Research Report

Interrater Reliability of the Movement Assessment Battery for Children

Bouwien CM Smits-Engelsman, Marlene J Fiers, Sheila E Henderson, Leslie Henderson

Background and Purpose

The Movement Assessment Battery for Children (M-ABC) is a widely used, standardized assessment of motor performance in children. The total score obtained on this test often is used to identify children who are either definitely impaired or at risk for motor impairment. The purpose of this study was to determine the interrater reliability of data for the M-ABC when scored by pediatric physical therapists working in routine clinical settings.

Subjects and Methods

For 9 children who were referred to clinical settings for an assessment of possible movement difficulties, performance on the appropriate age band of the M-ABC was videotaped. The 9 children, one at each age from 4 through 12 years, represented all ages covered by the test. The videotaped performances were rated according to the test instructions by 131 pediatric physical therapists with a range of experience and by an expert rater who developed the Dutch version of the test.

Results

The average agreement between therapists in their classification of the children was very high. The kappa coefficients for the 9 videos ranged from .95 to 1.00.

Discussion and Conclusion

Errors made by the therapists could be classified as those that might be common to all tests and those that are specific to the M-ABC.

BCM Smits-Engelsman, PhD, is Professor, Motor Control Lab, Division of Motor Control and Neuroplasticity, Department of Biomedical Kinesiology, Katholieke Universiteit Leuven, Leuven, Belgium; Professor, Department of Rehabilitation Medicine, Radboud University of Nijmegen, Nijmegen, the Netherlands; and Head, School of Pediatric Physical Therapy, Avans+, University for Professionals, Breda, the Netherlands. Address all correspondence to Dr Smits-Engelsman at: bouwiensmits@hotmail.com.

MJ Fiers, MPPT, is a practitioner of pediatric physical therapy in Veldhoven, the Netherlands. She was a master's student at Avans+, University for Professionals, at the time of data collection for this study.

SE Henderson, PhD, is Professor, School of Psychology and Human Development, Institute of Education, University of London, London, United Kingdom.

L Henderson, PhD, is Professor, School of Psychology, University of Hertfordshire and Imperial College Medical School, London, United Kingdom.

[Smits-Engelsman BCM, Fiers MJ, Henderson SE, Henderson L. Interrater reliability of the Movement Assessment Battery for Children. *Phys Ther*. 2008;88:286–294.]

The Movement Assessment Battery for Children (M-ABC)[1] contains 3 components: a standardized performance test, a teacher checklist, and a set of guidelines for intervention. Both the test and the checklist focus on the identification and description of impairments of motor function in children and complement each other in the way in which information is gathered. Whereas the test involves the child directly, the checklist requires an adult to rate the child's motor competence. Over the last few decades, the standardized performance test has become one of the most widely used instruments for the detection of mild to moderate movement difficulties in children.[2] When such relatively mild difficulties cannot be attributed to a known medical condition and are not simply part of a general pattern of retardation, the motor deficit often is described as "unexpected." Such difficulties recently were given a clearer and more explicit focus by the formal postulation of a developmental disorder now commonly known as "developmental coordination disorder" (DCD).[3] Although the original concept of the M-ABC precedes that of the diagnostic postulate DCD by a quarter of a century and its application is by no means limited to DCD, recognition of the existence of a developmental condition of concern to both the American Psychiatric Association and the World Health Organization[4] has fostered a growing need for the early identification of such movement difficulties.

At present, there is much debate about the best way to operationalize the criteria of the *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV*[5] for a diagnosis of DCD.[2,6,7] In particular, the formulation of criterion A, the primary inclusion criterion, is of interest to occupational therapists and physical therapists. Criterion A states: "Performance in daily activities that require motor coordination is substantially below that expected, given the person's chronological age and measured intelligence."[5(p147)] Because the M-ABC[1] contains a standardized assessment that furnishes an indication of the extent to which a child falls behind his age peers, providing norms for children aged 4 through 12 years, it seems well suited to a role in the operationalization of criterion A. In addition, the test is used for children with a range of other conditions, some physical, such as benign joint hypermobility syndrome[8] or congenital hypothyroidism,[9] and some essentially psychological, such as attention deficit-hyperactivity disorder[10] or Asperger syndrome.[11]

First described in 1972 as the Test of Motor Impairment,[12] the M-ABC[1] has been revised and restandardized on a regular basis. Originally a British-American instrument, the test has been translated into 6 European languages and 2 Asian languages. Of these, the translation into Dutch by Smits-Engelsman,[13] accompanied by a restandardization of the test instrument for the Dutch population, has had the greatest influence on local practice. It has been used extensively in the Netherlands by both occupational therapists and physical therapists. In the primary health care system, for example, Dutch pediatric physical therapists have judged it to be the most frequently used measure for both screening and descriptive purposes.[14–17]

The M-ABC was conceived as a test of motor impairment. Accordingly, some lack of resolution in the upper bandwidth of performance is tolerable, whereas uncertainties about performance as critical bands of impairment are approached are highly undesirable. These bands are the products of cutoff points at the 5th and 15th percentiles, 0 through 5th indicating children with frank abnormalities and 5th through 15th indicating children at risk, who require careful monitoring. In clinical practice, the results of this classification are used, along with other data, to make decisions about the allocation of resources. The reliability of such a classification, therefore, is very important.

Various estimates of the reliability of data for the M-ABC and its predecessor, the Test of Motor Impairment, are reported in the M-ABC manual.[1] Three examples of more recent studies must suffice to demonstrate the range of published estimates. These include data from the individual items of the test as well as the total score. In 2001, Croce et al[18] reported good test-retest reliability over all age bands of the test. A total of 106 children between the ages of 5 and 12 years were tested twice, 1 week apart. The kappa coefficient for all groups taken together was .95 and ranged from .92 to .98 for each age band separately.

In a more focused study, involving 138 children aged 4 through 6 years, Chow and Henderson[19] reported 3 estimates of reliability, 2 based on raw test scores and 1 based on a child's total score. Test-retest reliability was slightly lower than that reported by Croce et al[18] but nevertheless fell within the acceptable range. Intraclass correlation coefficients for the individual test items ranged from .64 to .86, with a mean of .77 for the test as a whole. For intertester reliability, intraclass correlation coefficients ranged from .80 to 1.00, with a mean of .96. Of particular interest in the study of Chow and Henderson[19] was the participation of 2 raters with quite different training, one an experienced educational psychologist and the other an inexperienced occupational therapist. When Chow and Henderson[19] divided their participants into those whose total scores fell above and

those whose total scores fell below the 15th percentile, the agreement between testers was perfect.

Although the above-mentioned studies showed that the M-ABC has, at the very least, moderate to good interrater and test-retest reliability, with the exception of the study of Chow and Henderson,[19] those studies have 2 limitations. First, they were conducted either with samples drawn from the standardization sample or with samples of children who were developing typically. Second, they focused on the stability of the item or total scores as opposed to the classification of children. Because 85% of the children with scores falling between the 100th and the 15th percentiles will, by definition, be classified in the "normal" category, for most it would take a great shift in percentile terms for their classification to change. The inherent likelihood of category change is much greater at the lower end of the distribution.

We know of only 2 studies of test-retest reliability in clinical populations. Leemrijse et al[16] drew samples from 2 special schools and a school for children who are chronically ill. They were able to test 23 children aged 6 to 8 years, all of whom had movement difficulties, 3 times in succession. The authors concluded that both the total scores and the cluster scores were reliable under these conditions and, therefore, that these scores might be used to detect change over time resulting from intervention. However, this reliability did not extend to the individual test items. Van Waelvelde et al[20] examined test-retest reliability in thirty-seven 4- to 5-year-old children chosen by their teachers as "the least motor skilled" in their class. Kappa coefficients for this study ranged from .65 to .93 for the individual items and reached .95 for the total scores.

In addition to studies that focus on populations of children with movement difficulties, there is a pressing need for studies in which the raters are "ordinary" clinicians, who lack a background in research but deliver therapy to children with movement difficulties on a daily basis. In the Netherlands, communication between therapists in research and clinical settings is exceptionally good. Accordingly, we were able to rectify most of these shortcomings in the study reported here.

The aim of this study was to assess the reliability of test results obtained by physical therapists who work in clinical settings, use the M-ABC on a regular basis, and use the results obtained to make decisions about children referred to them. All of the children involved were being assessed to determine whether intervention was recommended.

Method

Participants

Two groups of people participated in this study: the children being assessed and the raters.

Children. Nine children with movement difficulties participated in this study. The youngest child was 4 years of age, and the oldest child was 12 years of age, this being the age range for which normalized M-ABC test scores are currently available. Between these 2 children, one child represented each intervening year. The children were selected from a clinical population, having been referred to a pediatric physical therapy center in the Netherlands for assessment.

Table 1 shows the age, sex, and M-ABC scores for each child, along with some brief notes to characterize the range of children in the sample. Three children clearly met the criteria for DCD. A further 3 children had nonmotor problems that had already led to another diagnosis, namely, attention deficit-hyperactivity disorder, pervasive developmental disorder (not otherwise specified), and learning disability. Of the remaining 3 children, one child had hemophilia, a condition sometimes associated with poor motor performance.[21] Another child's score fell in the range described by the test authors as "at risk," and the last child's score fell on the borderline separating the normal category from the at-risk category (ie, on the 16th percentile). All parents gave informed consent, and the children assented to their performance being videotaped, with the recordings to be used for either research or educational purposes.

Raters. The 131 physical therapists who participated in this study were approached initially by telephone, e-mail, or advertisement. The group comprised 120 women and 11 men with a mean age of 41 years (range=23–59 years). Sixty-five percent had more than 10 years of clinical experience, and 15% had less than 4 years of clinical experience. Sixty-six percent had obtained compulsory postgraduate qualifications permitting the use of the title "Chartered Pediatric Physiotherapist."

The therapists were, in general, well acquainted with the M-ABC. Sixty-three percent administered the test between 1 and 4 times per month. A further 30% used the test more frequently. Most of the therapists (66%) had attended a short training session on the M-ABC (between a half-day and a full day), and 29% had taught themselves by using the manual or by working with a colleague over the learning period.

M-ABC

The Dutch translation of the M-ABC[13] was used in this study. The test accommodates 4 age bands from 4 to 12 years of age, with each age

Table 1.
Characteristics of the 9 Children Participating in the Study[a]

Age (y)	Sex	Mean TIS Assigned by All Raters	TIS Assigned by Expert	M-ABC Classification	Some Reasons for Referral and Comments on Associated Difficulties
4	M	15	15	At risk	General hypotonia, lack of muscle power
5	M	17	19	Definitely impaired	Motor coordination problems, hemophilia, poor concentration
6	F	12	11	At risk	Hypotonia, poor balance, PDD.NOS (autistic spectrum disorder)
7	M	8	9	Normal	Motor coordination problems, very nervous, insecure
8	F	12	11.5	At risk	Motor coordination problems
9	M	25.5	27.5	Definitely impaired	DCD, ADHD
10	M	18.5	20.5	Definitely impaired	Very slow in motor tasks, learning disability (normal IQ)
11	F	17	19	Definitely impaired	DCD, nonverbal learning disorder
12+	M	24	25	Definitely impaired	DCD

[a] TIS=total (motor) impairment score, M-ABC=Movement Assessment Battery for Children, M=male, F=female, PDD.NOS=pervasive developmental disorder (not otherwise specified), DCD=developmental coordination disorder, ADHD=attention deficit-hyperactivity disorder, IQ=intelligence quotient.

band covering the same types of skill. There are 8 items within each age band, divided into 3 major performance areas: manipulative skills (items 1, 2, and 3), ball skills (items 4 and 5), and balance skills (items 6, 7, and 8). Children can score between 0 and 5 on each of the 8 items. The total (motor) impairment score (TIS) is then calculated by summation of the 8 item scores, resulting in a range of scores between 0 and 40, with lower scores indicating better performance. This total is then interpreted by comparison with the norm tables, in which a TIS below the 5th percentile is considered to be indicative of a definite motor problem. Scores between the 5th and 15th percentiles suggest a degree of difficulty that is borderline and requires monitoring. Scores above the 15th percentile are deemed to indicate normality. Evidence of the validity of the M-ABC test is presented in the manual[1] and by Barnett and Henderson.[22] Three studies[18,23,24] provide data on the relationship between the M-ABC and the Bruininks-Oseretsky test, an instrument commonly used in the United States.

Procedure

All 9 children were tested by the second author (MJF) in a quiet room with one or both parents or guardians present. These sessions were videotaped by a technician, positioned so that as much of the child's performance as possible could be seen without distracting the child or the tester.

All video ratings took place over a 6-month period and took about 30 minutes per rating. Because the raters were distributed across the Netherlands, it was not possible to meet each one independently, nor was it possible to control the group size exactly. However, most of the ratings were made in small groups, ranging from a minimum of 5 therapists to a maximum of 15 therapists (median=7). In 65 cases, the group viewed one videotape, and in 66 cases, the group viewed 2 videotapes. The videotapes were selected to ensure that they were viewed approximately the same number of times and that, when 2 videotapes were being rated, each child was being assessed on a different age band of the test. In summary, the number of raters per videotape ranged from 16 to 34, yielding a total of 197 ratings (Tab. 2).

Once the group had been introduced to the organizer and the aim of the study had been explained, the therapists were told that the children to be seen on the videotapes had been referred for an assessment of their motor performance and that their task was to score each child's M-ABC performance as they would in their daily practice. They were shown the videotape item by item, with a short stop between the items, during which they entered their scores on forms. In the vast majority of cases, each item was viewed only once, as it would be in a real-life setting. Only if the therapists were not able to see an item properly was a fragment repeated. Because some of the manual

Table 2.
Number of Raters per Videotape, Differences in Total Impairment Scores (TISs), and Multiple Kappa Coefficients

Videotape, as Denoted by Child's Age	No. of Raters per Videotape	Mean (SD) Difference in Scores[a]	Kappa[b]
4	25	1.7 (1.5)	1.00
5	34	2.3 (1.6)	.98
6	23	1.1 (1.3)	1.00
7	24	2.9 (2.3)	.95
8	16	1.1 (0.7)	1.00
9	18	1.7 (1.4)	.99
10	18	1.6 (1.2)	.99
11	20	1.7 (1.2)	.99
12	19	2.8 (1.8)	.98
$\overline{X}$	22	1.9 (1.5)	.99

[a] Difference between the TIS assigned by each rater and the mean TIS for that child.
[b] All significant at $P<.01$.

dexterity items were difficult to judge on videotape because the child's body occluded his or her actions, the raters also were provided with hard copies of the product of the child's efforts (bicycle trail, flower trail, or cutout elephant, as appropriate for age) so that they could score errors as they normally would in their clinical practice.

After performance on the individual items had been viewed, subscores and total scores were calculated together with percentiles derived from tables provided after the viewing session had been completed. If a therapist participated in scoring of a second videotape, the entire procedure was repeated. All observations and measurements were made independently, and discussion among the therapists within a group was not permitted while they were viewing the videotapes. In this study, all 4 age bands of the M-ABC were used. The participating therapists did not know in advance which age bands would be shown to them.

The author of the Dutch version of the M-ABC (BCMSE) also scored the videotapes of the 9 children independently, thus providing an expert opinion on the scoring of all 32 items of the test.

Data Analysis

For each of the 131 therapists involved, the TIS assigned to each child and the classification of each child in the normal, at-risk, or "definitely impaired" category, according to the guidelines given in the test manual, formed the basis of the data analysis. In order to measure the degree of agreement among testers, multiple kappa statistics designed for use with ordinal responses were computed[25] with SAS 8.1 software.* In addition, to provide a measure of the confidence with which the M-ABC test scores could be interpreted, we computed: (1) the standard error of measurement (SEM); (2) the smallest detectable difference (SDD), calculated with the formula $1.96\times\sqrt{2}\times$SEM of TIS; and (3) the 95% confidence interval of the scores for each videotape (child) being assessed. These statistical analyses were followed by a detailed assessment of the types of errors made during the scoring of individual items of the test.

* SAS Institute Inc, PO Box 8000, Cary, NC 27513.

Results

Agreement Among Raters

As a starting point for our analyses, we considered, for each videotape (child), the total number of raters who observed that child. Agreement or disagreement was expressed in terms of the difference between the TIS assigned to each videotape by each rater and the mean TIS for that child on that particular videotape. Table 2 summarizes these data by showing the number of raters per videotape, the mean difference in the TIS assigned to each child by each rater, and the mean TIS for the child on that particular videotape, along with the kappa coefficients derived from these data. Averaged over all 9 videotapes (and 197 ratings), the mean difference among raters was only 1.9 points. The multiple kappa coefficients for the classification of the children in the normal, at-risk, or definitely impaired category fell in the range generally described by psychometricians as "excellent," the lowest value being .95.

Table 3.
Total Impairment Score (TIS) by Raters per Videotape, Standard Error of Measurement (SEM), Smallest Detectable Difference (SDD), and Limits of Agreement of the TIS[a]

Videotape, as Denoted by Child's Age	Mean (SD) TIS	SEM	SDD	95% Limits of Agreement	
				Upper	Lower
4	15.2 (2.3)	0.5	1.3	16.5	14.0
5	17.1 (2.8)	0.5	1.3	18.4	15.8
6	11.8 (1.8)	0.4	1.0	12.8	10.7
7	8.1 (3.7)	0.8	2.1	10.2	6.0
8	11.7 (1.3)	0.3	0.9	12.6	10.8
9	25.5 (2.3)	0.5	1.5	27.0	24.0
10	18.5 (2.0)	0.5	1.3	19.8	17.2
11	17.4 (2.1)	0.5	1.3	18.7	16.1
12	23.7 (3.4)	0.8	2.2	25.9	21.5
$\bar{X}$	16.3 (5.7)	0.4	1.1	17.4	15.1

[a] Classification cutoff scores for 6- to 12+-year-old children: TIS of 10.0=15th percentile, TIS of 13.5=5th percentile.

Table 3 shows the mean TIS assigned by the raters observing each videotape, the SEM associated with that score, the SDD, and the confidence interval or limits of agreement for each. Overall, these results inspired confidence in the test. Averaged over all 9 videotapes, the SDD was 1.1, ranging from 0.9 to 2.2 points. Examination of the 95% confidence interval of the scores for each videotape (child) revealed only one instance in which the degree of uncertainty was problematic. The child in question was the 7-year-old boy, who was observed by 24 therapists. This child was assigned a mean TIS of 8.1 by the therapists, but the agreement among them was slightly lower than that for the other 8 children (kappa coefficient=.95). A score of 8 would place him in the normal category, according to the test norms. However, if the limits of agreement for this age group are considered, then the range of scores associated with his TIS would be ±2, that is, between 6 and 10; the latter value would place him in the at-risk category.

Further investigation of the 197 observations revealed complete agreement on 85% of occasions. In 28 cases, a child was placed in a category one higher or lower than the expert's rating. In only 2 cases (1%) was a child rated as 2 categories removed from the expert's rating. We examined these cases very carefully, because a discrepancy of this size could have had serious practical consequences. Both misclassifications were caused by one misinterpretation, and both occurred during rating of the videotape of the 7-year-old boy. One misclassification was caused by the therapist giving the maximum impairment score (5) on a bimanual task, because the therapist considered the child to have made a procedural fault. In contrast, the expert considered the child's performance acceptable, recorded the time taken to complete the trial, and gave the child a score of 0. The second misclassification was caused by the "jumping-in-squares" item, in which a discrepancy of one jump resulted in a 3-point difference in the TIS.

Agreement by Videotape

Another way of interpreting the data was to determine whether there were systematic differences in ratings over the 9 videotapes that might be attributable to either the videotape or the child. As noted above, the children were correctly classified 85% of the time. More detailed analysis of agreement as a function of age showed perfect agreement on the videotapes of the children aged 9, 10, and 12 years. For those aged 8 and 11 years, only one therapist disagreed with the others. For the younger children, aged 4 through 7 years, there was slightly more disagreement, with the videotape of the 7-year-old boy generating the most disagreement (6 of the 24 raters).

Sources of Disagreement Among Raters: Analysis of Error Types

In order to determine the sources of disagreement among therapists or videotapes, all of the differences between the item scores of the therapists and those of the expert were examined individually. Almost all of the errors could be categorized into the following 3 groups.

Missing a procedural fault. The M-ABC manual specifies faults of procedure that result in a child "failing" a task. For example, in catching a ball, a child is not permitted merely to trap it against his or her clothing, and this action was missed by the observers.

Differences in timing or number of errors between rater and expert. On a very few occasions, the rater and the expert recorded different times or missed an error in a child's performance on a task. In most cases, a small difference (of 1 second or one error) would not affect a child's scaled score. However, in other cases, it could lead to a change in the scaled score. This situation occurred for the videotapes of the 5-, 6-, and 7-year-old children.

Errors in completing the form. On a few occasions, testers used the wrong age column (in the correct age band) to calculate the scaled score, failed to calculate a mean value for items involving both limbs, or failed to use the best performance to transform a raw score to a scaled score. Curiously, the most frequently made error was caused by misreading the digital stopwatch. When the display showed 1.06, some therapists recorded this value on the form as 106 seconds rather than the true 66 seconds.

Differences Among Therapists

In order to determine whether there was any systematic association between the number of errors, the type of errors, or both made by the therapists and their clinical experience, we compared the characteristics of all those whose classifications agreed with those of the expert examiner and those whose classifications differed. We found no significant differences between these 2 groups in age, basic training, work experience, number of times they used the test, or how they were trained on the test.

Discussion

Valid and reliable motor tests are essential tools used by clinicians to diagnose and evaluate motor performance in children with developmental disorders. In this study, we compared the extent to which "ordinary" therapists agreed with each other and with a highly experienced expert who had translated the M-ABC test into Dutch and collected local norms. We found that agreement in classifications was very high, with interrater reliability values ranging from .95 to 1.00.

Reliability and Classification

In clinical settings, it is common practice to use the M-ABC as a means of classifying children into 3 groups: normal, at risk, and definitely impaired. These data then are used in conjunction with clinical judgment to decide whether and how a child will receive intervention. Fundamental to any test used in this way is evidence of the reliability with which such decisions can be made. Previous studies showed that both the interrater reliability and the test-retest reliability of data for the M-ABC are good, provided that the whole range of scores is examined. However, there is a severe shortage of data on the stability of scores falling around the predetermined cutoff points indicating the likelihood of impairment. Although Chow et al[26] were able to show that 2 occupational therapists were in perfect agreement with each other on 2 occasions of testing, unfortunately, they restricted their attention to the 15th percentile as a means of dichotomizing children into impaired and not-impaired categories.

Distinctive Features of the Present Study

Features that merit special attention include the following. We included an unusually large and diverse sample of physical therapists as raters. To our knowledge, all previous reliability studies of tests of motor impairment that included therapists as raters were confined to judgments made by occupational therapists. Given that the knowledge base and training of these professionals are very different from those of physical therapists, the present investigation extended the findings of previous studies considerably. By focusing entirely on children with movement difficulties, we required raters to make judgments about performance that were more difficult than if the child being observed had no difficulties (eg, deciding whether a series of hops is continuous can be difficult if a child moves rather slowly and with a heavy footstep). Finally, previous studies tended to focus on younger children, whereas the children in the present study represented every age group covered by the test.

Video Recording and Sources of Error

Close inspection of the types of errors made in judging and scoring performance revealed some interesting findings. Some of these seemed likely to occur in almost any test. Observing and scoring any performance test from a videotape is different from observing a child in daily practice. When therapists examine a child in a clinical setting, they are free to move around and take the best possible viewpoint. Such positioning is not possible when a videotape is used, and this constraint certainly accounted for some of the errors made in the present study. Although videotapes of children being assessed with the M-ABC are invaluable for training purposes and are useful in clinical settings, it should be remembered that there is a cost for using videotapes, too. In the present study, the intermittent occlusion of a child's continuous tracing activity caused by a single viewpoint provides a simple example of a source of disagreement directly at-

tributable to the limitations of video recording.

In addition to the error types that might occur with any test involving the observation of movement, there were a few that seemed to be specific to the M-ABC. Sometimes, the source could be traced to an item description in the manual that was not quite clear; at other times, raters seemed to find the layout of the form confusing (the way in which a score for an item involving both limbs is derived needs clarification). Some of our participants reported that they worked in settings that limited the age range of the children that they assessed. Thus, those who worked primarily with younger children might not have used the older age bands of the test for some time. We did not tell the raters in advance which video they would see. With hindsight, this choice might have been a mistake, as conscientious testers who have not tested a child of a particular age for some time might well prepare themselves by reviewing the test manual and form for that age group.

Differences in Classification

There were very few instances in which therapists disagreed on their classification of a child and, when they did, it was usually possible to locate the source of the error quite easily. Nevertheless, one important lesson can be learned from the present study. With any test that classifies people around specified points, there is always a risk of making more errors close to the cutoff points. It is clear that it is important for clinicians to be acquainted with the factors that lead to misclassification, as the effect of misclassification on a child's future can be considerable. In addition, taking the SDD and the limits of agreement (confidence intervals) into account always helps clinicians judge the stability of a test score. In the present study, we found these values to be in the acceptable range for all age bands of the M-ABC.

There was some variation in the reliability coefficients obtained for the 9 videotapes used in this study. We found no evidence to suggest that this variation was the result of differences in the test items for different age groups of children. Rather, it seemed to be attributable to differences among the particular children chosen to represent the age groups. It is important to note, therefore, that different interrater reliability coefficients and different SDDs could be obtained for another 9 children of comparable ages.

Conclusion and Recommendations

In this study, we showed that the M-ABC provides a robust means of classifying children into the normal, at-risk, or definitely impaired category. The errors made, although numerous, were relatively minor and rarely resulted in major errors in the scoring of an individual item. However, users do need to be aware that the current manual does not provide confidence intervals around the specified cutoff points. Although our participants generally reported that they found the test manual and form clear, the present study revealed some areas in which improvements in wording, layout, or both could reduce errors. Other errors seemed to be more general and attributable to a lack of concentration or perhaps a lack of knowledge of the test instructions. In either case, these sources of errors could be addressed in courses that introduce the M-ABC to new users or in continuing professional development programs. Finally, it is important to remember that any decision about intervention for a child should never be based on a single test. A test is just an instrument to help a clinician make decisions and should not replace clinical judgment.

Dr Smits-Engelsman and Dr SE Henderson provided concept/idea/research design. Dr Smits-Engelsman, Dr SE Henderson, and Dr L Henderson provided writing. Ms Fiers provided data collection and subjects. Dr Smits-Engelsman provided data analysis. Dr Smits-Engelsman and Ms Fiers provided project management, fund procurement, facilities/equipment, and clerical support. All authors provided consultation (including review of manuscript before submission). The authors thank the children and all of the physical therapists, who gave so much of their time to this project, and Dr Chris Michels for his advice on psychometrics.

This article was submitted March 1, 2007, and was accepted September 18, 2007.

DOI: 10.2522/ptj.20070068

References

1 Henderson SE, Sugden DA. *Movement Assessment Battery for Children: Manual.* London, United Kingdom: The Psychological Corporation; 1992.

2 Geuze RH, Jongmans MJ, Schoemaker MM, Smits-Engelsman BCM. Clinical and research diagnostic criteria for developmental coordination disorder: a review and discussion. *Hum Mov Sci.* 2001;20:7–47.

3 *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV-TR.* 4th ed, text rev. Washington, DC: American Psychiatric Association; 2000.

4 *ICD-10: International Statistical Classification of Diseases and Related Health Problems*. Rev 10th ed. Geneva, Switzerland: World Health Organization; 1994.

5 *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV*. 4th ed. Washington, DC: American Psychiatric Association; 1994.

6 Henderson S, Henderson L. Toward an understanding of developmental coordination disorder. *Adapted Physical Activity Quarterly.* 2002;19:12–31.

7 Sugden DA, ed. *Leeds Consensus Statement: ESRC Seminar Research Series*. Cardiff, United Kingdom: Discovery Trust; 2006.

8 Kirby A, Davies R. Developmental coordination disorder and joint hypermobility syndrome: overlapping disorders? Implications for research and clinical practice. *Child Care Health Dev.* 2007;33:513–519.

9 Kooistra L, Laane C, Vulsma T, et al. Motor and cognitive development in children with congenital hypothyroidism: a long-term evaluation of the effects of neonatal treatment. *J Pediatr.* 1994;124:903–909.

10 Pitcher TM, Piek JP, Hay DA. Fine and gross motor ability in males with ADHD. *Dev Med Child Neurol.* 2003;45:525–535.

11 Green D, Baird G, Barnett AL, et al. The severity and nature of motor impairment in Asperger's syndrome: a comparison with specific developmental disorder of motor function. *J Child Psychol Psychiatry.* 2002;43:655–668.

12 Stott DH, Moyes FA, Henderson SE. *The Henderson Revision of the Test of Motor Impairment.* San Antonio, Tex: The Psychological Corporation; 1984.

13 Smits-Engelsman BCM. *Movement Assessment Battery for Children: Handleiding.* Lisse, the Netherlands: Swets & Zeitlinger; 1998.

14 Leemrijse C, Meijer OG, Vermeer A, et al. Detecting individual change in children with mild to moderate motor impairment: the standard error of measurement of the Movement ABC. *Clin Rehabil.* 1999;13:420–429.

15 Leemrijse C, Meijer OG, Vermeer A, et al. The efficacy of Le Bon Départ and Sensory Integration treatment for children with developmental coordination disorder: a randomized study with six single cases. *Clin Rehabil.* 2000;14:247–259.

16 Schoemaker MM, Niemeijer AS, Reynders K, Smits-Engelsman BC. Effectiveness of neuromotor task training for children with developmental coordination disorder: a pilot study. *Neural Plasticity.* 2003;10:155–163.

17 Niemeijer AS, Schoemaker MM, Smits-Engelsman BC. Are teaching principles associated with improved motor performance in children with developmental coordination disorder? A pilot study. *Phys Ther.* 2006;86:1221–1230.

18 Croce RV, Horvat M, McCarthy E. Reliability and concurrent validity of the Movement Assessment Battery for Children. *Percept Mot Skills.* 2001;93:275–280.

19 Chow SMK, Henderson SE. Interrater and test-retest reliability of the Movement Assessment Battery for Chinese preschool children. *Am J Occup Ther.* 2003;57:574–577.

20 Van Waelvelde H, Peersman W, Lenoir M, Smits-Engelsman BCM. The Movement Assessment Battery for Children: educational differences between the results of 4- and 5-year-old children from Flanders, the Netherlands, and the USA. *Adapted Physical Activity Quarterly.* 2007;24:59–69.

21 Schoenmakers MACG, Gulmans VAM, Helders PJM, et al. Motor performance and disability in Dutch children with haemophilia: a comparison with their healthy peers. *Haemophilia.* 2001;7:293–299.

22 Barnett A, Henderson SE. *An Annotated Bibliography of Published Studies Employing the TOMI/Movement ABC 1984–1996.* London, United Kingdom: The Psychological Corporation; 1998.

23 Smits-Engelsman BCM, Henderson SE, Michels CGJ. The assessment of children with developmental coordination disorders in the Netherlands: the relationship between the Movement Assessment Battery for Children and the Körperkoördinations Test für Kinder. *Hum Mov Sci.* 1998;17:699–709.

24 Van Waelvelde H, De Weerdt W, De Cock P, Smits-Engelsman BCM. Aspects of concurrent validity of the Movement Assessment Battery for Children. *Hum Mov Sci.* 2004;23:49–60.

25 Fleiss JL. *Statistical Methods for Rates and Proportions.* 3rd ed. New York, NY: John Wiley & Sons Inc; 2003.

26 Chow S, Chan LL, Chan CPS, Lau CHY. Reliability of the experimental version of the Movement ABC. *British Journal of Therapy and Rehabilitation.* 2002;9:404–407.