The developers of the COMPS designed a measurement tool to accurately measure the degree of elbow flexion. Whereas this item accurately measures the degree of asymmetric tonic neck reflex present, we found that it was also useful in identifying children who lack the postural stability to maintain a static quadrupedal position without locking their elbows or using an excessively wide base of support, regardless of the position of their head. This item could be alternatively called postural stability.

**Supine flexion posture.** This item was developed to be graded in a fashion similar to the Prone Extension Posture. It measures the ability to assume and maintain a flexed posture in the supine position and provides objective scoring criteria for each body segment, in addition to the ability to maintain the posture over time. Fisher et al. (1991) find this posture (especially the position of the neck) to be associated with somatodyspraxia.

### Method

#### Subjects

One hundred twenty-three children participated as subjects in this study: 67 children with DCD and 56 nondysfunctional children who served as the control group (see Table 1). The breakdown of subjects by sex and age is shown in Table 1. Sixty-nine of the children were recruited in Calgary and 54 in Hamilton, Ontario. The children with DCD were referred from the Department of Occupational Therapy at the Alberta Children's Hospital, the Occupational Therapy and Physiotherapy Services of the Calgary Board of Education, the Calgary Renfrew Early Childhood Services, a treatment efficacy study at the Behavioural Research Unit of Alberta Children's Hospital Research Centre, and the School Health Support Services Program at Chedoke-McMaster Hospitals. The children without motor problems were a convenience sample recruited through ads and acquaintances in both Calgary and Hamilton.

Subjects with DCD were defined as children who had been identified by their teachers as not performing comparably to their peers in motor skills or who had been referred for remedial services of occupational therapy or physiotherapy despite at least average intelligence, adequate sensorimotor systems, and adequate learning opportunities. All of the subjects in the final sample were identified by their teachers as having significant motor problems, with 72% of these children also having academic problems and 75% having social difficulties.

Confirmation that the nondysfunctional subjects of school age had no motor problems was provided by their teachers with the use of the following criteria: (a) at least average performance in all academic subjects, (b) no previous or present involvement in remedial services, (c) regular class placement, and (d) no apparent deficits in motor areas, such as balance, coordination, handwriting, or fine motor skills. Confirmation for nondysfunctional preschool children was provided by their nursery school or early childhood services teachers, who reported that

### Table 1

<table>
<thead>
<tr>
<th>Composition of Sample (N = 123) by Age and Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>5 years 0 months-5 years 11 months</td>
</tr>
<tr>
<td>6 years 0 months-7 years 11 months</td>
</tr>
<tr>
<td>8 years 0 months-9 years 11 months</td>
</tr>
<tr>
<td>% of Total</td>
</tr>
</tbody>
</table>

Note: DCD = developmental coordination disorder.
these children were performing comparably to their peers in preacademic and motor areas and that they were not involved in remedial services.

The subgroup of subjects whose assessments were used to measure construct validity were given the Motor Accuracy Test–Revised (Ayres, 1980) and the Developmental Test of Visual Motor Integration (Beery, 1982) to confirm the absence of a motor problem. This subgroup of nondysfunctional and DCD children were matched for age, sex, and socioeconomic status, as measured by Blishen and McRoberts (1976). Independent sample t tests of the entire sample revealed no significant group differences for age [t(120) = .29] or for socioeconomic status [t(114) = .99]. Socioeconomic status was obtained for only 116 subjects, because seven families were unwilling to disclose their occupations. A chi-square analysis revealed no significant group differences in sex [X^2(1) = 2.52].

Procedure

Testers. Ten qualified occupational therapists tested the children or scored the COMPS from videotapes. One therapist with 17 years of experience with adults was chosen for her lack of experience in pediatrics, to allow examination of interrater reliability between a therapist experienced in pediatrics and one with no pediatric experience. In addition, a third-year student in the undergraduate occupational therapy program at the University of Western Ontario was involved in scoring 40 of the subjects’ tapes to determine reliability with occupational therapy students. The other eight therapists had a mean of 7.9 years of pediatric experience (range = 3 to 13 years; SD = 4.34).

A brief in-service workshop on the COMPS was provided to all testers in Calgary and Hamilton, respectively. This consisted of viewing a demonstration tape showing the administration procedures and jointly marking one child’s performance. Following this in-service training, each tester scored two subjects’ videotaped performances independently, and these were compared to the score of the test developers. Agreement of 90% or higher was considered acceptable; lower agreement resulted in examination of areas of disagreement and further discussion with the therapist. The in-service training and discussion were restricted to one session, so that this study would evaluate the test and its manual rather than personal instruction by the developers of the test.

Testing. Consent was obtained from subjects and their parents. Testers followed administration procedures as outlined in the manual. When necessary, visual cues (demonstration) and verbal cues (extra instruction) were given to optimize the child’s performance. Children were encouraged to do their best and were given the opportunity to repeat an item whenever the tester felt that the level of motivation was below their motor ability.

All test sessions occurred in a quiet, well-lit room, free from distractions. Testing equipment involved two chairs, a mat to lie on, a stopwatch, a blindfold, and an acrylic plastic measuring tool for the asymmetric tonic neck reflex item.

Not all of the 123 subjects participated in all parts of this study. Fifty-three subjects were administered the COMPS twice within 2 weeks for evaluation of test-retest reliability. The second test was completed at the same time of day, by the same tester, and in the same location. The subjects’ motivation and cooperation were assessed on a 3-point scale to allow comparison of their behavior at the time of retesting. If a subject received an assessment of “uncooperative” on a test, that test score was eliminated from the analysis. Five subjects were eliminated from analysis of test-retest reliability because their cooperation or motivation was inadequate for testing on one or both tests. These five subjects represented 3 of the 31 nondysfunctional children and 2 of the 22 children with DCD, an equal ratio across the two groups. Three of the 5 subjects eliminated were 5-year-olds.

Seventy-two of the 123 subjects’ tests were videotaped and subsequently scored by several therapists for analysis of interrater reliability. Only 64 of these 72 videotaped tests were also used in the analysis of construct validity, because only these 32 pairs were matched one to one for age, sex, and socioeconomic status. For analysis of both the interrater reliability and the construct validity, the therapists were unaware of the child’s condition.

Data Analysis

Pearson product-moment correlations were used to provide information on the relationship between a pair of tests. However, Cicchetti and Conn (1976) stated that the Pearson correlation coefficient takes into account only the extent to which two independent observers put their measurements of the same subjects in the same order but tends to ignore the magnitude of the discrepancy on individual pairs of ratings. As a result, raters can be far apart on individual measurements, but, as long as the trend of their ranking is similar, the extent of the correlation will be high, giving the impression of better agreement than actually exists. (pp. 375-376)

Therefore, intraclass correlation coefficients (Fisher, 1967; Hartman, 1982; Streiter & Norman, 1990) were also used to provide an estimate of reliability, which accounted for changes in the standard deviation from one test to the next. An intraclass correlation is the ratio of the variance of interest over the sum of the variance of interest plus error. It allows us to measure how much of the variability among observers is due to the raters and how much is due to differences among the subjects and their test scores. There are several forms of intraclass correlation coefficients (cf. Shrout & Fleiss, 1979), the form of choice being determined by the design of the study. Because the period of time that elapses between testings...
could be expected to influence test-retest reliability in a systematic fashion, for test-retest analysis we chose the model for the intraclass correlation coefficient in which the effects of time were treated as fixed. This corresponds to Shrout and Fleiss’s (1979) intraclass correlation coefficient (Model 3,1):

\[ ICC(3,1) = \frac{BMS - EMS}{BMS + (K - 1) EMS}, \]

where ICC = intraclass correlation coefficient; BMS = between-targets mean square; EMS = residual mean square; K = number of repeated measures.

We did not calculate the pooled within-groups correlation because we considered there to be too many raters (7) and because the sample sizes for these raters ranged from 1 to 17.

Because the effects of randomly selected raters on the interrater reliability of a measure would not be expected to be systematic and because we wished to have an estimate of the reliability that could be expected among other randomly selected judges, we chose Shrout and Fleiss’s (1979) model of the intraclass correlation coefficient that treats judges as random effects (Model 2,1):

\[ ICC(2,1) = \frac{BMS - EMS/BMS + (K - 1) EMS}{BMS + (K - 1) EMS + K(EMS - EMS)/n}, \]

where ICC = intraclass correlation coefficient; BMS = between-targets mean square; EMS = residual mean square; JMS = between-judges mean square; K = number of judges; n = number of subjects.

In the random-effects model, the intraclass correlation coefficient value is usually lower than a Pearson correlation coefficient, and a value greater than .65 is generally an acceptable level (Kirshner & Guyatt, 1985). However, the fixed-effects model of intraclass correlation coefficient does not adjust for additive differences (differences in the mean), although it does adjust for multiplicative effects (interactions between time and subjects). With the fixed-effect model, the values are often comparable to those obtained with Pearson correlations.

Statistics were performed with the use of SPSS/PC+ version 4.0 (Norusis, 1990). For calculation of intraclass correlation coefficient, the appropriate mean squares were computed with a repeated-measures analysis of variance (ANOVA) through the use of a multivariate analysis of variance procedure (MANOVA). The intraclass correlation coefficients were then calculated from this information with the assistance of a spread sheet. All t tests, whether independent or paired, were performed with the t-test procedure. Pearson product-moment coefficients were calculated with the correlation procedure. Multiple regressions were calculated with the regression procedure. ANOVAs and MANOVAs were calculated with the use of the MANOVA procedure.

Results

Reliability

The stability and consistency of the COMPS over time (test-retest reliability) and between testers (interrater reliability) were analyzed with the intraclass correlation coefficients for the reasons given above. The accepted level of reliability for this study was set at .75.

The fixed-effects model of intraclass correlation coefficient used to analyze test-retest reliability revealed that reliability exceeded .75 when COMPS was readministered within a 2-week period for both groups and for the total sample. For the DCD group, the intraclass correlation coefficient was .85 (n = 20); for the control group, .79 (n = 28), and for the total sample, .92 (n = 48) (p < .001). Coefficients calculated for each age group also exceeded .75, as follows: (a) .88 for 5-year-olds; (b) .92 for 6- to 7-year-olds, and (c) .93 for 8- to 10-year-olds. Pearson coefficients were also computed and were almost identical in value, indicating that there were no multiplicative systematic differences across tests.

Intrarater reliability was analyzed for four different raters who had three different levels of experience: two experienced in pediatrics, one had experience in pediatrics, and one was an occupational therapy student. The random effects model used for this analysis takes into account the correlation as well as the absolute differences between scores. For this study, a correlation of .75 was considered to be acceptable. As shown in Table 2, correlations were greater than .75 for ratings by experienced therapists for both groups of children. Coefficients were .60 for the DCD group when a therapist with no pediatric experience was compared to a therapist with such experience, but were above .75 for the total group and the group of normal children. Interrater reliability between a student and an experienced pediatric therapist was below .75.

Another estimate of reliability, internal consistency, is the measure of how much each item contributes to the total assessment, or how well all items are measuring the same construct. Cronbach’s coefficient alpha was used, with .70 denoted as an acceptable level. The alpha coefficient for the total test was .75, which is a moderate but acceptable level of internal consistency. Item-total correlations were computed (Pearson correlation coefficients), which ranged from .20 for Schülker’s arm extension to .53

Table 2

<table>
<thead>
<tr>
<th>Group</th>
<th>OT With Pediatric Experience</th>
<th>OT With No Pediatric Experience</th>
<th>OT Student</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCD group</td>
<td>n</td>
<td></td>
<td>n</td>
</tr>
<tr>
<td></td>
<td>48</td>
<td>45</td>
<td>42</td>
</tr>
<tr>
<td>Nondysfunctional</td>
<td>29</td>
<td>30</td>
<td>-</td>
</tr>
<tr>
<td>Total sample</td>
<td>72</td>
<td>72</td>
<td>38</td>
</tr>
</tbody>
</table>

Note. OT = occupational therapist; DCD = developmental coordination disorder.
*p ≤ .001.
for finger-nose touching. Correlation alphas with each item deleted were also computed. Alphas ranged from .45 for finger-nose touching and prone extension posture to .55 for Schilder's arm extension. Schilder's arm extension had the lowest correlation compared with the total test score, and the deletion of this item resulted in a greater increase to the alpha coefficient than when other items were deleted (see Table 3).

**Construct Validity**

MANOVAs were calculated with age as a covariate (MANCOVA) to determine whether children with DCD and nondysfunctional children differed on their total COMPS score (see Table 4). Analyses within each of the three age groups did not involve age as a covariate. DCD children and nondysfunctional children differed significantly on the COMPS total score for the total sample and for each age level.

A discriminant analysis was done to assess the degree to which each item contributed to the ability of the COMPS to discriminate between nondysfunctional and DCD children for each age group. Because the SPSS/PC + discriminant procedure does not allow the use of covariates, the dependent variables were regressed on age and the residual scores analyzed. A forced-entry procedure was used for the analysis. As shown in Table 5, the COMPS items were generally quite strong in discriminating between the two groups. The univariate F was calculated after the items had been regressed on age; it examines the significance of the group differences on each separately analyzed item. The structure coefficients, or canonical variate correlates, measure how strongly each item is related to the discriminant function. The standardized discriminant function coefficients indicate the unique contribution of each item to the test's ability to discriminate between children with and without problems. The term antiques is used because the standardized coefficients indicate the contribution of the variable after the effects of the other variables have been controlled. Whereas these coefficients are affected by the intercorrelations among the variables and also by the variability of the variables with which they are associated, the standardized function coefficients are generally considered to be the indexes of choice in determining the contribution of each variable to the discriminant function (Bray & Maxwell, 1985; Klecka, 1980; Pedhazure, 1982). Across all analyses, Schilder's arm extension posture was the poorest at discriminating between groups.

The sensitivity and specificity of the total COMPS score for each age group was measured. The test's sensitivity (i.e., its ability to detect dysfunction when it is actually present) was 100% for 5-year-olds and for 8- to 10-year-olds and 85% for 6- to 7-year-olds. Specificity, or

### Table 4

<table>
<thead>
<tr>
<th>Age Group</th>
<th>DCD Group</th>
<th>Nondysfunctional Group</th>
<th>F (d.f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 years</td>
<td>M (SD)</td>
<td>M (SD)</td>
<td></td>
</tr>
<tr>
<td>0 months-5 years</td>
<td>42.04 (10.30)</td>
<td>55.32 (7.03)</td>
<td>12.46** (1, 20)</td>
</tr>
<tr>
<td>11 months (n = 22)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 years</td>
<td>52.64 (10.18)</td>
<td>63.68 (7.81)</td>
<td>8.15* (1, 20)</td>
</tr>
<tr>
<td>0 months-7 years</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 years</td>
<td>54.65 (7.55)</td>
<td>68.90 (6.58)</td>
<td>20.24*** (1, 18)</td>
</tr>
<tr>
<td>11 months (n = 20)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total sample</td>
<td>49.62 (10.77)</td>
<td>62.44 (8.05)</td>
<td>42.28*** (1, 61)</td>
</tr>
</tbody>
</table>

Note: Maximum possible score = 84. DCD = developmental coordination disorder. *p < .01. **p < .002. ***p < .001.

### Table 5

<table>
<thead>
<tr>
<th>Item</th>
<th>Univariate F</th>
<th>Structure Coefficient</th>
<th>Standardized Discriminant Function Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rapid forearm rotation</td>
<td>21.7**</td>
<td>.64</td>
<td>.51</td>
</tr>
<tr>
<td>Prone extension posture</td>
<td>15.1**</td>
<td>.53</td>
<td>.50</td>
</tr>
<tr>
<td>Asymmetrical tonic neck reflex</td>
<td>13.6**</td>
<td>.50</td>
<td>.51</td>
</tr>
<tr>
<td>Finger-nose touching</td>
<td>12.9**</td>
<td>.49</td>
<td>.16</td>
</tr>
<tr>
<td>Slow motion</td>
<td>8.7**</td>
<td>.40</td>
<td>.29</td>
</tr>
<tr>
<td>Supine flexion posture</td>
<td>6.9**</td>
<td>.35</td>
<td>.19</td>
</tr>
<tr>
<td>Schilder's arm extension</td>
<td>9.4</td>
<td>.09</td>
<td>-.03</td>
</tr>
</tbody>
</table>

*p < .01. **p < .001.
the extent to which the test identifies nondysfunctional children as such, was 85% for 5-year-olds, 63% for 6- to 7-year-olds, and 100% for 8- to 10-year-olds. Using a canonical discriminant analysis, which identifies cutoff scores for each age group, we found that the test correctly classified 73.44% of all the subjects. The classification percentages for individual age groups were high for the youngest and oldest groups and lower for the 6- and 7-year olds.

Discussion
The results of this study indicate that the COMPS is consistent and reliable across time and observers, particularly when experienced observers are used. Test-retest reliability was generally excellent. The lower test-retest correlations for the nondysfunctional and DCD groups compared with the total group were due to the fact that an overall wide range of scores in the total sample produced higher correlations than subgroup scores, which were more homogeneous. Scores for the second test tended to be slightly higher than the first test, which is likely related to a learning effect. The similar values obtained with the use of Pearson correlation coefficients indicated that there was no interaction between the time of testing and the test score; for example, children with high scores did not consistently score lower on the second test or vice versa. However, test-retest reliability may have been influenced by the fact that the two tests were administered by the same person. Although it is desirable to have the same tester on both occasions, which eliminates extraneous variance (the difference between raters), it allows for the possibility that the testers will remember, to some extent, the child’s performance on the first test when marking the second. This evaluator bias may have positively influenced reliability results. In an effort to control for this, we used a 2-week period between test sessions to decrease both the tester’s and the subject’s memory of the test.

The correlations obtained across raters were acceptable, with all but one correlation greater than .75, regardless of whether the therapists were experienced in pediatrics. The coefficients comparing a student rater to an experienced rater were not acceptable or adequate for a test of reliability. Even though interrater reliability was acceptable overall, examination of the mean scores across raters shows some systematic individual differences. The more experienced testers rated children approximately 5 points higher than the therapist who lacked experience in pediatrics. However, the occupational therapy student tended to give higher scores than any of the three therapists. This difference in scoring is an example of raters having “individual standards that interact with the examination items and/or examinees and are manifested as differing levels of severity” (Lunz & Stahl, 1990, p. 427). Lunz and Stahl (1990) demonstrated that the multiple-facets Rasch model can be used to adjust scores for the severity of the raters. Because the Rasch model is in its early developmental stages, it was not used in this study, but may be considered in the further development of this test.

The high internal consistency of the test indicates that the COMPS is measuring a single construct, presumably skills that underlie movement and posture. The item coefficients remained fairly constant when each item was deleted from the analysis, with the exception of Schilder’s arm extension, discussed below.

In the discriminant analysis, which examines how much each item contributed to the ability of the COMPS’ discriminant function to identify children with and without motor problems, six of the items were acceptable, based on both the significance tests and their high (> .30) structure coefficients. Analysis of the discriminant function of each item relative to the different age groups reveals that the slow motion, prone extension, and supine flexion items are the best discriminators for 5-year-olds. The finger-nose touching and rapid alternating movement items discriminate best for 6- and 7-year-olds. These latter two items plus prone extension and asymmetric tonic neck reflex discriminate best for 8- to 10-year-olds. These items represent all test items with the exception of Schilder’s arm extension.

The COMPS total score discriminates between nondysfunctional children and children with DCD. The relatively large standard deviation of the DCD group combined with the smaller difference between the means of the DCD and nondysfunctional groups resulted in an overlap of scores and less discriminative ability for the 6- to 7-year age range. It is likely that at the ages of 6 and 7 years there is a large amount of maturation of the skills that the COMPS assesses, because the items are chosen to reflect skills that matured by the age of 8. This would result in a wider distribution of scores within the normal range and would therefore result in more difficulty in discriminating between nondysfunctional children with low scores and children with DCD.

Most of the analyses of reliability and validity revealed the strength of all of the COMPS items except Schilder’s arm extension. Internal consistency showed more improvement with this item removed than with the removal of any other item. It was lowest in its ability to discriminate between children with and without DCD. The low correlation between Schilder’s arm extension and the total test score may be indicative of the poor relationship between this skill and the construct being assessed. However, it may also be related to the difficulty in administering this item correctly: children have a tendency to peek under the blindfold and move their feet as their heads are turned. These behaviors are difficult for an examiner to observe and the item is difficult to score from a videotape. The fact that all other items in the COMPS discriminated well between the groups and
The results of our sensitivity, specificity, and prediction value analyses are similar to the results of the MANOVA. The test discriminates well for all ages, with slightly lower values for the 6- and 7-year age group than for the other two age groups. The test accurately identifies children with motor problems.

King-Thomas and Hacker (1987) defined screening as the collection of preliminary information, and they differentiated it from evaluation that offers a delineation of strengths and weaknesses or a diagnosis or a need for intervention. The COMPS is designed to screen for the presence or absence of postural and motor problems. A low score indicates a need for careful examination of the items that are most problematic and consideration of further assessment in the areas identified. At this point in its development, the ability of the COMPS to predict problems or its ability to measure change over time has not been established. However, as a tool to identify motor dysfunction, it demonstrates higher reliability than other published tests, such as the ones described by Quittin, Rifkin, and Klein (1976) and Vitelli, Ricciuti, Stoff, Behar, and Denckla (1989). The evidence suggests that the COMPS correctly identifies a large percentage of children with motor and postural problems. In relation to the criteria of a good screening tool (King-Thomas & Hacker, 1987; Krishner & Guyatt, 1985; Law, 1987; Law & Polatajko, 1987), the COMPS has many properties of a good test: it is short (20 min), it is easy to administer, it is fun for the child to do, and it has adequate reliability and construct validity.

Implications for Occupational Therapy

This study demonstrates that the COMPS requires few revisions before it is ready for clinical use. The Schilder’s arm extension item will be dropped from the test, and the reliability and validity will be reanalyzed with this item removed. Given that these analyses will be as good or better when the least useful item is eliminated, we will complete the test manual and make the test available in the near future. Further work is underway to assess the concurrent validity of the COMPS, comparing this new test to other measures of postural control and motor function.

The generalizability of this study is limited by the relatively small sample size and the fact that the control subjects were a convenience sample and therefore may not be representative of the whole population. Because this was not a normative study and the test is not a diagnostic test, the COMPS does not offer much descriptive information on the nature of the occupational performance problem. However, as a screening measure, it will be valuable in identifying children who have motor problems with a postural component and who may require direct treatment. As screening and assessment time available to therapists becomes more limited and the need to determine priorities for treatment becomes paramount, the COMPS will provide an assessment that is more reliable and valid than previous instruments used. Although not a new measurement tool, it is an improvement over the clinical observations currently in use, and further research can provide us with more information on the domain or domains of motor skills that it assesses.

Acknowledgments

We thank the children who consented to participate, the therapists who provided valuable feedback at the inception of this test, and the therapists and research assistants involved in this study, particularly Robin Hall in Hamilton and Susan Crawford in Calgary. This study was completed with the financial support of the Hospital for Sick Children Foundation, Toronto, Ontario, and the Alberta Children’s Hospital Foundation, Calgary, Alberta.

References

Cicchetti, D. V., & Conn, H. O. (1976). A statistical analysis...
of reviewer agreement and bias in evaluating medical abstracts. *Yale Journal of Biology and Medicine, 49*, 373–383.


