CONFIDENTIAL

Accepted for publication in Journal of Motor Development and Learning

Reliability Assessment of Scores from Video-Recorded TGMD-3 Performances

Rintala, P., Sääkslahti, A. & Iivonen, S.

Abstract

This study examined the intrarater and interrater reliability of the *Test of Gross Motor Development—Third Edition* (TGMD-3). Participants were 60 Finnish children aged between 3 and 9 years, divided into three separate samples of 20. Two samples of 20 were used to examine the intrarater reliability of two different assessors, and the third sample of 20 was used to establish interrater reliability. Children's TGMD-3 performances were video recorded and later assessed using an intraclass correlation coefficient, a kappa statistic and a percent agreement calculation. The intrarater reliability of the locomotor subtest, ball skills subtest, and gross motor total score ranged from 0.69 to 0.77, and percent agreement ranged from 87% to 91%. The interrater reliability of the locomotor subtest, ball skills subtest, and gross motor total score ranged from 0.56 to 0.64. Percent agreement of 83% was observed for locomotor skills, ball skills, and total skills, respectively. Hop, horizontal jump and two-hand strike assessments showed the most difference between the assessors. These results show acceptable reliability for the TGMD-3 to analyze children's gross motor skills.

Key words: Children, Early childhood, Motor development, Pediatrics

Introduction

Fundamental motor/movement skills (FMS) are needed to manage motor challenges generated by everyday life (Gallahue, Ozmun, & Goodway, 2012). Gallahue et al. (2012) divided motor skills into balance skills (e.g., balancing on one foot), locomotor skills (e.g., walking, running, and hopping), and manipulative skills (e.g., ball handling skills). These FMS create a foundation for children to learn more specific skills necessary for games or different sport activities (Gallahue et al., 2012). Children's motor competence becomes visible through children's FMS performances, and FMS performance is positively associated with children's physical activity level (Stodden et al., 2008). However, many children's motor competence and physical activity levels are low (Reilly, 2010; Roth et al. 2010). Therefore, it is important to follow the development and level of children's motor competence using valid and reliable observational tools to measure children's motor competence. Psychometrically valid tools will help researchers and teachers monitor change, the impact of interventions, and the impact of policies. Moreover, measurement tools are needed not only for diagnostic purposes but also to find associations between and understand the significance of motor skills in overall development, daily wellbeing, and health (Robinson et al. 2015). The importance of regular data collection of FMS and the fact that test choice depends on specific purpose of test use is well justified in the study by Cools, Martelaer, Samaey, and Andriens (2009), who analyzed seven different movement skill measurements. In addition, cultural comparisons require measurement tools that are not overly sensitive to cultural differences (Cools et al., 2009).

When doing research with children, ethical aspects require careful consideration. Observation as a research method is unobtrusive and is thus warranted as an ethical method for use with children. Unfortunately, the reliability of observational tools is in question. Earlier studies on FMS have used either live assessments or video recordings. The *Test of Gross Motor Development—Second Edition* (TGMD-2) (Ulrich, 2000) was used in a study by Barnett, Minto, Lander, and Hardy (2014). They reported reliability based on live observation for interrater reliability in six object-control skills. Specifically reliability as intraclass correlation coefficient (ICC) for object control skills was 0.93, varying in individual skills from 0.71 (catch) to 0.94 (dribble). Another study by Slotte, Sääkslahti, Metsämuuronen, and Rintala (2015) analyzed children's motor skills through video recordings and reported intrarater reliability for 24 children's motor skills. In their study, reliability as ICC was 0.978 for

locomotor skills and 0.995 for object-control skills. Additional reliability studies will provide valuable information for test developers about the characteristics of such tests and will inform future test development.

The *Test of Gross Motor Development—Third Edition* (TGMD-3) (Ulrich, 2013), which was used in this study, is a process-oriented measurement, wherein children's FMS performances are observed and scored by a rater. The TGMD-3 is a new version of the TGMD-2; it gathers observations of both locomotor and object-control (called ball skills) FMS skills, but it differs from TGMD-2 in some individual skill components (Ulrich, 2013). In locomotor skills, leaping is replaced with skipping, and in ball skills, underhand roll is replaced with underhand throwing. Moreover, forehand strike is added, for a total of six locomotor skills and seven ball skills. As in the TGMD-2, the scores for each skill are based on the sum of either the presence or absence of the skill performance criteria (3–5 criteria depending on the skill). A more precise description of this tool can be found in another article (see Ulrich, 2013).

The TGMD-3, similar to its earlier version, is likely to be used by different professionals in practical settings such as schools (Cools et al., 2009), but also for research purposes when collected data must be as reliable as possible (Ulrich, 2013). Video recordings allow more detailed scrutiny and flexibility when carrying out assessments. Videos may be replayed several times if needed, and slow speed replay can assist the observation of performance criteria that are difficult to observe without slow motion. Finding the most and least challenging skills to score from video reliably also helps practitioners prepare for live observations. In addition, reliability values found for the TGMD-2 using either video recordings or live observations are not necessarily applicable to the TGMD-3 (Ulrich, 2013).

The purpose of this study was to assess the reliability of the TGMD-3 when used with video-recorded performances. First, the consistency of ratings within two independent assessors and, secondly, the consistency of ratings between two different assessors in each of the TGMD-3 individual skills were studied. In addition, the performance criteria shown to be the most challenging to rate consistently underwent a more detailed analysis.

Methods

Participants and Settings

Participants of this study were randomly selected from a larger study conducted among children at six elementary schools and eight day care centers/kindergartens (n = 374, 3–10 years) who had performed the TGMD-3 in Central Finland. The performances of 40 children were used to study intrarater reliability of two assessors (A and B). Assessor A analyzed performances from 10 boys, ranging from 6 to 9 years (M = 7.8 ± 1.2) and 10 girls, ranging from 5 to 9 years (M = 7.4 ± 1.2). Assessor B analyzed performances eight boys, ranging from 4 to 7 years (M = 6.6 ± 1.4) and 12 girls, ranging from 3 to 7 years (M = 6.1 ± 1.6). The performances of an additional 20 children were randomly chosen for interrater reliability. These children included 10 boys, ranging from 4 to 6 years (M = 5.9 ± 0.7) and 10 girls, ranging from 5 to 6 years (M = 6.2 ± 0.5). Institutional approval of the research protocol and informed consent from parents were obtained prior to the study, which was approved by the university ethics committee. All children also had the right to refuse to participate and to refrain from testing at any time. None of the assessed children had a known disability and/or impairment.

Procedure and Data Collection

All trials were conducted in school gymnasiums or similar locations that were suitable for the administration of the TGMD-3 according to the test instructions. In a few cases, the space did not allow the full running distance suggested by the test instructions. Children performed the TGMD-3 under the administration of two professionals, a trained physical education professional (one of the authors) and one graduate student (five altogether).The professionals were intimately familiar with administering the TGMD-2 and had used the test before, and the students (five altogether) received a two-hour training on how to administer the test. One of the two test administrators instructed the performer and the other video recorded the performance. The camera was placed optimally (i.e., side view, frontal view, or rear view) to best detect the particular skill performance whenever circumstances permitted. The skills were administered in the order of the scoring sheet, as depicted in Table 1. Preceding assessment, an accurate demonstration of the skill was performed by the test administrator. Participants were tested in groups of 3 or 4 and were

given one practice trial to ensure that the child understood what to do. One additional demonstration was given if a child did not seem to understand the task. Each participant performed two trials individually for each gross motor skill.

Two physical education teachers with master's degrees (different from the test administrators) assessed the test performances from the videos. Both teachers had a good knowledge base about children's motor skills and experience assessing the motor skills of several hundred children using TGMD-3. These assessors also participated in a two-hour training session organized by the first author for elaborating on the performance criteria. The physical education teachers had also established 80% reliability in scoring with the TGMD-3 author through electronic videos. The scoring system used to rate performances was as follows: a score of 1 meant the criterion was performed accurately, and a score 0 meant the criterion was not performed accurately or not performed at all.

To determine intrarater reliability, first, the two assessors both coded the skill performances of 20 children twice. Approximately three months elapsed before their second coding. Using these two sets of scores, the ability of both assessors to score the performance criteria of the 13 individual skills similarly between the first and second evaluation was analyzed. To determine interrater reliability the two assessors (A and B) independently coded the same videos from 20 children.

Statistical Analysis

To determine intrarater and interrater reliability, intraclass correlation coefficients (ICC) (a one way model for consistency for single measures), a kappa statistic (Cohen, 1960) and a percent agreement calculation were used. For ICC the following guidelines were used: when $r < .40$, the level of significance is poor; between .40 and .59, it is fair; between .60 and .74, it is good; and $\geq .75$, it is excellent (Cicchetti, 1994). As in a previous study (Barnett et al. 2014) assessing the reliability of measuring children's gross motor skills with TGMD-2, we used the magnitudes suggested for kappa by Landis and Koch (1977) for characterizing the resulting reliability statistics: a kappa statistic <0.20 was considered slight agreement; between 0.21 and 0.40, fair; between 0.41 and 0.60, moderate; and 0.61 and above was considered substantial agreement. Percent agreement was also calculated for each sub-skill. Significance level was set at 0.05. Data were analyzed using SPSS (version 22 for Windows).

Results

Intra- and interrater ICCs, kappa coefficients (κ) and corresponding percentages of agreement (% Agr.) of the assessments for individual skills, subtests of locomotor skills (LS), ball skills (BS), and gross motor test total score (TS) are provided in Table 1. The intrarater reliability ICCs and kappa coefficients of TS ranged from 0.73 to 0.75, suggesting good or excellent agreement. Similarly, ICC and kappa coefficients for LS and BS were also good or excellent (range from 0.69 to 0.77). Intrarater percent agreement for LS, BS, and TS varied from 87% to 91%. When the intrarater reliability for the individual skills was examined, all values were at least fair.

Table 1 about here

For interrater reliability, ICCs and kappa coefficients for LS, BS, and TS between the two assessors varied from fair to good, ranging from 0.56 to 0.64. Percent agreement for LS, BS, and TS were all 83% (Table 1).

Based on ICCs, kappa coefficients and/or percent agreement between the assessors, the individual skills most reliably scored were skip (ICC = 0.87, κ = 0.87, % Agr. = 93), two-hand catch (ICC = 0.84, κ = 0.84, % Agr. = 94), and one-hand stationary dribble (ICC = 0.81, κ = 0.81, % Agr. = 93). Three individual skills with the lowest reliability scores were hop (ICC = 0.13, κ = 0.19, % Agr. = 73), horizontal jump (ICC = 0.37, κ = 0.39, % Agr. = 79), and two-hand strike (ICC = 0.32, κ = 0.32, % Agr. = 72), characterized as poor level of consistency (Table 1).

A more detailed examination of these three skills with the lowest reliability scores was performed (Table 2). For the hop, these criteria were "arms flex and swing forward to produce force" (κ = 0.13, 63%) and "foot of non-hopping leg remains behind hopping leg" (43%). In the latter criterion, both raters scored the same number of 1s and 0s on the same criteria; therefore, the Kappa statistic could not be calculated for this

criterion. Also, for the fourth criterion, “hops four consecutive…”, assessor A scored all cases “1” in both trials and assessor B scored similarly except for one case, which again did not allow the kappa statistic to be calculated. However, the percent agreement in this criterion was high (98%). In assessing the horizontal jump, the most inconsistent performance criterion was “arms extend forcefully forward and upward reaching above the head” ($\kappa = 0.21$, 65%). In the two-hand strike, “preferred hand grips bat above non-preferred hand,” the kappa coefficient indicated slight ($\kappa = 0.07$, 60%) consistency between assessors (assessor B scored more 1s). Fair consistency was found in “non-preferred hip/shoulder faces straight ahead” ($\kappa = 0.31$, 83%) and in “steps with non-preferred foot” ($\kappa = 0.31$, 68%). In both criteria, assessor B scored more 1s, but in the first criteria, the assessors agreed on 83% of the cases.

Table 2 about here

Discussion

The main purpose of this study was to assess the intra- and interrater reliability of the TGMD-3 video performances of children from 3 to 9 years of age. The results as intraclass correlation coefficients and percent agreement showed good or excellent intrarater reliability and fair to good interrater reliability for locomotor skills, ball skills and total score. In terms of individual skill reliability, especially the interrater values, there was large variability among three skills (hop, horizontal jump, and two-hand strike) with poor ICC values. Those skills, in particular, appear to have some performance criteria that are challenging to assess.

The reliability values for the total score (ICC = 0.62 to 0.75) were good or excellent (Cicchetti, 1994). Moreover, percent agreement ranged from 83 to 91 percent. These high values were expected from assessors A and B, who had established reliability with an expert before they began the analysis.

All the children's performances were recorded on videos. Although the test protocol does not assume videotaping, in this case it allowed assessors to score the same performance twice and to compare their own scoring of the same children. Similarly, videotaping has been successfully used in earlier studies (Rintala & Linjala, 2003; Parkkinen & Rintala, 2004; Rintala & Loovis, 2013) with earlier TGMD versions. Analyzing videotaped performances has pros and cons: it allows several viewings to decide whether the criteria were met, but it is time consuming, and it is not suitable for every day school or daycare life evaluation. However, it is good for research purposes, since one can re-analyze the data if necessary.

When looking at specific individual skills such as 'two-hand strike on a stationary ball' (Table 1), we see a large difference between the intrarater ICC values of assessors A and B (0.84 vs. 0.47) and percent agreement (94% vs. 80%). Especially their interrater reliability values (ICC = 0.32; % Agr = 72) showed that they scored this skill differently. In this case, one potential challenge arises from not establishing the child's preferred hand; as a result without knowing the preferred hand, the assessor is unable to determine the score on the first criterion, "child's preferred hand grips bat above non-preferred hand." Similar challenges might have been faced by Barnett et al. (2014). Their interrater values for different performance criteria of two-hand strike varied from 0.73 to 0.92 (ICC), from 0.27 to 0.92 (kappa), and agreement percentages ranged from 78 to 97.

The interrater reliability scores of this study showed that hop, two-hand strike, and horizontal jump were the most challenging skill performances for two different assessors to observe and interpret unambiguously. The ICC value (0.13) and kappa value (0.19) for the hop was the lowest of all the skills. It was also supported by a low percent agreement (73%). These low values may originate from the criterion "foot of non-hopping leg remains behind hopping leg," which may be hard to observe if the skill is not yet automated. The difference may also come from the fact that one assessor may interpret the criterion literally, (i.e., another foot cannot pass the other leg at any point during hopping, while another assessor may think that if the foot stays behind for most of the time it is acceptable). Similarly low values were found for "arms flex and swing forward to produce force" when there are different kinds of 'flexed arms' and the pendulum movement varies in length.

The two-hand strike also had some performance criteria with poor or fair interrater reliability values, especially in "preferred hand grips bat above non-preferred hand" ($\kappa = 0.07$; 60%), that might indicate that it was sometimes difficult to see whether the criterion was fulfilled. It was not always possible even upon watching the video to decide which hand gripped above the other. Sometimes, especially among younger children, their hands were on top of each other, making the decision difficult. However, there was no indication of similar difficulties in Barnett et al.'s (2014) study, in which "hip and shoulder rotation during swing" had the lowest kappa values (0.27 and 0.32). It is notable that they used live observation.

In the horizontal jump, the criterion "arms extend forcefully forward and upward reaching above the head" produced the lowest kappa (0.21). In this case, the assessors among themselves may have set different limits for acceptable performance, (i.e., it is acceptable if hands are at the height of the face, or both hands need to reach above head as directed in the criterion).

The study by Barnett et al. (2014) revealed that low kappa values may not necessarily mean low values of agreement. In our study, those two values, however, seem to reflect one another. Namely, the lowest kappa values, as presented above, correspond to the same lowest percent agreement values. This distinct phenomenon needs more research to be more fully understood. Differences between the study of Barnett et al. (2014) and the current study may be explained, for example, through the scoring protocol and the children's different skill levels. Namely, it is easier to give accurate scores when a child's skill performance level is high in comparison to scoring children who are just learning the skill. The similarity of these two values in our study may be due to the position of the video camera. From an ecological validity standpoint, it is necessary to disturb children as little as possible. In this study, this meant that the position of the video camera was as constant as possible. This may have made it difficult to see all body movements as precisely as in a live observation situation. In live observation, the observer may change his/her visual angle naturally, without disturbing children's performance. In general, it can be assumed that two assessors, even with similar training backgrounds, will always have slightly different views, experience, and potential to assess motor skills.

The test instructions and the criteria used to assess fundamental movement skills of children should be unambiguous, easy to use even by non-professionals, and simple enough that the test will actually be used in daily routines. The TGMD-3 has potential to serve in this capacity all over the world, not just in the United States, where it already has a 30-year established reputation. With the development of several national norms for other countries, the test will grow in popularity and find its way into practitioners' tool kits.

Ecological validity was a strength of this study. Children's movement skills could be measured at the child's own childcare center, kindergarten, or school with familiar educators around them. Children felt comfortable and the testing situation did not cause extraordinary stress. The two independent assessors in this study were not aware of the research questions and carried out their observations and scoring based on their understanding of the performance criteria.

In the analysis from the videos, it was possible to use slow-speed replays of the test performances. Afterwards when discussing the skills that were challenging to score, the assessors realized that they utilized the videos differently in some instances. Assessor A reported using slow-speed replays when assessing especially young children and in unclear situations in specific skills. Assessor A used slow speed replay on random occasions when an assessed child was very young (3-4 years old), and/or child's performance skill was not yet 'fully' developed, and therefore more difficult to assess. In this interrater reliability sample there were no three year olds and only one four year old child. Assessor B only used the video at normal video speed. This was a limitation of the study and might have affected the interrater reliability ratings. In future video-based performance assessments, this speed replay option and its use should be determined before the beginning of the analysis.

Limited gym sizes in some childcare centers may be another limitation of the study. The size of the gym did not allow the full distance for running and galloping. During live observations, assessors may need the full distance to observe all criteria. On one hand, this problem can be minimized by videotaping, because the performance can be observed as many times as needed. On the other hand, it is difficult to change the angle of the camera in a small space, or there may only be one optimal location for the camera. In such situations, there will always be hidden spots and not all criteria may be visible.

This study demonstrated that TGMD-3 is a reliable and useful tool to analyze children's gross motor skills. The criteria are well described, and they can be learned quickly during a relatively easy familiarization period. When familiarizing assessors with different observation criteria, special attention should be paid to very quick movements such as those in the two-hand strike. Moreover, the criteria for hop and horizontal jump should be recognized as challenging to observe. Additional studies with different kinds of reliability analyses, based either on live observation or video recording, are needed to determine the most reliable gross motor skill measurement practices. In addition, studies addressing cultural differences in interpreting different performance criteria are warranted.

References

Barnett, L.M., Minto, C., Lander, N., & Hardy, L.L. (2014). Interrater reliability assessment using the Test of Gross Motor Development-2. *Journal of Science and Medicine in Sport, 17*, 667-670. doi:10.1016/jsams.2013.09.013

Cicchetti, D.V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284-290.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

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

Gallahue, D.L., Ozmun, J.C., & Goodway, J. (2012). *Understanding motor development: infants, children, adolescents, adults* (7th ed.). Dubuque, IA: McGraw-Hill.

Landis, J.R., & Koch, G.G. (1977). The measurement of observer for categorical data. *Biometrics, 33*, 159-174. doi:10.2307/2529310

Parkkinen, T., & Rintala, P. (2004). Primary school teachers' and physical education teachers' accuracy in assessing children's gross motor performance. *European Bulletin of Adapted Physical Activity, 3*. http://www.bulletin-apa.com/Brief_Communications.htm

Reilly, J.J. (2010). Low levels of objectively measured physical activity in pre-schoolers in child care. *Medicine & Science in Sports & Exercise, 42*, 502-507.

Rintala, P., & Linjala, J. (2003). Scores on test of gross motor development of children with dysphasia: A pilot study. *Perceptual and Motor Skills, 97*, 755-762.

Rintala, P., & Loovis, E.M. (2013). Measuring motor skills in Finnish children with intellectual disabilities. *Perceptual and Motor Skills, 116*, 294-303.

Robinson, L.E., Stodden, D.F., Barnett, L.M., Lopes, V.P., Logan, S.W., Rodrigues, L.P., & D'Hondt, E. (2015). Motor competence and its effect on positive developmental trajectories of health. *Sports Medicine, 45*, 1273-1284. doi:10.1007/s40279-015-0351-6

Roth, K., Ruf, K., Obinger, M., Mauer, S., Ahnert, J., Schneider, W., …Hebestreit, H. (2010). Is there a secular decline in motor skills in preschool children? *Scandinavian Journal of Medicine and Science in Sports, 20*, 670–678. doi:10.1111/j.1600-0838.2009.00982.x

Slotte, S., Sääkslahti, A., Metsämuuronen, J., & Rintala, P. (2015). Fundamental movement skills proficiency and body composition measured by dual energy X-ray absorptiometry in eight-year-old children. *Early Child Development and Care, 185*, 475-485.

Stodden, D., Goodway, J., Langendorfer, S., Robertson, M., Rudisill, M., & Garcia, C. (2008). A developmental perspective on the role of motor skill competence in physical activity: An emergent relationship. *Quest, 60*, 290–306.

Ulrich, D. (2000). *Test of Gross Motor Development* (2nd ed.). Austin, TX: Pro-ed.

Ulrich, D. (2013). The Test of Gross Motor Development-3 (TGMD-3): Administration, scoring, & international norms. *Hacettepe Journal of Sport Sciences, 24*(2), 27-33.

244 Table 1. Intra- and interrater reliability results for individual skills, subtests, and total scores for TGMD-3 videotaped test performance ratings.

	Intrarater: Rater A			Intrarater: Rater B			Interrater		
	κ	% Agr.	ICC (LB–UB)	κ	% Agr.	ICC (LB–UB)	κ	% Agr.	ICC (LB–UB)
Run	0.58	94 %	0.58 (0.47–0.68)	0.42	94 %	0.41 (0.28–0.53)	0.63	93 %	0.63 (0.53–0.72)
Gallop	0.80	93 %	0.80 (0.74–0.85)	0.77	89 %	0.77 (0.70–0.83)	0.62	82 %	0.61 (0.50–0.70)
Hop	0.51	92 %	0.51 (0.39–0.62)	0.62	82 %	0.62 (0.51–0.70)	[!]0.19	73 %	[!!]0.13 (0.0–0.28)
Skip	0.75	88 %	0.75 (0.66–0.82)	0.86	93 %	0.86 (0.81–0.90)	0.87	93 %	0.87 (0.81–0.91)
Horizontal jump	0.61	89 %	0.61 (0.51–0.70)	0.68	84 %	0.68 (0.58–0.75)	[!]0.39	79 %	[!!]0.37 (0.23–0.49)
Slide	0.58	89 %	0.58 (0.47–0.67)	0.61	87 %	0.61 (0.50–0.70)	0.45	80 %	0.45 (0.31–0.57)
Two-hand strike on a stationary ball	0.84	94 %	0.84 (0.79–0.88)	0.47	80 %	0.47 (0.36–0.57)	[!]0.32	72 %	[!!]0.32 (0.19–0.44)
One-hand forehand strike on self-bounced ball	0.70	86 %	0.70 (0.61–0.77)	0.73	86 %	0.73 (0.64–0.79)	0.64	82 %	0.64 (0.54–0.72)
One-hand stationary dribble	0.67	83 %	0.67 (0.56–0.76)	0.72	88 %	0.73 (0.63–0.80)	0.81	93 %	0.81 (0.74–0.87)
Two-hand catch	0.90	98 %	0.90 (0.86–0.93)	0.81	92 %	0.81 (0.74–0.87)	0.84	94 %	0.84 (0.78–0.89)
Kick a stationary ball	0.62	83 %	0.62 (0.52–0.71)	0.76	88 %	0.76 (0.69–0.82)	0.52	76 %	0.50 (0.37–0.60)
Overhand throw	0.84	95 %	0.84 (0.78–0.88)	0.68	84 %	0.68 (0.59–0.76)	0.65	83 %	0.65 (0.55–0.73)
Underhand throw	0.85	94 %	0.85 (0.80–0.89)	0.84	94 %	0.84 (0.79–0.88)	0.63	87 %	0.62 (0.52–0.71)
Locomotor Skills (LS)	0.69	91 %	0.69 (0.65–0.72)	0.73	88 %	0.73 (0.70–0.76)	0.57	83 %	0.56 (0.52–0.61)
Ball Skills (BS)	0.77	90 %	0.77 (0.75–0.79)	0.73	87 %	0.73 (0.70–0.76)	0.64	83 %	0.64 (0.61–0.68)
Total Skills (TS)	0.75	91 %	0.75 (0.73–0.76)	0.73	87 %	0.73 (0.71–0.75)	0.62	83 %	0.62 (0.59–0.65)

Note: Both assessors rated the same 20 children's performances for interrater reliability and each assessor rated another 20 children's performances twice for intrarater reliability. κ = kappa statistic; % Agr = percent agreement; ICC = Intraclass correlation coefficient; LB = Lower bound; UB = Upper bound; [!]Denotes fair or slight agreement; [!!]Denotes poor agreement

Table 2. Interrater reliability for the performance criteria of three individual skills for TGMD-3 videotaped test performance ratings

	κ	% Agr.
Hop		
1. Non-hopping leg swings forward in pendular fashion to produce force	0.72	90 %
2. Foot of non-hopping leg remains behind hopping leg (does not cross in front of)	-	43 %
3. Arms flex and swing forward to produce force	0.13	63 %
4. Hops four consecutive times on the preferred foot before stopping	-	98 %
Horizontal jump		
1. Prior to take off both knees are flexed and arms are extended behind the back	0.49	88 %
2. Arms extend forcefully forward and upward reaching above the head	0.21	65 %
3. Both feet come off the floor together and land together	0.45	75 %
4. Both arms are forced downward during landing	0.40	88 %
Two-hand strike on a stationary ball		
1. Preferred hand grips bat above non-preferred hand	0.07	60 %
2. Non-preferred hip/shoulder faces straight ahead	0.31	83 %
3. Hits ball sending it straight ahead	0.43	70 %
4. Hip and shoulder rotate and derotate during swing	0.46	80 %
5. Steps with non-preferred foot	0.31	68 %

Note: for the two Hop components, a κ could not be calculated as one rater only scored positively