The Movement Assessment Battery for Children (Movement ABC; Henderson & Sugden, 1992) is now recognized as one of the most widely used tests of motor impairment in the world (Geuze, Jongmans, Schoemaker, & Smits-Engelsman, 2001). It contains three components: a standardized test, a teachers’ checklist, and a set of intervention guidelines. The standardized test comprises four age bands, suitable for children aged 4–6, 7–8, 9–10, and 11–12 years. The items in each age band are designed to assess manual dexterity, ball skills, and balance, and item difficulty increases as age band ages increase. Norms are provided for ages 4 through 12 years (Henderson & Sugden).

The Movement ABC has been translated into six European languages and published studies employing it in Europe and North America are numerous (see Barnett & Henderson, 1998). However, studies emanating from the Far East have been few. Accordingly, we have initiated a series of investigations into its suitability for use in Hong Kong, Taiwan, and Greater China. In view of the importance now attached to early identification and intervention for children with movement difficulties, we decided to begin with an examination of the youngest age band (ages 4–6 years). Partly due to the early start of preschool education in Hong Kong, the preschool period also seemed likely to be the one in which the motor experience of Chinese children was most likely to differ from that of their Western counterparts. The present investigation is concerned with reliability. In another parallel study (Chow, Henderson, & Barnett, 2001), we sought to detect cross-cultural differences in the level of performance on the component items.

Here, we seek to determine whether cultural factors are reflected in the agreement between different testers (intrarater reliability) or in the short-term variability of a child’s performance from occasion to occasion (test–retest reliability). The Movement ABC does not contain a standard set of verbal instructions. Rather, the tester is encouraged to use any combination of physical demonstration and verbal guidance they consider appropriate to ensure that a child understands the tasks. While this characteristic makes the test par-

**Brief Report**

Interrater and Test–Retest Reliability of the Movement Assessment Battery for Chinese Preschool Children

Susanna M. K. Chow, Sheila E. Henderson

**Objective.** The aim of this paper was to check on the reliability of the Movement Assessment Battery for Children (Movement ABC) in preparation for its standardization in Hong Kong. Interrater and test–retest reliability are reported for Age Band One, designed for use with children ages four to six.

**Method.** Interrater reliability of the Movement ABC was estimated using two trained observers with 79 children. Test–retest reliability was assessed using 75 children who were tested twice by the same tester over a 2- to 3-week interval.

**Results.** Agreement between testers was good with a mean intraclass correlation of 0.96 across items. A value of 0.77 was obtained for test–retest reliability.

**Conclusion.** These results support the use of this component of the Movement ABC in Hong Kong.

particularly suitable for children with cognitive and attention problems, it may also be a source of disagreement (variability) between scoring by testers. Moreover, we thought it useful to supplement the reliability data presented in the manual, which was exclusively from the United States. Since stability of a test’s scores is a crucial attribute, we assessed our test–retest group on two occasions separated by 2 to 3 weeks, a period of sufficient length to avoid boredom by repeated testing and yet brief enough to avoid maturational changes.

In sum, the aim of this study was to determine the interrater agreement and test–retest stability for Age Band 1 of the Movement ABC. Two Chinese assessors were involved, both with Cantonese as their first language. One was a psychologist, generally experienced in assessing children; the other was a final year occupational therapy student with some experience of children with movement difficulties. Neither had used the Movement ABC before.

Method

Participants

The 138 children participating in this study formed part of a larger cohort of 255 children, selected to be representative of the Hong Kong population as a whole (see Chow et al., 2001 for further details). Parental permission was obtained for all children in the study after ethics committee approval had been granted by the university.

The children in this study attended eight preschools distributed throughout the three main regions in Hong Kong (Hong Kong Island, Kowloon, and the New Territories). Within each school, children were randomly selected from class lists provided by the head teachers, with the proviso that there be roughly equal numbers of boys and girls at each age. For the interrater component of the study, the sample consisted of 79 children: 27 4-year-olds (12M, 15F), 26 five-year-olds (13M, 13F), and 26 six-year-olds (13M, 13F). For the test–retest component, another 75 children were selected: 26 four-year-olds (13M, 13F), 25 five-year-olds (13M, 12F), and 24 six-year-olds (12M, 12F). Sixteen children took part in both parts of the study. Any child with a known medical condition or severe emotional or behavioral problem was excluded prior to sample selection.

Materials

Table 1 provides a brief description of the 10 test items on the Movement ABC and the units of measure for education. Details of the practice given, the number of formal trials administered, and the scoring of items can be found in the manual (Henderson & Sugden, 1992).

Procedure

As suggested in the manual, the trainee testers were first briefed about the purpose and background of the test before carrying out a series of practice assessments on a small number of children who were not part of the study. Videotapes of the practice sessions were then viewed with the first author and any problems of administration and interpretation clarified. Children in the main study were tested individually in the familiar surroundings of their own schools, following the manual instructions precisely. Depending on the age of the child, test duration ranged from 20 to 35 minutes.

For the interrater component, both testers were present in the room and scored each child’s performance simultaneously but independently. Half of the children at each age were tested by one tester while the other observed. The testers then reversed roles. In order to assess test–retest reliability, the same tester (the psychologist), tested all children twice in the same setting.

Testing sessions were a minimum of 2 weeks and a maximum of 3 weeks apart.

Scoring and Data Analysis

Performance on the Movement ABC test can be scored in various ways. Raw scores, such as the number of seconds taken to complete a task or the number of catches made, are always noted. These scores are then converted into scaled scores, which range from 0 to 5 for each item, with higher scores indicating poorer performance. In the two items involving both the preferred and nonpreferred limbs, the scale scores of each limb are averaged to yield one score. A total impairment score (maximum 40) is obtained by summing the scores across all eight items. The manual recommends that children whose total scores fall below the 5th percentile be treated as having a definite impairment, and those falling below the 15th percentile be considered as “at risk.” Since the appropriateness of the scaled scores for a Chinese population had yet to be determined, in the present study we used raw scores to examine the reliability of performance at the item level. In view of the importance of examining agreement and stability of a child’s “pass or fail” classification on the test as a whole, however, we have used the U.S. norms advisedly.

Reliability estimates for the raw scores were obtained using analysis of variance-based intraclass correlations (ICCs) (Suen & Ary, 1989). Prior to the calculation of ICCs, the Bartlett-Box test for homogeneity of variance (Bray & Maxwell, 1985) was applied and in all cases, the assumptions

<p>| Table 1: Interrater Reliability (ICC 2, 1) for the Group as a Whole and for Each Age Group |</p>
<table>
<thead>
<tr>
<th>Test Items</th>
<th>Scores</th>
<th>All ages</th>
<th>4 years</th>
<th>5 years</th>
<th>6 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Posting coins—preferred hand</td>
<td>Number of seconds</td>
<td>.98</td>
<td>.97</td>
<td>.96</td>
<td>.98</td>
</tr>
<tr>
<td>Posting coins—nonpreferred hand</td>
<td>Number of seconds</td>
<td>.97</td>
<td>.99</td>
<td>.90</td>
<td>.97</td>
</tr>
<tr>
<td>Threading beads</td>
<td>Number of seconds</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Bicycle trail</td>
<td>Number of errors</td>
<td>.94</td>
<td>.89</td>
<td>.91</td>
<td>1.00</td>
</tr>
<tr>
<td>Catching bean bag</td>
<td>Number of successful catches</td>
<td>.98</td>
<td>.96</td>
<td>.99</td>
<td>.99</td>
</tr>
<tr>
<td>Rolling ball</td>
<td>Number of successful goals</td>
<td>.98</td>
<td>.98</td>
<td>.96</td>
<td>.99</td>
</tr>
<tr>
<td>Balance on preferred leg</td>
<td>Number of seconds</td>
<td>.98</td>
<td>.99</td>
<td>.93</td>
<td>.98</td>
</tr>
<tr>
<td>Balance on nonpreferred leg</td>
<td>Number of seconds</td>
<td>.98</td>
<td>.97</td>
<td>.98</td>
<td>.98</td>
</tr>
<tr>
<td>Jumping over cord</td>
<td>Whether passed on 1st, 2nd, or 3rd trial</td>
<td>.99</td>
<td>.98</td>
<td>.99</td>
<td>1.00</td>
</tr>
<tr>
<td>Walking heels raised</td>
<td>Number of correct steps</td>
<td>.80</td>
<td>.79</td>
<td>.74</td>
<td>.88</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>.96</td>
<td>.95</td>
<td>.94</td>
<td>.98</td>
</tr>
</tbody>
</table>

All items significant at $p < .001.$

$n = 27$

$n = 26$

$n = 26$
were met. An ICC of .75 and above was considered good, those between .75 to .50, moderate, and those below .50, poor (Portney & Watkins, 1993). To detect practice effects between test and retest, we compared the two values for each age group using matched-pairs t tests, with Bonferroni corrections for multiple tests. The stability of the pass or fail criteria was determined by subjecting the total impairment scores to the chi-square analysis.

Results
Generalizability of the Results
In order to ensure that the two samples of children in this study were comparable we performed a multivariate analysis of variance on all of the items of the test obtained on first testing only. This revealed no difference between the groups (F = 1.45, p = 0.17).

Interrater Reliability of Raw Scores
Table 1 shows the ICCs for each age and for the entire sample. The agreement between raters was very high, with an overall mean ICC of 0.96. For individual items, ICCs ranged from 0.74 for the 5-year-olds walking along a line with heels raised to 1.00 for the timed task of threading beads (all significant at p < 0.001).

Stability of Test Scores Over Time
Paired t tests showed a statistically significant improvement on retest performance only on the Threading Beads item (t = 3.30, p < .001). ICCs for these scores are also shown in Table 2. Averaged over all items and ages, the mean ICC was 0.77 (p < 0.001). According to Portney and Watkins’ (1993) guidelines, these ICCs could be considered moderate or good for all items–ages, except for the Rolling Ball item, at age 6.

Agreement on Stability of Pass or Fail Criteria
The total scores obtained by the children in this study, ranged from 0 (no impairment) to 29 (severe impairment). Of the 79 children observed by two independent raters, application of the U.S. norms revealed that 73 (92.4%) obtained scores above the 5th percentile and 6 children below. There was perfect agreement between the testers on this classification. Of the 75 children in the test–retest component, 69 “passed” on both occasions. The remaining 6 “failed” on both occasions.

Discussion
In the interrater component of our study, a psychologist and a trainee occupational therapist, both Chinese-speaking, explained and demonstrated the tasks to the children. If they differed from their American counterparts in their use of language or style of presentation, these differences were not reflected in the level of agreement reached on the children’s scores. Only one item (Walking with Heels Raised) fell short of Portney and Watkins’ (1993) criterion for reliability to be described as good rather than merely moderate, and this was only the case for the 5-year-olds. Some children walk very rapidly along the line making it difficult to count the number of successful steps. The authors of the test have previously encountered this problem when employing relatively inexperienced testers (Chow, Chan, Chan, & Lau, 2002).

Probably due to the presence of more sources of error variance, test–retest reliability usually yields lower reliability coefficients than the interrater paradigm. Although the test–retest values obtained in this study were moderate to good, they were lower, overall, than those reported in the manual or in the recent study of the Movement ABC by Croce, Horvat, and McCarthy (2001). The latter reported ICCs ranging from 0.92 to 0.98 for test–retest reliability over a 1-week period for American children ages 5 and 6 (similar values obtained for all age bands). One explanation of the lower test–retest reliabil-

---

Table 2: Mean (SD) and ICC (1, 1) Values of Test and Retest Reliability

<table>
<thead>
<tr>
<th>Test 1 mean (SD)</th>
<th>Test 2 mean (SD)</th>
<th>ICC Test 1 mean (SD)</th>
<th>ICC Test 2 mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Posting coins—</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>preferred hand</td>
<td>17.95 (4.2)</td>
<td>17.72 (4.0)</td>
<td>0.78 (0.81)</td>
</tr>
<tr>
<td>nonpreferred hand</td>
<td>20.65 (4.2)</td>
<td>20.52 (4.0)</td>
<td>0.77 (0.81)</td>
</tr>
<tr>
<td>Threading beads</td>
<td>44.72 (13.51)</td>
<td>41.39 (13.51)</td>
<td>0.83 (0.81)</td>
</tr>
<tr>
<td>Bicycle trail</td>
<td>0.72 (0.94)</td>
<td>0.67 (0.94)</td>
<td>0.86 (0.94)</td>
</tr>
<tr>
<td>Catching bean bag</td>
<td>7.33 (1.96)</td>
<td>7.53 (1.96)</td>
<td>0.74 (1.96)</td>
</tr>
<tr>
<td>Balance on preferred leg</td>
<td>15.11 (6.22)</td>
<td>14.92 (6.22)</td>
<td>0.81 (6.22)</td>
</tr>
<tr>
<td>Balance on nonpreferred leg</td>
<td>13.49 (6.45)</td>
<td>15.11 (6.45)</td>
<td>0.69 (6.45)</td>
</tr>
<tr>
<td>Jumping over cord</td>
<td>0.33 (0.76)</td>
<td>0.40 (0.76)</td>
<td>0.84 (0.76)</td>
</tr>
<tr>
<td>Walking heels raised</td>
<td>14.45 (3.0)</td>
<td>14.33 (3.0)</td>
<td>0.78 (3.0)</td>
</tr>
</tbody>
</table>

*p value significant at p < .05; *ICC values significant at p < .01, all other ICC values significant at p < 0.001

September/October 2003, Volume 57, Number 5
ity obtained here might be cultural. We had the impression that the children in our sample were more variable in their response to being assessed twice. On one hand, the culture in Hong Kong is one that encourages achievement motivation from an early age. We know that some children did try to practice the tasks between test and retest. On the other hand, there were children for whom motor competence was not highly regarded. Some even seemed to resent the time spent on these activities.

Regardless of cross-cultural implications, the present sample is much larger than in other published reliability studies of the Movement ABC. It therefore adds considerably to the findings of studies conducted in the West (e.g., Croce et al., 2001; Riggen, Ulrich, & Ozmun, 1990). One aspect, however, requires comment. Our sample was drawn from preschools serving typically developing children. Since the proportion of children failing the test was consistent with those obtained in other studies (e.g., Riggen et al., 1990; see Barnett & Henderson, 1998 for a review), we concluded that preset pass or fail criteria were fairly valid and reliable across cultures. On the other hand, however, the number of children with movement difficulties in our studies was relatively small, in absolute terms. None obtained scores at the extreme end of the severity scale (> 30), and none exhibited the kinds of associated problems commonly encountered in a clinical setting. Consequently, there is still a need to assess the reliability of scores at the extremely impaired end of the distribution, by undertaking a specially designed study in which children with a wider range of difficulties are mixed with typically developing children and tested “blind” by experienced assessors.

Use of the Movement ABC is restricted to a limited number of professionals who are deemed to be qualified to use tests of this type. However, the authors have always declined to prescribe a mandatory training program. This decision appears to be vindicated by the demonstration, here, that relatively inexperienced testers can be trained to use the test quite reliably by studying the manual and testing children of widely differing ability. We have found videotapes invaluable for clarifying procedural uncertainties as well as for illustrating and resolving problems of interpretation. We recommend that all of these devices be employed when the test is being introduced to a new user.

Conclusion

Although the Movement ABC does not advocate the use of set verbal instructions as in most standardized motor tests, it was found to be reliable for young Chinese children when tested over time and by novice testers. ▲

Acknowledgments

This study was partly funded by the Hong Kong Polytechnic University. We are grateful to all the participating preschools and children. We thank Professor Leslie Henderson for comments on the manuscript.

References


