Is the Movement Assessment Battery for Children-2nd edition a reliable instrument to measure motor...

Article *in* Research in developmental disabilities · February 2011

DOI: 10.1016/j.ridd.2011.01.031 · Source: PubMed

CITATIONS
32

READS
571

3 authors:

Bouwien Smits-Engelsman
University of Cape Town
187 PUBLICATIONS **4,216** CITATIONS
SEE PROFILE

Anuschka S Niemeijer
Martini Ziekenhuis
25 PUBLICATIONS **480** CITATIONS
SEE PROFILE

Hilde Van Waelvelde
Ghent University
59 PUBLICATIONS **893** CITATIONS
SEE PROFILE

Research in Developmental Disabilities 32 (2011) 1370–1377

Contents lists available at ScienceDirect

Research in Developmental Disabilities

Is the Movement Assessment Battery for Children-2nd edition a reliable instrument to measure motor performance in 3 year old children?

Bouwien C.M. Smits-Engelsman [a,b,*], Anuschka S. Niemeijer [c], Hilde van Waelvelde [d]

[a] Avansplus, University for Professionals, Masteropleiding Kinderfysiotherapie, Heerbaan 14-40, 4817 NL Breda, The Netherlands
[b] Motor Control Laboratory, Research Center for Movement Control and Neuroplasticity, Dep. of Biomedical Kinesiology, K.U. Leuven, Gebouw De Nayer kamer 02.11, Tervuurse Vest 101, 3001 Leuven, Belgium
[c] Sint Maartenskliniek, afdeling Research Development and Education, Postbus 9011, 6500 GM Nijmegen, The Netherlands
[d] Hilde van Waelvelde, Revalidatiewetenschappen en Kinesitherapie, Arteveldehogeschool en Universiteit Gent, Campus Heymans 2B3, De Pintelaan 185, 9000 Gent, Belgium

ARTICLE INFO

Article history:
Received 14 April 2010
Received in revised form 3 January 2011
Accepted 15 January 2011
Available online 23 February 2011

Keywords:
Motor test
Movement ABC-2
Test–retest reliability
Agreement
Feasibility

ABSTRACT

Formal testing of 3 year old children is a new feature in the revised version of the Movement Assessment Battery for Children (Movement ABC-2). Our study evaluated the reliability and explored the clinical applicability of the Movement ABC-2 Test in this young age group. A total of 50 typically children were given two trials of the test within a one to two week interval by two physical therapists: same assessor ($n = 28$ children) and different assessors ($n = 22$ children). Psychometric properties were evaluated by calculating internal consistency (Cronbach α), intra-class correlation (ICC), the standard error of measurement (SEM), the smallest detectable difference (SDD) and Kappa values for classification agreement. The results are promising for future implementation of the Movement ABC-2 in clinical practice. The children's performance was highly reproducible when tested by the same assessor (ICC .94) The SEM was 1.7 or 2.1 standard scores for 90% or 95% confidence intervals respectively, making the test sensitive enough to detect individual changes. If two different assessors tested the children the ICC was .76. In conclusion, the revised test can be applied to assess motor performance in typically developing 3-year old children. Future studies are needed to confirm if the same can be said for children with motor delays.

1. Introduction

Early identification of children with mild to moderate motor impairments is crucial for many reasons. One of the most important reasons is that support for the children and their parents can begin before the child starts school (Jongmans, 2005). To identify mild motor disorders, the Movement Assessment Battery for Children (Movement ABC) is the most commonly used test in clinical practice in Europe (Geuze, Jongmans, Schoemaker, & Smits-Engelsman, 2001). The Movement ABC is a norm referenced test that requires a child to perform a series of motor tasks in a strictly specified way (Henderson & Sugden, 1992).

Recently the Movement ABC has been revised and the age range for the test extended from 4–12 year olds to 3–16 year olds (Henderson, Sugden, & Barnett, 2007). The reliability of the original Movement ABC in the age range 4-12 has been

* Corresponding author at: Motor Control Laboratory, Research Center for Movement Control and Neuroplasticity, Dep. of Biomedical Kinesiology, K.U. Leuven, De Nayer 02.11, Tervuurse Vest 101, 3001 Leuven, Belgium. Tel.: +32 16 329078; fax: +32 16 329197.
E-mail addresses: bouwiensmits@hotmail.com (Bouwien C.M. Smits-Engelsman), a.niemeijer@mzh.nl (A.S. Niemeijer), hilde.vanwaelvelde@ugent.be (H. van Waelvelde).

0891-4222/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ridd.2011.01.031

established in several studies. For example, Smits-Engelsman, Fiers, Henderson, and Henderson (2008) calculated the agreement between raters to be very high (.95–1.00) when therapists classified children of all age groups using videotapes. Croce, Horvat, and McCarthy (2001) reported test–retest reliability to be good across all age bands. They twice tested 106 children between the ages of 5 and 12, with a one week assessment interval. The Kappa coefficient for all groups taken together was .95 and ranged from .92 to .98, for each age band separately. Van Waelvelde, Peersman, Lenoir, and Smits-Engelsman (2007) established a good test–retest reliability of the total M-ABC score for 4- and 5-year-old children.

Although it was not expected that the reliability of the Movement ABC-2 would be significantly different from the original test, there have been no published studies confirming this. Importantly, no studies are available that evaluate the test–retest reliability of children in the youngest age group (3 year olds) since this age group is a new addition in the Movement ABC-2. In young children, test results may be highly dependent on the way the children are instructed and kept motivated, since they have a limited attention span. As the manual states "Establishing good rapport with a child during assessment is critical, the examiner is the one that has to make the child feel at ease and enjoy the testing experience... If the child says he or she cannot perform a test, the examiner should encourage the child... However, none of the strategies to encourage a child should be pushed too far. In those cases an R for refused can be noted. (Manual MABC-2, p. 16)." This implies there is a possibility that examiners differ in the way they instruct and motivate young children.

It could also be argued that aspects such as the ability to understand the task of younger children may lead to differences in performance between two test occasions and therefore decrease reliability of the test results. The first purpose of our study was to evaluate the clinical applicability of the Movement ABC-2 for 3-year-old children. Another important question in this study was how stable the test scores were over repeated testing. For example, if large differences were found within a relatively short period (2 week interval), the results might indicate that different capacities were being measured. In such a scenario, the first measurement might not be a valid measure of motor performance but would instead be associated with "getting acquainted with a formal test situation" or learning the meaning of terms used in the test, like "as fast, or as neat as you can".

We also evaluated the classification metric, which should be reliable, since a score $\leq$15th percentile is essential for diagnosis and assignment to intervention in children with motor delays. Finally, we investigated whether young children were highly susceptible to the way they were approached by different assessors by comparing reliability values in a test–retest design with the same assessor (intra tester reliability) and two different assessors (inter tester reliability).

2. Methods

2.1. Participants

In this study 50 healthy children aged 36–48 months were tested twice. Of these children 27 were boys (mean age 41 months; SD 2.6 months) and 23 girls (mean age 41.5 months; SD 3 months). The children were recruited from three Dutch pre-schools. The first 50 children, whose 50 parents responded to our invitation to participate, were enrolled in the study. The parents were informed about the purpose of the study and consent for all children was obtained. Of these 50 children 27 children were between 36 and 41 months old (young group) and 23 between 42 and 47 months (older group). Children who had no obvious physical, cognitive and/or sensory disabilities (such as cerebral palsy, Down syndrome, and blindness) were eligible for this study. However no children had to be excluded.

Their parents were informed about the purpose of the study and consent for all children was obtained. Parent education was representative for the Dutch population; the highest education for the father (mother) was low for 18% (14), middle or vocational for 53% (59) and 29% (27) of the parents went to university. The study received ethical approval from the Local Committee of the University of Nijmegen in the Netherlands.

2.2. Movement Assessment Battery for Children 2nd edition (Movement ABC-2)

The Movement ABC-2 (Henderson et al., 2007) is a revision of the Movement ABC (Henderson & Sugden, 1992). The revised Test and Checklist make it possible to identify and describe impairments in motor performance of children in the age range of 3–16. In this study, the Dutch translation of the test items was used (Smits-Engelsman, 2010).

The Movement ABC-2 Test consists of eight items in each of three age bands (3–6 years; 7–10 years; and 11–16 years). These items measure different aspects of motor ability which are divided into three major performance areas called components: manual dexterity (three items), aiming and catching (two items) and static and dynamic balance (three items). For all items, besides aiming and catching, two trials are given if needed and the best trial is used to rate an item. The manual provides clear guidelines for how to instruct the task and when to record a performance as 'R-refused', 'I-inappropriate', or 'M-missing'. A well-coordinated child with an average intelligence can complete the Test in 20–40 min.

For each child, the raw item scores (e.g. the number of seconds or times a task is performed well) can be transformed into item standard scores (ISS), component standard scores (CSS), a total standard score (TSS) and a total percentile score. For age band 1, used in this study, ten ISSs can be calculated because two items are performed with both sides of the body. However, a raw score to calculate an ISS can be lacking for several reasons. If a child could not perform any trial of an item as prescribed in the manual, due to procedural faults, an 'M' is recorded. For example, a procedural fault is recorded when a child picks up more than one coin at the time or changes hands while putting the coins in the box. If a procedural fault was recorded, it was

replaced by the poorest score of the reference group. For the Aiming and Catching items, the child gets 10 trials. If all attempts on the Aiming and Catching tasks are failed (including procedural faults and missed catches/throws) the child receives a zero score. For the balance items counting the seconds, steps or jumps stops when an error is made, so a zero score in these items indicates procedural faults and/or inability to perform the task. If a procedural fault was recorded on the Manual Dexterity items (number of ISS = 4) or zero catches or hits on the Aiming and Catching items (number of ISS = 2) or in case of a zero score on the balance items (number of ISS = 4), these items will be referred to in this paper as "missings scores" or "failed items". A maximum of 10 missings scores could be achieved.

The total standard score (TSS) can be used for classification purposes. To make it easy to use in clinical practice the "traffic light system" is introduced in Movement ABC-2. A TSS below or equal to 56 points places a child at or below the 5th percentile, in a red zone. Scores between 57 and 67 inclusive place a child between the 5th and 15th percentile, in an amber zone. Performance falls within the green zone, normal range, if the TSS score is above the 15th percentile, above 67. In this green zone there are no movement problems detected. Children who score $\leq$15th percentile (TSS $\leq$ 67) are classified as children with potential motor problems (at risk or impaired). In clinical practice, this $\leq$15th percentile cut off score is used to make diagnostic and intervention decisions.

Inter-observer reliability of the original version of the Movement ABC was high (1–.95) (Smits-Engelsman et al., 2008). For Movement ABC-2, 16 young children (age band 1) were tested when 2 assessors were present in the same room. They scored each child simultaneously but independently. Each assessor tested 8 children while the other person only looked from a certain distance, so s/he would not interfere with the testing procedure and rating of the items. The ICC between the two ratings of the 16 children was .96 (Smits-Engelsman, 2010).

For all children non-motor factors that might have prevented them from demonstrating their true movement capability were recorded as provided on Movement ABC-2 scoring forms.

2.3. Procedure

Each child was assessed individually in a quiet room with a Dutch translation of the second version of the Movement ABC Test (Smits-Engelsman, 2010). Within a one to two week interval the child was tested again either by the same (normal test–retest or intra-tester design) or by another assessor (inter-tester test–retest design). It is important to distinguish between intra and inter-rater reliability in scoring the test (usually based on videotaped cases) and intra/inter-tester designs for test–retest reliability which by definition not only includes the scoring but also includes the way in which the assessor coaxes the best performance from the child. The two assessors were trained as pediatric physical therapist and had more than 5 years of clinical experience. Both assessors attended a full day certified training encompassing guided practice of all items with children in the appropriate age groups and passed the video criterion test where they had to rate all the items and calculate standard scores. They could submit any queries to the author of the Dutch Movement ABC-2 manual (BSE).

Questionnaires were filled out by the parents to collect data on medical history and to gather background information (earlier medical treatment, therapies, and motor, language or behavioural problems).

2.4. Data analysis

The raw item scores were transformed into standard scores (mean = 10, SD = 3) according to the tables provided in the manual (Henderson et al., 2007).

2.4.1. Feasibility

The number of failed items or missing scores was counted. The impact of missing scores was evaluated and possible association with age and gender was investigated by comparing the number of missing scores between the two age and gender groups using independent *t*-tests. Age groups were based on the half year norms provided in the manual (3.0–3.5 and 3.6–3.11 years of age). In addition, the items that were difficult and caused children to make procedural faults were examined.

2.4.2. Internal consistency

The extent to which items are measuring the same construct was examined using Cronbach's alpha (α) coefficient. This coefficient was calculated based on 10 ISS per measurement occasion using the data from all 50 children.

2.4.3. Reliability

The statistical analyses were executed in two different ways: (1) including *all* children (n = 50) and (2) including only children that had 4 or less missing items or failed attempts (n = 42 '*clean*' dataset). To calculate test–retest reliability, the data of children tested by the same tester were used (intra-tester test–retest design; n_{all} = 28 and n_{clean} = 23). Next consequences on reliability of testing by two assessors are explored (inter-tester test–retest design; n_{all} = 22 and n_{clean} = 19).

2.4.3.1. Stability of test scores. Intra-class correlation coefficients (ICCs) were calculated with a 2-way random effect model. This model allows the results to be generalized to assessors not participating in the study (Shrout & Fleiss, 1979).

2.4.3.2. Standard error of measurement (SEM). The SEM was applied to determine the precision of the total score of the MABC-2. The SEM describes the error in interpreting an individual's test score. The SEM allows for estimation of the 'true' test performance using a reliability coefficient and is computed by the standard deviation of the total score multiplied by the square root of one minus its reliability coefficient [SEM = SD $\times \sqrt{1 - \text{ICC}}$] (Nunnally & Bernstein, 1994).

2.4.3.3. Smallest detectable difference (SDD). In clinical practice it is important to know whether test–retest differences on an individual basis are at or over one SDD level. The SDD is considered the minimal amount of change that is not likely to be due to chance variation in measurement. If the found change is larger than the SDD it may be called a real change. SDD can also be used to determine the proportion of the study group that achieved at least the minimal amount of reliable change (i.e., not likely due to measurement error) (Haley & Fragala-Pinkham, 2006). The SDD of the total score was computed as $1.65 \times \sqrt{2} \times \text{SEM}$ and $1.96 \times \sqrt{2} \times \text{SEM}$ to obtain a 90 or 95% confidence interval.

2.4.3.4. Classification. Whether the diagnostic category based on the traffic light system between both test occasions is stable, is evaluated with a Kappa statistic for multiple ratings per subject. This statistical method measures classification agreement between different occasions taking chance agreement into account. In addition, the classification of children for who failed on more than four items was examined. Characteristics of the children who changed categories between both occasions are analyzed more precisely because a change in classification from normal to at risk/impaired or vice versa is clinically important.

3. Results

3.1. Feasibility

Half of the children had no failed items (Table 1). During the first test 92% or 46 of the children could do more than 6 out of 10 items (including Catching and Aiming). During the second test this percentage was 90 or 45 children. No significant differences between the number of failed items were found between younger (mean 3.3, SD 3.6) and older children (mean 2.2, SD 3.4; $p = .28$) or between boys (mean 2.9, SD 3.6) and girls (mean 2.6, SD 3.5; $p = .72$). Eight children had more than 4 failed items on one of the test occasions. Based on the numbers of missing items a "clean sample" was made, consisting of children which had $\leq$4 failed or missing items ($n = 42$).

Table 2 shows the items for which a procedural fault was recorded. The number of children that scored 0 on the items Catching and Aiming is also shown. Procedural faults were frequently made during the performance of item 3 (drawing bicycle trail) and item 7 (walking on a line heels raised). A large number of children could not yet catch a beanbag or throw a beanbag to hit a target.

Table 1
Number of missings out of 10 items per child for each test occasion.

Number of missings out of 10	First test occasion			Second test occasion		
	Number of children	Percent	Cumulative percent	Number of children	Percent	Cumulative percent
0	25	50	50	27	54	54
1	11	22	72	7	14	68
2	4	8	80	6	12	80
3	5	10	90	2	4	84
4	1	2	92	3	6	90
5	1	2	94	0	0	0
6	1	2	96	1	2	92
7	0	0	96	0	0	92
8	2	4	100	4	8	100

Table 2
Number of children for whom an item was missing ($n = 50$).

Item	Item in age band 1	First test occasion		Second test occasion	
1a	Posting coins – best hand	4	8%	4	8%
1b	Posting coins – other hand	5	10%	5	10%
2	Threading beads	4	8%	6	12%
3	Drawing trial 1	**18**	36%	**17**	34%
4	Catching beanbag	**13**	26%	**8**	16%
5	Throwing beanbag onto mat	**14**	28%	**12**	24%
6a	One-leg balance – best leg	9	18%	9	18%
6b	One leg balance – other leg	8	16%	9	18%
7	Walking heels raised	**12**	24%	**17**	34%
8	Jumping on mats	5	10%	8	16%

Bold: $\gg$20% of the children could not perform the task.

Table 3
Estimates for internal consistency of the MABC-2; Cronbach's α.

	$n_{all} = 50$	$n_{clean} = 42$[a]
First test occasion	.81	.70
Second test occasion	.87	.76

[a] Clean data set: only children who have ≤4 missing items.

3.2. Internal consistency

Table 3 shows the estimates for internal consistency of the Movement ABC-2. Cronbach's αranges from .70 through .87. Cronbach's αs above .70 are generally regarded as acceptable, over .80 as good, and over .90 as excellent (Vangeneugden, Laenen, Geys, Renard, & Molenberghs, 2005). The present results suggest that the internal consistency can be considered acceptable to good, even if 4 or more items are missing or replaced. These Cronbach's α scores demonstrate a sufficient homogeneity of all 10 items in the test.

3.3. Stability of test scores

Table 4 shows the ICC for component scores. The ICCs varied between .67 and .85. The ICC of the total score was .94 for the clean sample ($n = 42$) and .83 for the whole sample of test–retest children ($n = 50$). Based on the values suggested by Portney and Watkins (2000) to interpret the ICC values (>.75 good; .50–.75 moderate;,<.50 poor), the findings for the TSS of the 3 year old children indicate good reliability.

3.4. Standard error of measurement (SEM)

The SEM ranged between .73 or 1.47 standard scores (Table 5). The SEM doubled if children were tested by different assessors. As criterion for an acceptable precision of the SEM, a value ≤SD/2 was used (Wyrwich, Nienaber, Tierney, & Wolinsky, 1999). The standard scores of the Movement ABC-2 have a mean of 10 and SD of 3. All SEM values for the total score attained the criterion of less than 1.5 standard scores, suggesting an acceptable measurement precision of the Movement ABC-2 even in less optimal conditions (two different assessors) and data with many replaced items (Table 5).

3.5. Smallest detectable difference (SDD)

For the intra-tester test–retest design, SDD values are 1.7 or 3.4 depending on the desired percent confidence (Table 5). For the inter-tester design, the SDD went up to 3.2 or 4.1 standard scores. Clinical change can more easily be determined if a child is tested by the same assessor twice and when less then 4 missing items are replaced by the worst score of the age group.

3.6. Agreement on classification

Of the 50 children, 84% or 42 were classified in the same category on both test occasions (Kappa .66). The performance of 27 children was categorized twice in the green zone and 15 children scored in the amber/red zone. Of the 42 children in the clean sample 88% or 37 were categorized the same with a Kappa coefficient of .70. The intra-tester and inter-tester test–

Table 4
Intra-class correlation coefficients for the component scores for test–retest outcomes.

MABC-2 component	Test–retest $n_{all} = 28$	Test–retest $n_{clean} = 23$[a]
Manual dexterity	.85	.84
Aiming and catching	.74	.67
Balance	.75	.68

[a] Clean data set: only children who have ≤4 missing items.

Table 5
Test–retest results for 3 year old children on Movement ABC-2 divided into intra-tester and inter-tester test–retest.

Sample	ICC	SEM	SDD 90%	SDD 95%
Intra-tester test–retest ($n_{clean} = 23$)[a]	.94	.73	1.71	2.04
Intra-tester test–retest ($n_{all} = 28$)	.83	1.24	2.89	3.43
Inter-tester test–retest ($n_{clean} = 19$)[a]	.76	1.47	3.43	4.07
Inter-tester test–retest ($n_{all} = 22$)	.79	1.37	3.21	3.81

[a] Clean data set: only children who have ≤4 missing items.

Table 6
Percentage of agreement and Kappa coefficients for 3 year old children on Movement ABC-2 divided into intra-tester and inter-tester test–retest.

Sample	% agreement	Kappa
Intra-tester test–retest ($n = 23$)[a]	91	.81
Intra-tester test–retest ($n = 28$)	86	.71
Inter-tester test–retest ($n = 19$)[a]	84	.58
Inter-tester test–retest ($n = 22$)	82	.60

[a] Clean data set: only children who have $\leq$4 missing items.

retest Kappa statistic are shown in Table 6. The Kappa coefficient can range from minus 1 to 1; minus 1 indicates perfect disagreement, 0 indicates chance agreement, values between 0 and .2 present poor agreement, between .21 and .40 fair agreement, between .41 and .60 moderate agreement, between .61 and .80 good agreement, and 81 or higher excellent agreement (Landis & Koch, 1977). Thus, the present results indicate moderate to excellent agreement.

Finally, the scoring pattern of eight children was examined. These children had more than four out of eight item scores missing and were excluded to make a clean data set. Of these children, 4 scored in the impaired (red) range on both occasions and were probably not yet ready to be tested or had really poor motor coordination. They were clearly two times at the bottom end of the distribution and 95% of their peers performed better. One child improved from impaired to within the normal range. And 3 children were first classified as performing within normal range (green zone), but during retest they scored at risk (amber) or impaired (red) because they did not adhere to the test instructions in many items. Because of small numbers no statistical comparison was made between the whole sample and the 8 children. No clear pattern was found in the characteristics of these eight children. Half of them were boys and half girls, five were in the youngest (3:0–3:5) and three in the oldest age group (3:6–3:11 years). Six children were right handed, and two were left handed. Non-motor factors listed on the MABC-2 record forms for these 8 children were: hesitant 1×, timid 4×, anxious 2×, impulsive 1×, distractible 2×, lack of persistence 1×. However, for none of the children the assessors recorded that they thought that non-motor factors prevented the child from demonstrating his or her true movement capability.

Parents reported the following characteristics about these 8 children with more than four missings (number of parents reporting this characteristic for all 50 children is in brackets). One child (7/50) had seen a medical specialist for earlier health problems. One parent (3/50) reported motor problems, zero (1/50) vision or hearing difficulties, two parents (5/50) reported developmental speech or language problems and none (1/50) behavioural problems, none of the children (0/50) received any kind of motor or speech therapy.

In addition to the eight children already described, two more children in the clean sample were classified differently on the two test occasions. They both went from the 5th percentile (red zone) to the 25th (green zone). Noteworthy is that the better performance of these two children could not be explained by missing item substitution.

4. Discussion

The first aim of this study was to explore whether every day motor skills can be formally tested in children at the age of three. Our results show that the majority of 3-year-olds were able to complete the tasks in the Movement ABC-2 Test in accordance with the guidelines. Around 90% of these typically developing children were willing and able to perform more than half of the tasks. One of the most difficult tasks was holding a pen and drawing a trace between two boundary lines. That task has very specific instructions, e.g. not lifting the pen, not turning the paper more than 45°, and not changing drawing direction. Some of the children ($n = 6$) were not yet able to hold a pen in an age appropriate way (e.g. fist grip instead of any form of grip involving the fingers). Other items that children had difficulty with were walking on a line with heels raised and getting in balance on one-leg without the free foot touching the floor, or hooking it around the other leg. Since the Movement ABC-2 is a norm-referenced test, the item-, component- and total standard scores are adapted to the fact that young children are not yet able to perform the motor tasks well. It is important to realize that if a child for instance cannot jump or walk on a line on its toes, it will get a maximum low score for zero jumps or zero steps. Therefore the test will not differentiate between an attempt to perform an item or not even trying.

The second aim was to verify that 3 year olds could be tested (and retested) reliably. For this, the data of the normal test–retest (same assessor) design were used. The ICC showed excellent reliability (.94). Importantly, the analyses of the intra-tester reliability showed that including children with failed items did not deflate the reliability of the component scores (range .74–.85) and only slightly the total score (.83). Kappa scores for the classification into a "normal developing" or green category and group "at risk or with motor impairments" were good to excellent (.71 and .81). An explanation of the low impact of replaced scores might be that half of the excluded children (clean sample) were classified as very poor on both occasions and many children had failed items on the same tasks at both occasions. This means the replaced items were indeed failed items because most of these children could do the task properly. Conversely, it might be argued that other aspects such as the ability to understand the task or the limited attention span of 3-year-old children may cause less consistent performance and therefore increased measurement error. Yet that is not what was found in this study.

The SDD represents the smallest change in individual scores that reflects real change rather than measurement error. The minimal change in total test score between two test occasions needs to be at least 1.7 (90%CI) or 2.0 (95% CI) above (or below) before one may conclude that an individual 3 year old child has really improved (or worsened). These values indicate that

Movement ABC-2 is sensitive enough to detect individual changes. The SDD was much larger if replaced scores for the failed items are used, namely 2.9 (90%) and 4.1 (95%) standard scores. Although outcomes were still reliable, substituting more than 4 items does influence the variability of the outcomes of the individual child.

In Movement ABC-2-Test, substantial adaptations have been made to obtain a reliable score by improving the standardized description and procedures for scoring each item. The content is more or less the same as in the previous edition. Interestingly, the reliability values of the 3 year old sample are not less precise than found in the literature so far. For instance the mean SDD values over all age groups (3–16) reported in the Movement ABC-2 manual are 2 (90% CI) and 3 (95% CI) standard scores (Henderson et al., 2007).

Two studies tested children close to our age group with the first edition of the test. Chow and Henderson (2003) tested 138 Chinese preschool children aged 4–6 years and reported a test–retest ICC of .77 for the total test score when children were tested twice by the same tester. Van Waelvelde et al. (2007) studied the test–retest reliability in a group of young children with low motor performance and obtained an ICC of .88. These values were lower than the .94 found in the present study. It appeared that 3 year old children could perform comparably if tested two times. This is still the case if a number of failed item scores is replaced by the poorest score of the reference group, as instructed in the manual (ICC .83). However, the therapist should take into account that substituted scores may not always give the most valid impression of the child's motor competence. Occasionally it may reflect more a child's behavioural response to being tested or asked to do tasks by an unfamiliar person. It may take extra time for a few of these young children to get acquainted with the test situation.

Our results confirm that all the reliability values of a test–retest design, with one assessor which is a standard procedure, are within the good to excellent ranges. We expected that being tested by different examiners would have a high impact on children in this age group, as young children might be more sensitive to the way they are instructed and encouraged. Our data only partially confirm this assumption. If approached by another examiner, the children were still guided to very similar task performance on both test occasions (ICC .79 and .76). Although more variance is introduced, reliability is still graded as good. Both ICC values are similar to those reported in the Chow study (2001) for items in the experimental version of the Movement ABC-2, which ranged from .69 to .92 (mean ICC .79).

The present study was limited by the fact that the children did not exhibit any clear motor impairment. Although children over the full range of scores were included, no definite conclusions can be drawn with regard to the reliability of the test when used in children with motor impairments and co-occurring conditions. Second, the assessors were experienced professionals used to test children at risk or with motor impairments. Therefore caution is necessary if applying these results to other assessors that are less familiar with guiding young children to optimal motor performance.

Despite the excellent test–retest reliability, two children were misclassified and went from the orange/red to the green zone. This most likely means that they had to get used to the test circumstances and items. They may have needed more time to get used to the tester, test situation or test items. These changes represent the normal variability when testing young children. It is therefore very important that the interpretation of the classification of very young children using the traffic light system is done very carefully, especially when this is based on many replaced scores for failed items. In order to draw conclusions about a child's motor competence, it is recommended to retest a young child before deciding about starting intervention. The results of both test occasions together with careful history taking can be used for diagnostic purposes. A small gain in age can make children more task-oriented and during a second test occasion children are more familiar with the test situation.

5. Conclusion

The Movement Assessment Battery for Children-2 is a reliable instrument to measure motor performance in 3 year old children. Despite the young age and a number of tasks not yet performed according to the instructions, the test results are highly reproducible. We have shown that daily movement skills in very young children like threading beads, jumping or drawing can be tested reliably, even with replaced scores for procedural faults. It may be concluded that substituting scores does not have a large impact on reliability. It is most likely that children that are unable to perform a task as instructed, would not have the motor competence to do so. It is recommended that test–retesting be performed by the same therapist and to try to redo items to get a valid impression if children cannot complete the task or need more time to get acquainted with a formal testing situation. In all cases poor results in 3 year olds should be confirmed by a second testing and labelling a child as motor impaired without confirmation by history taking, parental questionnaires (e.g. Movement ABC-2 Checklist) should be avoided.

Acknowledgements

Great appreciation goes to all the parents and children participating in this study. We wish to thank Mrs. Jantine Booij and Mrs. Marcia Backer for testing the children, and Dr. Theodore Sana for proof-reading our manuscript.

References

Chow, S. M. K., & Henderson, S. E. (2003). Interrater and test–retest reliability of the Movement Assessment Battery for Chinese Preschool Children. *American Journal of Occupational Therapy, 57*, 574–577.

Croce, R. V., Horvat, M., & McCarthy, E. (2001). Reliability and concurrent validity of the Movement Assessment Battery for Children. *Perceptual Motor Skills, 93*, 275–280.

Geuze, R. H., Jongmans, M. J., Schoemaker, M. M., & Smits-Engelsman, B. C. M. (2001). Clinical and research diagnostic criteria for Developmental Coordination Disorder: A review and discussion. *Human Movement Science, 20*, 7–47.

Haley, S. M., & Fragala-Pinkham, M. A. (2006). Interpreting change scores of tests and measures used in physical therapy. *Physical Therapy, 86*, 735–743.

Henderson, S. E., & Sugden, D. A. (1992). *Movement Assessment Battery for Children: Manual*. London: Psychological Corporation.

Henderson, S. E., Sugden, D. A., & Barnett, A. L. (2007). *Movement Assessment Battery for Children – second edition (Movement ABC-2); examiner's manual*. London: Harcourt Assessment.

Jongmans, M. J. (2005). Early identification of children with Developmental Coordination Disorder. In D. A. Sugden & M. Chambers (Eds.), *Children with Developmental Coordination Disorder* (pp. 155–167). London: Whurr.

Landis, J. R., & Koch, G. G. (1977). The measurement for observer agreement for categorical data. *Biometrics, 33*, 159–174.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.

Portney, L. G., & Watkins, M. P. (2000). *Foundations of clinical research: Applications to practice* (2nd ed.). NY: Pratice Hall Health.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations – uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428.

Smits-Engelsman, B. C. M. (2010). *Handleiding Movement ABC-2-NL [Manual Movement ABC-2-NL]*. Amsterdam: Pearson.

Smits-Engelsman, B. C. M., Fiers, M. J., Henderson, S. E., & Henderson, L. (2008). Interrater reliability of the Movement Assessment Battery for Children. *Physical Therapy, 88*, 286–294.

Vangeneugden, T., Laenen, A., Geys, H., Renard, D., & Molenberghs, G. (2005). Applying concepts of generalizability theory on clinical trial data to investigate sources of variation and their impact on reliability. *Biometrics, 61*, 295–304.

Van Waelvelde, H., Peersman, W., Lenoir, M., & Smits-Engelsman, B. C. M. (2007). The reliability of the Movement Assessment Battery for Children for preschool children with mild to moderate motor impairment. *Clinical Rehabilitation, 21*, 465–470.

Wyrwich, K. W., Nienaber, N. A., Tierney, W. M., & Wolinsky, F. D. (1999). Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. *Medical Care, 37*, 469–478.