Routledge
Taylor & Francis Group

Measurement in Physical Education and Exercise Science

ISSN: 1091-367X (Print) 1532-7841 (Online) Journal homepage: http://www.tandfonline.com/loi/hmpe20

Inter-rater and test–retest (between-sessions) reliability of the 4-Skills Scan for dutch elementary school children

Willem G. van Kernebeek, Antoine W. de Schipper, Geert J.P. Savelsbergh & Huub M. Toussaint

To cite this article: Willem G. van Kernebeek, Antoine W. de Schipper, Geert J.P. Savelsbergh & Huub M. Toussaint (2017): Inter-rater and test–retest (between-sessions) reliability of the 4-Skills Scan for dutch elementary school children, Measurement in Physical Education and Exercise Science, DOI: 10.1080/1091367X.2017.1399891

To link to this article: https://doi.org/10.1080/1091367X.2017.1399891

MEASUREMENT IN PHYSICAL EDUCATION AND EXERCISE SCIENCE
https://doi.org/10.1080/1091367X.2017.1399891

Inter-rater and test–retest (between-sessions) reliability of the 4-Skills Scan for dutch elementary school children

Willem G. van Kernebeek[a], Antoine W. de Schipper[a], Geert J.P. Savelsbergh[b,a], and Huub M. Toussaint[a]

[a]Faculty of Sports and Nutrition, Amsterdam University of Applied Sciences, Amsterdam, The Netherlands; [b]Department of Human Movement Sciences, Faculty of Behavioural and Movement Sciences, Vrije Universiteit, Amsterdam, The Netherlands

ABSTRACT

In The Netherlands, the 4-Skills Scan is an instrument for physical education teachers to assess gross motor skills of elementary school children. Little is known about its reliability. Therefore, in this study the test–retest and inter-rater reliability was determined. Respectively, 624 and 557 Dutch 6- to 12-year-old children were analyzed for test re-test and inter-rater reliability. All tests took place within the school setting. The outcome measure was age-expected motor performance (in years). Results showed a small practice effect of .24 years for re-test sessions and assessment of motor skills was possible with acceptable precision (standard error of measurement = .67 years). Overall, intraclass correlation coefficient (ICC) was .93 (95% confidence interval: .92–.95) for test–retest reliability and .97 for inter-rater reliability. For the repeated measures, the smallest detectable change (SDC) was 1.84 and limits of agreement were −1.60 and 2.08 years. It can be concluded that the 4-Skills Scan is a reliable instrument to assess gross motor skills in elementary school children.

KEYWORDS

motor skill assessment; physical education; primary school children; psychometrics; reliability

Introduction

A major aspect of a child's development is the acquisition of gross motor skills. Longitudinally, studies show that the mastery of fundamental movement skills (FMS) and gross motor skills at an early age appears to be an important factor for their physical activity and well-being during childhood and later in life (Bryant, James, Birch, & Duncan, 2014; Gallahue, Ozmun, & Goodway, 2012; Jaakkola, Yli-Piipari, Huotari, Watt, & Liukkonen, 2016; Lopes, Rodrigues, Maia, & Malina, 2011; Lubans, Morgan, Cliff, Barnett, & Okely, 2010; Payne & Isaacs, 2012). Beneficial effects with respect to cognitive and social well-being have also been mentioned (Payne & Isaacs, 2012).

Physical education (PE) lessons can play a central role in the acquisition of FMS and gross motor skills (Babin, Katić, Ropac, & Bonacin, 2001; Morgan et al., 2013; Siedentop, 2009; Wrotniak, Epstein, Dorn, Jones, & Kondilis, 2006) and time spent in physical activity (Meyer et al., 2012). Currently, it is mandatory for PE teachers to monitor the progression and development of their pupils. Motor development logically makes part of that student tracking system.

Several instruments are available for assessing the level of gross motor skills, some of which are widely used and accepted, such as the Movement Assessment Battery for Children-2 (MABC-2; Henderson, Sugden, & Barnett, 2007). However, most of these instruments are designed to detect delayed or abnormal motor development and are unsuitable for monitoring purposes where multiple measurements are desired to follow development longitudinally. The MABC-2 is found to be sufficient sensitive in its ability to monitor treatment progress in children with motor impairment and Developmental Coordination Disorder (DCD) (Wuang, Su, & Su, 2012). However, more research is needed to assess its responsivity for children without motor problems (general population). Besides, due to the relative high standard error of measurement (SEM) and possible learning effect, repeated testing at short time intervals is not recommended (Van Waelvelde, Peersman, Lenoir, & Smits Engelsman, 2007). Also, discontinuation of the scales over the age bands might interfere with longitudinal monitoring (Blank, Smits-Engelsman, Polatajko, & Wilson, 2012). Furthermore, many tests appear to be unsuitable for a school setting because of the considerable amount of time it takes to assess all children. The Test of Gross Motor Development—Second Edition (TGMD-2) (Ulrich, 2000) and Körperkoordinationstest für Kinder (KTK; Kiphard & Schilling, 2007) have been mentioned as feasible for PE lessons in a school setting (Vandorpe et al., 2011). These tests still take approximately 15 minutes per child and

CONTACT Willem G. van Kernebeek w.g.van.kernebeek@hva.nl Faculty of Sports and Nutrition, Amsterdam University of Applied Sciences, Dr. Meurerlaan 8, 1067 SM. Amsterdam, The Netherlands.

require specific testing materials, such as beanbags, balance beams, or boxes, that are not part of the standard inventory of every sports hall. Since the choice of a motor skills test depends on the context of use, it has been challenging to find an instrument that better suits the PE lesson setting. Over the past decade, an instrument was developed that can be used during PE classes and specifically fits the context of PE lessons (Van Gelder & Stroes, 2010). Assessment of gross motor skills with this 4-Skills Scan results in a performance-based outcome (PerfO) in terms of Motor Age. The 4-Skills Scan was developed by Van Gelder and Stroes and finds its origin in the motor development theory of Ayres (1963) and Gesell (1975). It includes the age-range around the transition from FMS to context specific skills at about the age of 7 (Clark & Metcalfe, 2002). The outcome measure of the 4-Skills Scan is "motor age." This "Motor Age" is an age-expected motor skills level and easy to interpret when compared with calendar age. An advantage of this outcome measure is the easy comparison of children across age-bands. The main aim of the 4-Skills Scan is to serve as an instrument that monitors children's motor skill development. PE teachers have developed this instrument and, in addition, feasibility for the PE context was an important criterion. These characteristics make it possible to assess the motor skill level for all elementary school children in a quick and feasible manner during PE lessons. As a result, the 4-Skills Scan gained popularity among PE teachers in The Netherlands.

Overall, the 4-Skills Scan seems to be a well-substantiated test (Van Gelder & Stroes, 2010), and years of iterations have preceded in order to determine the sequence of test items and to match the difficulty levels with calendar age. Also, the three main FMS categories, locomotion, manipulative or object control, and stability skills (Gallahue et al., 2012), seem to be covered by the four subscales of this test (Van Gelder & Stroes, 2010). Until recently, little was known about the validity. However, in a study conducted by Van Kernebeek, De Kroon, Savelsbergh, and Toussaint (manuscript submitted for publication), the test was concluded to be valid for assessing gross motor skills.

It is important to keep in mind a test's design and context of use (Kielhofner, 2006) when interpreting reliability values. For example, the Bruininks-Oseretsky Test of Motor Proficiency (BOTMP), the MABC-2, and the Peabody Developmental Motor Scales (PDMS-2), each have a minimum of 8 test items and take at least 15 minutes to conduct (Cools, De Martelaer, Samaey, & Andries, 2009; Wiart & Darrah, 2001). The 4-Skills Scan, on the other hand, merely consists of 4 subscales that take 2 minutes each and is conducted in a boisterous setting.

In order to get a clearer picture of the value of data collected with the 4-Skills Scan, it is important to get better insight into its reliability. It might also stimulate communication between youth healthcare professionals, such as pediatricians, pediatric physiotherapists, and PE teachers. Therefore, the aim of this study is to (1) assess the test–retest reliability; and (2) the inter-rater reliability of the 4-Skills Scan when conducted by specially trained test conductors.

Method

Participants

For test–retest reliability, a representative sample of the general school population, consisting of 629 third to eighth grade children (6- to 12-year-old, 48.9% boys), were recruited at three elementary schools in Amsterdam, The Netherlands, throughout November and December 2014. These three schools were selected because of the presence of double-sized sports halls. Five pupils were excluded, due to incorrect or incomplete motor skills assessment as a result of measurement error (n = 4) or conflicting injury (n = 1). Since the number of missing data was low, list-wise deletion of missing values was deemed legitimate and resulted in complete data for 624 pupils (see Table 1 for a description of the study sample). Data derived from 557 children, recruited at eight different schools, were analyzed for inter-rater reliability. Children with injuries or other physical impairments were excluded from analyses. Children gave assent to participation and informed consents were obtained from their parents. The study protocol was approved by the Ethics Committee Human Movement Sciences, VU University Amsterdam (2014-59).

Instrument

For assessing motor skills, the 4-Skills Scan of Van Gelder and Stroes (2010) was used. This is an easy to conduct, quantitative motor skills test and and specifically

Table 1. Descriptives of the study population for test–retest and inter-rater reliability.

Study sample	N	Gender % boys	Height (cm) ± SD	Weight (kg) ± SD
Test–retest sample	624	49.0	135.3 ± 11.8	31.5 ± 9.4
6-years-old	104	43.3	122.3 ± 4.9	24.0 ± 3.6
7-years-old	133	54.9	127.3 ± 6.1	26.8 ± 4.7
8-years-old	139	41.7	133.0 ± 5.8	30.9 ± 10.1
9-years-old	82	50.0	138.8 ± 6.0	31.8 ± 5.0
10-years-old	64	57.8	145.9 ± 6.0	37.1 ± 6.3
11-years-old	87	52.9	152.1 ± 7.0	42.0 ± 10.3
12-years-old	15	40.0	154.1 ± 7.8	44.2 ± 7.2
Interrater sample	561	53.8	134.7 ± 11.4	31.4 ± 8.8
Junior grades (age 5–8)	331	54.0	127.4 ± 7.5	26.8 ± 5.3
Senior grades (age 9–12)	230	53.5	144.6 ± 7.8	37.7 ± 8.6

developed for the PE lesson context. The four subscales of the 4-Skills Scan were administered to all participants: "standing-still," "jumping-force," "jumping-coordination," and "bouncing-ball" (Van Gelder & Stroes, 2010; see appendix). The outcome of the test is a PerfO in terms of "Motor Age." It is a multi-dimensional concept consisting of four rating subscales each made up of nine items that vary in difficulty but assumed replications of the same construct. Thus, the nine items within each subscale are similar and measure the same underlying constructs, but become more difficult across age bands. For example, hopping while covering distance is assumed to be more difficult that without covering distance (jumping-force) and an alternating "shuffle-jump" pattern with anteflexion and retroflexion of the hip (shuffle-jump) is assumed to be more difficult than making synchronous ski-jumps (jumping-coordination).

Procedure

The motor skills assessments took place during regular PE classes. All test-conductors were students or already graduated PE teachers, physiotherapists, or Human Movement scientists. All were familiar with the 4-Skills Scan beforehand and were provided with additional training in order to assure scoring consistency and protocol compliance. For test–retest reliability, 19 test-conductors at 3 different schools contributed in an alternating test-team composition. For inter-rater reliability, 18 test-conductors at eight different schools contributed, also in an alternating test-team composition. The PE lessons started with a brief explanation to the children about the purpose of the test. Test-conductors were informed about the age, class, and gender of the children. Children were told "to do the best they could given their own ability." The pupils of each class were then divided alphabetically over the test stations of the four subscales. Thus, children started at different subscales, and therefore, any possible crossover practice effect from one subscale to the next was evened out. An extra demonstration of the task to be performed was given at each station. To ensure a positive experience, the initial difficulty for each subscale task was age-dependent and generally well below the expected maximal performance of the tested child. Children were given two attempts per difficulty levels. Children proceeded to the next station when execution of the motor task appeared too complex or when the end of the scale was reached.

Test–retest reliability

In order to assess test–retest reliability, two connected sports halls were both divided into four compartments in which the assessments of each subscale took place. Due to practical considerations, a time interval of 30 minutes between test and retest was the option chosen. This way, both the test and retest could take place during one PE lesson for grades three to eight.

In accordance with the usual protocol, each pupil performed tasks on every subscale and after registration of their highest achievement the pupil was directed to the next test station. No feedback was given about their test scores during the PE lesson, so the second test-conductors were unaware of the scores achieved during the first test. As a consequence of this design, retest scores were assessed by different test-conductors.

Inter-rater reliability

Although the inter-rater reliability of an instrument is based on the combined scores of individual items, inter-rater reliability evaluation for individual item is often omitted (Orsi, Drury, & Mackert, 2014). In this study we address the importance of assessing inter-rater reliability for individual items and assessed inter-rater reliability by determining the inter-rater reliability per item. For this, a practical and suitable design was chosen, where at one of the subscales, two test-conductors were present and scored each child's achievement simultaneously but independently. The protocol was slightly adjusted for this occasion: rather than having the children "quit at failure," attempts were made on every task, up to the highest difficulty level.

Outcome measures and data analyses

Motor age

In order to calculate a child's Motor Age, the four subscale scores where averaged by the following formula:

$$Motor\ Age = \frac{\begin{pmatrix} level,'balance' + level,'jumping\ force' \\ +level'\ jumping\ coordination' \\ +level'\ bouncing(ball)' \end{pmatrix}}{4}$$

Test–retest and inter-rater reliability

Since for analyses, the ICC is preferred (Portney & Watkins, 2008; Streiner, Norman, & Cairney, 2014), a Two-Way-Random effect ICC ($ICC_{absolute}$ 2.1) was calculated for both test–retest and inter-rater reliability in accordance with Shrout and Fleiss (1979). According to Portney and Watkins (2008), an ICC of .75 is considered good, .75 to .50 as moderate, and below .50 as poor. For both test–retest and inter-rater reliability, agreement was based on differences in the Motor Age. For test–retest analysis, besides

calculating ICC-values for the whole cohort, a stratified analysis by age was done. Regarding the inter-rater reliability, two analyses were made: one for the junior grades (6- to 8-years-old) and one for the senior grades (age 9- to 12-years-old). This is done to take account for the fluctuation in performance that can happen at young age due to distraction or misunderstanding of the purpose of the test (Blank et al., 2012). At a later age, children tend to perform more stable on tests.

Measurement error and limits of agreement (LoA)

For assessing measurement error, both the SEM ($SEM_{consistency}$) and LoA were calculated. Bland and Altman plots have been recommended by Lamb (1998) since these give a visual understanding of the range of individual differences. Here, the difference of the two test occasions is plotted against the mean of the two test occasions. Absolute reliability was determined by the SEM and calculated as follows: $SEM = \frac{SD_{diff}}{\sqrt{2}}$ (De Vet, Terwee, Knol, & Bouter, 2006), where $SEM \leq \frac{SD}{2}$ was taken as the criterion for acceptable precision (Wyrwich, Nienaber, Tierney, & Wolinsky, 1999). The Standard Error of the difference was calculated as follows:$SE_{diff} = \sqrt{2 * SEM^2}$. LoA were calculated as:$LoA = mean\ difference\ (bias) \pm 1.96 * SD_{diff.}$. Here, the mean difference (bias) is the systematic error and the standard deviation of the difference ($SD_{diff.}$)is the random error.

Practice effect

For calculating the practice effect, the mean difference score (MDS) is calculated. A paired *t*-test was performed between test and retest values in order to detect a possible practice effect.

SDC

A more clinical relevant measure for reliability is the SDC (SDC_{95}), which was calculated as follows: $SDC_{95} = SEM * 1.96 * \sqrt{2}$(De Vet, Beckerman, Terwee, Terluin, & Bouter, 2006; Mokkink et al., 2010; Stratford, 2004). This can be seen as an outcome measure with a confidence interval of 95%, with 1.96 representing the z-score and $\sqrt{2}$ as a means to account for accumulating errors associated with repeated measures.

Results

Motor age

The results for 624 pupils were analyzed to assess test–retest reliability on the Motor Age with the 4-Skills Scan. Data for the individual subscale were also analyzed. The descriptive statistics and test–retest reliability scores are summarized in Tables 1 and 2.

Test–retest reliability

ICC's ($ICC_{absolute}$) for all ages ranged from .76 to .94, with an overall score of .93 ($p < .01$; see Table 2). According to Portney and Watkins (2008), these ICC's can be considered as good outcomes for all subscales. ICC's calculated per age band (see Table 3) showed comparable values (ranging from .82 to .87), with the exception for 6-year-old children (ICC = .74).

Measurement error and agreement

For the Motor Age and the subscales "jumping-force" and "bouncing-ball," corresponding SEM-values met the criterion for acceptable precision. The subscales "standing-still" and "jumping-coordination" however, did not meet the criterion for acceptable precision. Figure 1 presents a Bland and Altman plot for the test–retest sessions. The re-test shows a bias of –.24 years, meaning that the second examiner scored systematically higher than the first examiner. Individual retest scores above the upper range of +2.09 years or below the range of –1.60 years can be interpreted as a true change in the Motor Age. As can be seen in Figure 1, there are children in both the upper and lower ranges that show considerable changes within the two test sessions.

Table 2. Test–retest reliability (ICC 2.1) for *N* = 628.

	Test		Re-Test		Bland & Altman						
4-Skills Scan	Mean	SD	Mean	SD	Mean diff [a]	SD	95% LoA	$ICC_{absolute}$	95%-CI	SEM	SDC_{95}
Standing-still	8.58	3.24	8.89	3.3	.31**	2.85	−5.9 + 5.3	0.76**	0.72–0.80	2.02	5.59
Jumping-force	8.53	2.42	8.55	2.43	.02	1.15	−2.3 + 2.2	0.94**	0.93–0.95	.81	2.25
Jumping-coordination	9.82	2.09	10.21	2.25	.39**	1.81	−3.9 + 3.2	0.78**	0.74–0.82	1.28	3.55
Bouncing-ball	7.61	1.82	7.86	1.94	.25*	1.19	−2.6 + 2.1	0.89**	0.86–0.90	.84	2.33
Motor Age	8.63	1.87	8.88	1.98	.24**	0.94	−2.1 + 1.6	0.93**	0.92–0.95	.67	1.84

$ICC_{absolute}$: Intraclass Correlation Coefficient; SEM: Standard Error of Measurement; SDC_{95}: Smallest Detectable Change at the 95% Confidence Interval.
[a]difference scores were calculated by subtracting the baseline score from the retest score
*items significant at $p < .05$; **items significant at $p < .01$.

Table 3. Test–retest reliability (ICC 2.1) per age-band.

4-Skills Scan	6	7	8	9	10	11	12	All Ages
N	104	133	139	82	64	87	15	624
ICC	.74	.82	.85	.87	.84	.87	.84	.93
SEM	.57	.60	.70	.71	.72	.66	.86	.67

$ICC_{absolute}$: Intraclass Correlation Coefficients for every age band; SEM: Standard Error of Measurement.

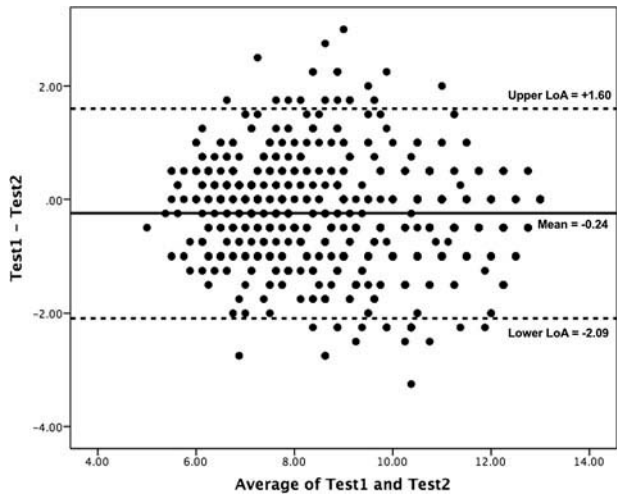

Figure 1. Bland and Altman graph with 95% LoA. The differences between test and retest plotted against their mean for each subject for Motor Age (years) for 628 participants. LoA = mean difference (bias) ± 1.96 SD.

Practice effect (stability of test scores over time)

An average practice effect by the children of .24 years ($p < .01$, 95%-CI: .17 – .32) was found on the Motor Age. With the exception of "jumping-force," *t*-tests for mean differences between test and retest for the separate subscales showed significant practice effects (Table 2).

SDC

The SDC_{95} for the Motor Age was 1.84, indicating that a change over time of more than 1.84 years can be seen as a true change in the Motor Age in 95% of the pupils (Table 2).

Inter-rater reliability

A total of 557 children were included for analyses for the inter-rater component (see Table 1). Table 4 presents overall averaged ICC's for both the test as a whole and each separate subscale. In addition, also ICC's per age band are shown in Table 4. A high overall ICC-value was found for the total test (ICC = .97 ± .01). Also, for the separate subscale, ICC's of .95 to .98 were found. Analyses per age-band sometimes showed lower, but generally similar ICC-values, ranging from .90 to .98. All ICC-valueswere above .75, indicating a good inter-rater reliability for the 4-Skills Scan as well as the separate subscales for each age band.

Discussion

The motor development is an important aspect of children's healthy development. PE in primary school is—among other things—aimed at enhancing motor skill levels. Recently, monitoring the development of motor skills became mandatory in the Netherlands. The need for a quick and feasible instrument with good test–retest and inter-rater reliability led to the selection of the 4-Skills Scan as a potential instrument. In order to evaluate the merits of the 4-Skills Scan, it was studied using ICC's for test–retest and inter-rater reliability, the Bland and Altman method, SEM, and the systematic error.

Test–retest reliability

The test–retest reliability was assessed with an overall ICC of .93 (95%-CI: .92 – .95) for the 4-Skills Scan. This can be interpreted as a high test–retest reliability and is comparable to established motor skills tests such as the BOTMP (ICC = .86 to .89; Moore, Reeve, & Boan, 1986; Wiart & Darrah, 2001), the MABC-2 (ICC = .77 to .97; Chow & Henderson, 2003; Wuang et al., 2012), and the KTK (correlation coefficient = .97; Vandorpe et al., 2011). A study by Houwen, Hartman, Jonker, and Visscher (2010) on the reliability of the 12-items TGMD-2 among 75 6- to 12-year-old children showed an ICC of .92 (95%-CI: .88 – .98) for test–retest reliability with regard to the test as a whole. A study by Croce, Horvat, and McCarthy (2001) showed similar ICC's for the MABC (ICC = .95) with 106 participants and a 1-week time interval.

The stability of the Motor Age as the outcome of the 4-Skills Scan was determined by using MDS (retest—baseline score). For the Motor Age, a significant systematic error of .24 year was found (see Table 2). "Jumping-force" appears to be the most stable subscale with a non-significant systematic difference of .02 year. "Jumping-coordination" was most susceptible for a practice effect, resulting in a significant systematic difference of .39 year. A possible mechanism behind these differences in practice effects might be that the instruction for a familiar task like hopping mostly refers to an inherent ability where the quality of performance may differ between children. Instructions for less familiar and complex jumping-coordination tasks refer to acquirable skills that some

Table 4. Inter-Rater reliability (ICC 2.1) for the whole group, for junior, and senior grades.

	Whole Group			Junior Grades			Senior Grades	
4-Skills Scan	*N*	ICC	95%-CI	*N*	ICC	95%-CI	*N*	ICC	95% CI
Subscales									
Standing-still	114	.97	.96–.98	79	.94	.91–.96	35	.95	.91–.98
Jumping-force	148	.98	.98–.99	92	.97	.96–.98	56	.97	.95–.98
Jumping-coordination	109	.95	.93–.97	54	.90	.82–.94	55	.98	.97–.99
Bouncing-ball	212	.96	.95–.97	120	.94	.91–.96	92	.95	.92–.97
Average		.97			.94			.96	

All items significant at $p < .001$.

children instantly—or after some practice—can master.

Due to practical considerations, test–retest reliability was evaluated based on same-day scores with a time interval of half an hour. The found systematic difference can be seen as a practice effect rather than a learning effect. According to Beglinger et al. (2005), the magnitude of the practice effect may be influenced by the time span between two measurements. A biological recall or motor memory of the movement might be the responsible mechanism behind this practice effect, since this mechanism can occur even after little practice (Singer, 1980). A larger time-interval for retest sessions will most likely result in a smaller systematic difference.

Given the context of use and the speed at which this test can be conducted, test–retest reliability as well as the SEM can be interpreted as good for this instrument. A SEM of .67 for the total 4-Skills Scan means that a child's individual Motor Age can be calculated with a precision of .67 year and that it meets the criterion for acceptable precision (Wyrwich et al., 1999). The SEM of the 4-Skills Scan is slightly higher than the MABC-2 (SEM = .53; Wuang et al., 2012), but this was expected given the larger number of test items of the MABC-2.

For the subscales, the ICC's ranged from .76 to .94. This indicates good test–retest reliability for the separate subscales. Little research has been done on test item reliability for other motor skills tests. However, a study by Henderson et al. (2007) showed MABC-2 test item ICC's ranging from .73 to .84. Another study, by Wuang et al. (2012), on the reliability of the MABC-2 for children with DCD, showed test-item ICC's ranging from .88 to .99.

For the subscale "jumping-coordination" and "standing-still," the SEM turned out to be relatively high. The SEM for these two subscales did not meet the criterion for acceptable precision. The increased SEM for "jumping-coordination" might be a result of the rating scale used for the PerfO. Test items on this subscale differ, but are considered to be similar motor tasks. The identified difficulty levels, however, might not follow the typical developmental sequence of motor skills for every child. The order in which some motor skills develop, might be the result of practice or dissimilar cultural factors, and consequently result in individual variation in the sequence of motor skill acquisition (Kamm, Thelen, & Jensen, 1990). For "standing-still," wobbling results in a lower test score of 1 or 2 years (see appendix). Besides the fact that wobbling is unwanted while carrying out the task of "standing still," it is a rather subjective aspect to be judged by the test-conductor, and is therefore, prone to causing inter-rater discrepancies. Subjective judging plays hardly a role evaluating "jumping-force" and "bouncing-ball."

When considering gross motor skills as a single construct, it is important to interpret the test result of the Motor Age rather than the single sub-scale scores. The individual subscales add up to a complete picture of the child's gross motor skills. In addition, a more precise test result is given by the average of all subscales, since the SEM is relatively low (see Table 2). For the test as a whole, the SEM met the criteria for acceptable precision. The Bland and Altman graph shows the LoA, ranging from –1.60 to 2.08 for test–retest reliability. All scores outside of these ranges can be considered as true changes of the child's Motor Age. Although the test–retest reliability of the 4-Skills Scan is good, extra assessments or multiple assessments throughout the year will improve the test's accuracy and reduce the absolute measurement error, since repeated testing is an effective strategy for averaging outliers.

Inter-rater reliability

The results of 557 children were included in the analyses for the inter-rater reliability. The results indicate a good inter-rater reliability with an overall ICC of .97, ICC's of .95 or higher for the individual subscales, and ICC's of .85 or higher across the age-bands. The present

inter-rater reliability scores compare well to those of other motor skill tests. For instance, Chow and Henderson (2003) assessed the inter-rater reliability for the MABC-2 using a similar method to the current study and also found a high overall $ICC_{interrater}$ of .96. Houwen et al. (2010) used video recording to assess inter-rater reliability. Two independent examiners evaluated video recordings, resulting in $ICC_{interrater}$ of .89 (95%CI: .81 – .93) for the total test of TGMD-2.

The 4-Skills Scan has a high test–retest and inter-rater reliability that is comparable to the established motor skills tests. This is an interesting observation when bearing in mind that the 4-Skills Scan consists of only four subscales. With the exception of the KTK, all other mentioned motor skill tests consist of more subscales. In addition, the test–retest assessment was carried out by two different raters. Therefore, both errors—test–retest and inter-rater—contribute to the total error reported. Compared to other motor skills tests, the SEM is somewhat higher for the 4-Skills Scan. This is mainly due to the non-objective wobbling aspect in the "standing-still" task. However, with due consideration of the previously mentioned criterions (such as feasibility and necessary testing material), the results of this study support the use of this test as an instrument for monitoring motor development and the effects of PE lessons on motor development on an individual level.

Some worthwhile directions for future research can be pointed out. Since PE teachers are the professionals who will most often use the 4-Skills Scan, examining of the reliability done by PE teachers might result in more ecologically valid study results. Furthermore, determining the validity would be of value to legitimize the use of this instrument to evaluate the motor development of children.

For many years PE teachers did not have the tools to adequate assess gross motor skills within their lessons. This study gives insight in the reliability of the 4-Skills Scan. Knowing the value of data gathered with this instrument not only empowers the PE teacher to monitor children's progression but also improves communication between youth healthcare professionals, such as pediatricians, pediatric physiotherapists, and PE teachers.

In conclusion, this study shows that the 4-Skills Scan is a reliable tool for assessing quantitative aspects of gross motor skills in elementary school children. With its unique context of use, it is suitable as an instrument for PE teachers for assessing motor skills level and the progress thereof as a result of PE lessons.

Funding

The current study is funded by RAAK-PRO/SIA (ref. 2014-01114PRO), part of Nederlandse Organisatie voor Wetenschappelijk Onderzoek.

References

Ayres, A. J. (1963). The development of perceptual-motor abilities: A theoretical basis for treatment of dysfunction. *American Journal of Occupational Therapy*, *17*, 221–225.

Babin, J., Katić, R., Ropac, D., & Bonacin, D. (2001). Effect of specially programmed physical and health education on motor fitness of seven-year-old school children. *Collegium Antropologicum*, *25*(1), 153–165.

Beglinger, L., Gaydos, B., Tangphaodaniels, O., Duff, K., Kareken, D., Crawford, J., … Siemers, E. (2005). Practice effects and the use of alternate forms in serial neuropsychological testing. *Archives of Clinical Neuropsychology*, *20* (4), 517–529. doi:10.1016/j.acn.2004.12.003

Blank, R., Smits-Engelsman, B., Polatajko, H., & Wilson, P. (2012). European Academy for Childhood Disability (EACD): Recommendations on the definition, diagnosis and intervention of developmental coordination disorder (long version): EACD Recommendations. *Developmental Medicine & Child Neurology*, *54*(1), 54–93. doi:10.1111/j.1469-8749.2011.04171.x

Bryant, E. S., James, R. S., Birch, S. L., & Duncan, M. (2014). Prediction of habitual physical activity level and weight status from fundamental movement skill level. *Journal of Sports Sciences*, *32*(19), 1775–1782. doi:10.1080/02640414.2014.918644

Chow, S. M., & Henderson, S. E. (2003). Interrater and test–Retest reliability of the movement assessment battery for Chinese preschool children. *The American Journal of Occupational Therapy*, *57*(5), 574–577. doi:10.5014/ajot.57.5.574

Clark, J. E., & Metcalfe, J. S. (2002). The mountain of motor development: A metaphor. *Motor Development: Research and Reviews*, *2*, 163–190.

Cools, W., De Martelaer, K., Samaey, C., & Andries, C. (2009). Movement skill assessment of typically developing preschool children: A review of seven movement skill assessment tools. *Journal of Sports Science & Medicine*, *8*, 154–168.

Croce, R., Horvat, M., & McCarthy, E. (2001). Reliability and concurrent validity of the movement assessment battery for children. *Perceptual and Motor Skills*, *93*, 275–280. doi:10.2466/pms.2001.93.1.275

De Vet, H. C. W., Beckerman, H., Terwee, C. B., Terluin, B., & Bouter, L. M. (2006). Definition of clinical differences. *The Journal of Rheumatology*, *33*(2), 434–434.

De Vet, H. C. W., Terwee, C. B., Knol, D. L., & Bouter, L. M. (2006). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology*, *59*(10), 1033–1039. doi:10.1016/j.jclinepi.2005.10.015

Gallahue, D., Ozmun, J., & Goodway, J. (2012). *Understanding motor development: Infants, children, adolescents, adults* (7th ed.). New York, NY: McGraw-Hill Education.

Gesell, A. L. (1975). *Gesell and Amatruda's Developmental diagnosis: The evaluation and management of normal and abnormal neuropsychologic development in infancy and early childhood* (3rd ed.). Hagerstown, MD: Medical Dept, Harper & Row.

Henderson, S. E., Sugden, D. A., & Barnett, A. L. (2007). *Movement assessment battery for children-2 second edition [Movement ABC-2]*. London, UK: The Psychological Corporation.

Houwen, S., Hartman, E., Jonker, L., & Visscher, C. (2010). Reliability and validity of the TGMD-2 in primary-school-age children with visual impairments. *Adapted Physical Activity Quarterly*, *27*(2), 143–159. doi:10.1123/apaq.27.2.143

Jaakkola, T., Yli-Piipari, S., Huotari, P., Watt, A., & Liukkonen, J. (2016). Fundamental movement skills and physical fitness as predictors of physical activity: A 6-year follow-up study: Motor skills, fitness, and physical activity. *Scandinavian Journal of Medicine & Science in Sports*, *26* (1), 74–81. doi:10.1111/sms.12407

Kamm, K., Thelen, E., & Jensen, J. L. (1990). A dynamical systems approach to motor development. *Physical Therapy*, *70*(12), 763–775. doi:10.1093/ptj/70.12.763

Kielhofner, G. (2006). *Research in occupational therapy: Methods of inquiry for enhancing practice*. Philadelphia, PA: F.A. Davis.

Kiphard, B. J., & Schilling, F. (2007). *Körperkoordinationstest für Kinder 2. Überarbeitete und ergänzte Auflage*. Weinheim, Germany: Beltz Test GmbH.

Lamb, K. (1998). Test-retest reliability in quantitative physical education research: A commentary. *European Physical Education Review*, *4*(2), 145–152. doi:10.1177/1356336X9800400205

Lopes, V. P., Rodrigues, L. P., Maia, J. A. R., & Malina, R. M. (2011). Motor coordination as predictor of physical activity in childhood: Motor coordination and physical activity. *Scandinavian Journal of Medicine & Science in Sports*, *21*(5), 663–669. doi:10.1111/j.1600-0838.2009.01027.x

Lubans, D. R., Morgan, P. J., Cliff, D. P., Barnett, L. M., & Okely, A. D. (2010). Fundamental movement skills in children and adolescents. Review of associated health benefits. *Sports Medicine*, *40*(12), 1019–1035. doi:10.2165/11536850-000000000-00000

Meyer, U., Roth, R., Zahner, L., Gerber, M., Puder, J. J., Hebestreit, H., & Kriemler, S. (2012). Contribution of physical education to overall physical activity: Physical activity during physical education. *Scandinavian Journal of Medicine & Science in Sports*, n/a-n/a. doi:10.1111/j.1600-0838.2011.01425.x

Mokkink, L. B., Terwee, C. B., Knol, D. L., Stratford, P. W., Alonso, J., Patrick, D. L., … De Vet, H. C. (2010). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. *BMC Medical Research Methodology*, *10*(1), 22. doi:10.1186/1471-2288-10-22

Moore, J. B., Reeve, T. G., & Boan, T. (1986). Reliability of the short form of the Bruininks-Oseretsky Test of Motor Proficiency with five-year-old children. *Perceptual and Motor Skills*, *62*(1), 223–226. doi:10.2466/pms.1986.62.1.223

Morgan, P. J., Barnett, L. M., Cliff, D. P., Okely, A. D., Scott, H. A., Cohen, K. E., & Lubans, D. R. (2013). Fundamental movement skill interventions in youth: A systematic review and meta-analysis. *Pediatrics*, *132*(5), e1361–e1383. doi:10.1542/peds.2013-1167

Orsi, R., Drury, I. J., & Mackert, M. J. (2014). Reliable and valid: A procedure for establishing item-level interrater reliability for child maltreatment risk and safety assessments. *Children and Youth Services Review*, *43*, 58–66. doi:10.1016/j.childyouth.2014.04.016

Payne, V. G., & Isaacs, L. D. (2012). *Human motor development: A lifespan approach* (8th ed.). New York, NY: McGraw-Hill.

Portney, L. G., & Watkins, M. P. (2008). *Foundations of clinical research: Applications to practice* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86* (2), 420–428. doi:10.1037/0033-2909.86.2.420

Siedentop, D. L. (2009). National plan for physical activity: Education sector. *Journal of Physical Activity & Health*, *6* (2), SS168–S180. doi:10.1123/jpah.6.s2.s168

Singer, R. N. (1980). *Motor learning and human performance : An application to motor skills and movement behaviors* (3rd ed.). New York, NY: MacMillan.

Stratford, P. W. (2004). Getting more from the literature: Estimating the standard error of measurement from reliability studies. *Physiotherapy Canada*, *56*(1), 027. doi:10.2310/6640.2004.15377

Streiner, D. L., Norman, G. R., & Cairney, J. (2014). *Health measurement scales: A practical guide to their development and use* (5th ed.). Oxford, UK: Oxford University Press.

Ulrich, D. A. (2000). *Test of gross motor development* (2nd ed.). Austin, TX: PRO-ED.

Van Gelder, W., & Stroes, H. (2010). *Leerlingvolgsysteem Bewegen en Spelen. Over observeren, registeren en extra zorg [Pupil tracking system Moving and Playing. About observing, registering, and extra care]* (2nd ed.). Amsterdam, The Netherlands: Elsevier.

Van Kernebeek, W. G., De Kroon, M. L. A., Savelsbergh, G. J. P., & Toussaint, H. M. (manuscript submitted for publication). *The validity of the 4-skills scan: A double validation study*.

Van Waelvelde, H. V., 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*(5), 465–470. doi:10.1177/0269215507074052

Vandorpe, B., Vandendriessche, J., Lefevre, J., Pion, J., Vaeyens, R., Matthys, S., … Lenoir, M. (2011). The KörperkoordinationsTest für Kinder: Reference values and suitability for 6-12-year-old children in Flanders. *Scandinavian Journal of Medicine & Science in Sports*, *21*(3), 378–388. doi:10.1111/j.1600-0838.2009.01067.x

Wiart, L., & Darrah, J. (2001). Review of four tests of gross motor development. *Developmental Medicine & Child*

Neurology, *43*(4), 279–285. doi:10.1017/S0012162201000536

Wrotniak, B. H., Epstein, L. H., Dorn, J. M., Jones, K. E., & Kondilis, V. A. (2006). The relationship between motor proficiency and physical activity in children. *Pediatrics*, *118*(6), e1758–e1765. doi:10.1542/peds.2006-0742

Wuang, Y., Su, J., & Su, C. (2012). Reliability and responsiveness of the movement assessment battery for children-second edition test in children with developmental coordination disorder. *Developmental Medicine & Child Neurology*, *54*(2), 160–165. doi:10.1111/j.1469-8749.2011.04177.x

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*(5), 469–478. doi:10.1097/00005650-199905000-00006

Appendix

SCORE SHEET 4-SKILLS SCAN Name _____ Date of birth _____

	Level -I 2.0 yrs	Level 0 3.0 yrs	Level I 4.0 yrs	Level II 5.0 yrs	Level III 6.0 yrs	Level IV 7.0 yrs	Level VI 9.0 yrs	Level VIII 11.0 yrs	Level X 13.0 yrs
Standing still	can step over 4 cm ☐	can shoot a ball without falling ☐	stand on one leg for 3 s ☐	can stand on one leg for 10 s (wobbling allowed) right ☐ left ☐	can stand on one leg for 10 s (stable) right ☐ left ☐	can stand on one leg for 30 s (wobbling allowed) right ☐ left ☐	can stand on one leg for 30 s (stable) right ☐ left ☐	can stand on one leg for 60 s (stable) right ☐ left ☐	can stand on one leg with eyes closed for 10 s right ☐ left ☐
Jumping-Force	steps from a 20 cm high platform ☐	jumps from a bench (30 cm) and stands still ☐	hops 3 times ☐	hops with preference leg 10 times ☐	hops with non-preference leg 10 times ☐	hops over 9 m distance 11 x ☐ 12 x ☐	hops over 9 m distance 9 x ☐ 10 x ☐	hops over 9 m distance 7 x ☐ 8 x ☐	hops over 9 m distance 5/6 x ☐ 6/7 x ☐
Jumping-Coordination	tramples when child wants to jump ☐	jumps with 2 legs synchronously ☐	jumps forward as a kangaroo 3 times ☐	skips ☐	makes ski jump 10 times ☐	makes quick shuffle-jump ☐	can skip and clap in hands synchronously ☐	can cross-spread-cross jump + clap hands at cross ☐	can cross-spread-cross jump + clap hands at spread ☐
Bouncing (ball)	hits (regularly) a well aimed balloon ☐	keeps balloon in the air 3-5 times ☐	keeps balloon in the air 6 times ☐	drops - bounce - catch ☐	bounce with preferred hand 15 times right ☐ left ☐	bounce with non-preference hand 15 times ☐	can dribble a figure 8 10 times in 30 s ☐	bounce without looking at ball 15 times right ☐ left ☐	can dribble a figure 8 12 times in 30 s ☐