Journal of Motor Learning and Development, 2023, 11, 225–244
https://doi.org/10.1123/jmld.2022-0068

Human Kinetics
ORIGINAL RESEARCH

Interrater Reliability of the Test of Gross Motor Development—Third Edition Following Raters' Agreement on Measurement Criteria

Aida Carballo-Fazanes,[1,2] **Ezequiel Rey,**[3] **Nadia C. Valentini,**[4] **Cristina Varela-Casal,**[1,3] **and Cristian Abelairas-Gómez**[1,2,5]

[1]CLINURSID Research Group, Universidade de Santiago de Compostela, Santiago de Compostela, Spain; [2]Simulation, Life Support, and Intensive Care Research Unit (SICRUS) of the Health Research Institute of Santiago de Compostela (IDIS), Santiago de Compostela, Spain; [3]REMOSS Research Group, Faculty of Education and Sport Sciences, University of Vigo, Pontevedra, Spain; [4]Department of Physical Education, Physical Therapy, and Dance, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil; [5]Faculty of Education Sciences, Universidade de Santiago de Compostela, Santiago de Compostela, Spain

We aimed to calculate interrater reliability of the Test of Gross Motor Development—Third Edition (TGMD-3) after raters reached a consensus regarding measurement criteria. Three raters measured the fundamental movement skills of 25 children on the TGMD-3 at two different times: (a) once when simply following the measurement criteria in the TGMD-3 manual and (b) after a 9-month washout period, following the raters' consensus building for the measurement criteria for each skill. After calculating and comparing the interrater reliability of these three raters across these two rating times, we found improved interrater reliability after the raters' consensus-building discussions on ratings of both locomotor skills (moderate-to-good reliability on two of six skills initially and at least moderate-to-excellent on four of six skills following criteria consensus building) and ball skills (moderate-to-good reliability on one of seven skills initially

Carballo-Fazanes https://orcid.org/0000-0001-6615-9821
Rey https://orcid.org/0000-0003-4770-2694
Valentini https://orcid.org/0000-0001-6412-5206
Varela-Casal https://orcid.org/0000-0001-6797-2670
Abelairas-Gómez (cristian.abelairas.gomez@usc.es) is corresponding author, https://orcid.org/0000-0002-1056-7778

and at least moderate-to-excellent reliability on four of seven skills following criteria consensus building). For subtest scores and overall test scores, raters achieved at least moderate-to-good reliability on their second, postconsensus-building ratings. Based on this improved reliability following consensus building, we recommend that researchers include rater consensus building before assessing children's fundamental movement skills or guiding curriculum interventions in physical education from TGMD-3 data.

Keywords: assessment, child development, motor competence, gross motor skills, TGMD

Motor development is a complex process that occurs throughout life, in which the acquisition and development of fundamental motor skills (FMS) are crucial (Goodway et al., 2021). After the reflexive movement phase (involuntary movements) and the rudimentary movement phase (first forms of voluntary movement), the FMS phase is the moment in which children explore and experiment the movement potential of their bodies. FMS are considered as the building blocks for more complex motor skills and movement patterns (Goodway et al., 2021). Thus, without a correct FMS development, children will not reach the minimum level of competence necessary to participate in many childhood physical activities (Stodden et al., 2008). This has led to an increase in the importance of assessing children's motor competence over recent decades (Bardid et al., 2019; Scheuer et al., 2019). FMS testing can be valuable for identifying children with low levels of motor competence, permitting comparisons of motor proficiency levels across different populations, or guiding appropriate school interventions to promote children's healthy development (Scheuer et al., 2019; Tamplain et al., 2020).

A raft of FMS assessment tools is available for clinical, educational, and research purposes (Eddy et al., 2020). These tools can be broadly classified into (a) quantity/product-oriented tests, those that offer quantifiable measurements of the product or outcome of children's movements (e.g., distance jumped); (b) quality/process-oriented tests, those that measure the quality of the movement process to determine whether children have yet attained some predefined behavioral criteria (e.g., judgments of arm movement quality during the jump); and (c) hybrid tests with scoring methods that combine both approaches (Bardid et al., 2019). Process-oriented tests can provide valuable qualitative information to guide teaching children how to accomplish a new motor movement (Barnett et al., 2020). However, scoring an observer's judgments of children's movements can be complex, and it may require raters to have extensive knowledge of FMS or specific skill training in their accurate assessment (Klingberg et al., 2019). Regardless of the type of FMS assessment tool (product oriented, process oriented, or hybrid), a prerequisite is that it has to be valid and reliable (DeVellis, 2012; Streiner & Norman, 2008). While validity refers to the appropriateness of the tool in a population of interest, reliability refers to the degree a test produces consistent results (Barnett et al., 2020).

Among process-oriented assessment tools, the Test of Gross Motor Development—Third Edition (TGMD-3; Ulrich, 2019) and its predecessors TGMD and TGMD-2 are based on rater observations that are organized into two subscales—locomotor and ball skills. A recent systematic review suggested that the TGMD, in its various editions, is one of the most frequently used tools for measuring children's FMS proficiency, and also that has been validated thoroughly in different settings (Scheuer et al., 2019). Despite the psychometric strengths of the results derived from the FMS assessment with TGMD-3, another systematic review found that interrater reliability values were generally lower than those observed for intrarater reliability (Rey et al., 2020), probably due to rater difficulty achieving a consensus in the interpretation of scoring criteria for some of the test's skill components (Barnett et al., 2014; Carballo-Fazanes et al., 2021; Houwen et al., 2010). Interrater differences relate to the rater's varied viewpoints, interpretations, and assessment methods. The TGMD-3 test manual provides clear instructions to rate whether a child meets certain performance criteria (Ulrich, 2019), but these judgments will always be subject to each rater's discretion (Cano-Cappellacci et al., 2015), unless there has been interrater consensus building that is, ideally, widely disseminated and shared.

Barnett et al. (2014) examined the interrater reliability of the TGMD-2 object control subtest by live observation and found some problematic, hard-to-identify definitions of performance criteria that need to be clarified and discussed among raters in a consensus-building process. For example, in the assessment of the overhand throw, they found low agreement on three of the four performance criteria referred to the *windup*, *rotation of hips and shoulders to a point where the nonthrowing side faces the wall* and *weight transfer*. In terms of rotation, it may be observed at different stages throughout the throw, which could explain the raters' disagreement (Barnett et al., 2014). Although the fact that FMS assessment has been carried out throughout live observation might be one of the causes for the lower interrater reliability in some performance criteria, this TGMD weakness might imply a need for more accurate TGMD measurement obtained by reducing the subjectivity bias of each rater with criteria consensus prior to assessment with this tool. In fact, studies aimed to analyze interrater reliability of the different versions of the TGMD described that raters established agreements to ensure scoring precisions (Allen et al., 2017), reviewed the performance criteria prior assessments (Maeng et al., 2017), or had been previously trained in TGMD assessment (Aye et al., 2017; Estevan et al., 2017). However, although training the raters or revision of the performance criteria may likely occur in practice prior to FMS assessment, the consensus-building process, and how this consensus is reached, is not reported in previous research. Thus, our aims in this study were as follows: (a) to report the consensus-building process of three raters regarding performance criteria of the TGMD-3 and (b) to compare interrater reliability before and after the consensus-building process in the same sample of school children.

Methods

Participants

Twenty-five healthy primary school children participated in this study as TGMD-3 examinees (15 girls and 10 boys: $M = 9.16$ years, $SD = 1.31$). We obtained informed written consent from all the children's parents or guardians, and we obtained informed verbal assent from the participants. This study followed the Helsinki Convention's ethical principles, and it was approved by the Ethical

Committee of the Faculty of Education and Sport Sciences (University of Vigo, Spain). In the rest of the manuscript, “participants” refers to school children.

Study Design

In the first stage, five raters used 13 video-recorded skills performed by 25 participants on the TGMD-3 to assess their motor competence (Carballo-Fazanes et al., 2021). Initially, each skill was explained, one by one, to the participants. Next, participants viewed a video, at normal speed and in slow motion, produced by the author of the TGMD, showing the correct execution of the skills (Ulrich & Webster, 2014). Participants then performed three trials of each skill: The first one was a practice trial permitting us to be sure they understood what they had to do, and the other two were video-recorded (camera Nikon D5300) so that the FMS could be measured by the raters later. This motor competence assessment relied upon the TGMD-3 manual guidelines (Ulrich, 2019) for rating the 13 skills.

In the second stage of this research, carried out 9 months later, three out of the five initial raters reached an agreement about assessing the performance criteria of each skill on the TGMD-3 following an interrater consensus-building process. Fifteen days later, the three raters measured the same 13 video-recorded skills of the same 25 participants included in the first stage of the study.

Raters

In the first stage, five raters (convenience sample) assessed the TGMD-3 skills performed by the 25 participants. Two raters were experts in TGMD assessment (more than 5 years of experience using TGMD) and the other three novices. Of these three, one had physical education background; one was a nurse with no physical education and sports sciences training, but PhD student in terms of physical literacy and motor competence; and the last one was a primary schoolteacher. Before assessing school children’s motor competence performance, novice raters reviewed the content of the TGMD-3 manual (Ulrich, 2019), and according to this, they practiced the assessment of three children (different from those of the 25 children whose FMS were assessed to study the interrater reliability).

In the second stage, the five raters were invited to participate, but just three were available to be included. This sample of three raters was composed of one of the expert raters and two novices (with physical education background and the PhD student). After reaching the agreement regarding the TGMD-3 performance criteria, the three raters assessed the motor competence of the same participants assessed in the first stage.

Interrater TGMD-3 Criteria Consensus Building

Between their initial TGMD-3 ratings of participants and their second TGMD-3 ratings of the same participants after a 9-month washout, our three raters discussed TGMD-3 criteria together and reached a consensus for how to best score the participants’ performance on the TGDM-3. The 9-month interval was determined according to the availability of the raters, since we wanted to guarantee a time interval appropriate between assessments but, at the same time, that once the second phase was initiated, the three raters could perform all the tasks related to this stage of the study uninterruptedly (reach the consensus, wait for a washout between reaching the agreements and the second assessment, and perform the second assessment). This was necessary to assuring that test conditions were similar for all raters.

First, each rater reviewed, individually, the performance criteria for the 13 skills (Figure 1). They had a repository of video recordings (different from those 25 children that participated in the present study) that they could use for the revision of the performance criteria. The raters took notes regarding those performance criteria that they considered more subjective. Subsequently, they met to discuss about their perceptions and notes regarding each performance criteria of each skill. They spent 4 hr, split into two halves: one half to discuss locomotor skills and another one about ball skills. They reached general agreements on factors that could apply universally to all skills, and they reached specific agreements that were applicable only to a particular skill (Tables 1 and 2). Finally, raters waited 2 weeks before rating the 25 children in the second stage of the study.

Assessment Measure—TGMD-3

The TGMD-3 is a process-oriented test for assessing young children's (aged 3–10 years) gross motor skill performance (Ulrich, 2019). It is organized into two

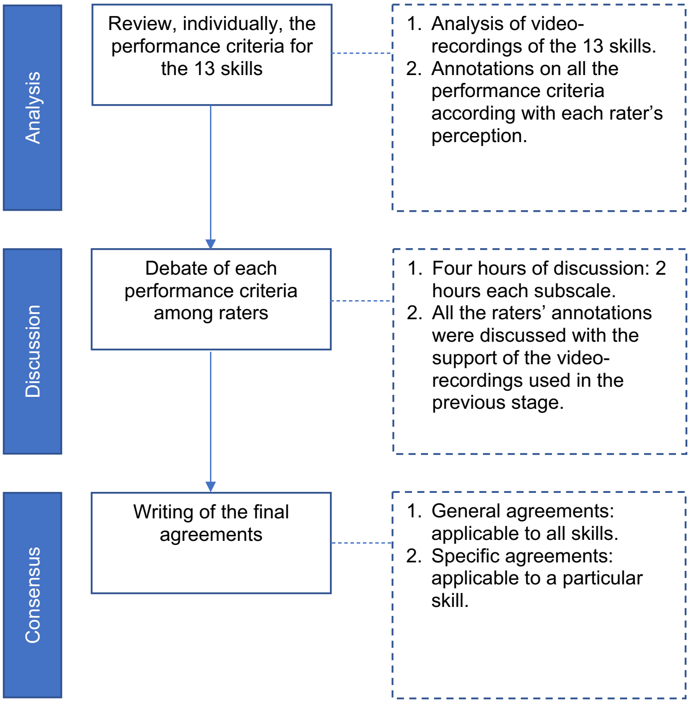
Analysis
Review, individually, the performance criteria for the 13 skills
1. Analysis of video-recordings of the 13 skills.
2. Annotations on all the performance criteria according with each rater's perception.
Discussion
Debate of each performance criteria among raters
1. Four hours of discussion: 2 hours each subscale.
2. All the raters' annotations were discussed with the support of the video-recordings used in the previous stage.
Consensus
Writing of the final agreements
1. General agreements: applicable to all skills.
2. Specific agreements: applicable to a particular skill.

Figure 1 — Consensus-building process.

Table 1 Consensus Reached Regarding Locomotor Skills

General agreements for all skills

1. In rhythmic skills with several patterns (i.e., gallop), performance criteria will be scored with "1" when is maintained during half (pair patterns) or half +1 (odd patterns) patterns.
2. If the rater has doubts in the assessment of a performance criterion, it will be scored with "1".

Specific agreements regarding locomotor subtest

Skill	Performance criteria	Consensus
Gallop	1. Arms flexed and swinging forward	Consider every takeoff, not just the first one.
	2. A step forward with lead foot, followed by the trailing foot landing beside or a little behind the lead foot (not in front of the lead foot)	If the trailing foot clearly surpasses the lead foot, this performance criterion will be scored with "0."
	3. Brief period where both feet come off the surface	—
	4. Maintains a rhythmic pattern for four consecutive gallops	If Criteria 1, 2, and 3 are scored "0," Criterion 4 will also be scored "0."
Slide	1. Body is turned sideways, so shoulders remain aligned with the line on the floor (score on preferred side only)	Body aligned with the direction of the movement.
	2. A step sideways with the lead foot followed by a slide with the trailing foot where both feet come off the surface briefly (score on preferred side only)	Although trailing foot touches the lead foot, this performance criterion will be scored "1."
	3. Four continuous slides to the preferred slide	—
	4. Four continuous slides to the nonpreferred slide	—

(continued)

Table 1 ***(continued)***

Specific agreements regarding locomotor subtest		
Skill	**Performance criteria**	**Consensus**
Hop	1. Nonhopping leg swings forward in pendular fashion to produce force	The knee of the nonhopping leg should surpass the trunk line with a pendular movement.
	2. Foot of nonhopping leg remains behind hopping leg (does not cross in front of)	If the foot of nonhopping leg clearly surpass the hopping leg, this performance criterion will be scored “0.”
	3. Arms flex and swing forward to produce force	—
	4. Hops four consecutive times on the preferred foot before stopping	—
Skip	1. A step forward followed by a hop on the same foot	—
	2. Arms are flexed and move in opposition to legs to produce force	—
	3. Completes four continuous rhythmical alternating skips	—
Horizontal jump	1. Prior to takeoff, both knees are flexed and arms are extended behind the back	Arms extended behind the trunk, taking glutes as reference.
	2. Arms extend forcefully forward and upward, reaching above the head	It is not necessary to reach a fully arm extension. If arms swing forward and hands reach the head height, this performance criterion will be scored “1.”
	3. Both feet come off the floor together and land together	—
	4. Both arms are forced downward during landing	During landing, arms should swing downward, not necessary to move behind the trunk. If performance Criterion 2 is scored “0,” this performance criterion too.
Run	1. Arms move in opposition to legs with elbows bent	If a criterion fails more than three times, this performance criterion will be scored “0.”
	2. Brief period where both feet are off the surface	
	3. Narrow foot placement landing on heel or toes (not flat-footed)	
	4. Nonsupport leg bent about 90°, so foot is close to buttock	

Table 2 Consensus Reached Regarding Ball Skills

General agreements for all skills

1. In rhythmic skills with several patterns (i.e., one-hand stationary dribble), performance criteria will be scored with “1” when is maintained during half (pair patterns) or half +1 (odd patterns) patterns.
2. If the rater has doubts in the assessment of a performance criterion, it will be scored with “1”.

Specific agreements regarding ball subtest

Skill	Performance criteria	Consensus
Two-hand strike of a stationary ball	1. Child’s preferred hands grips bat above nonpreferred hand	
	2. Child’s nonpreferred hip/shoulders face straight ahead	Alignment: Feet–hips–shoulders facing striking area.
	3. Hip and shoulder rotate and derotate during swing	Special attention in the hips. If Performance Criterion 2 is scored “0,” this performance criterion too.
	4. Steps toward ball with nonpreferred foot	If Performance Criterion 3 is scored “0,” this performance criterion too.
	5. Hits ball, sending it straight ahead	
Kick a stationary ball	1. Rapid, continuous approach to the ball	No walking and focus in continuous approach.
	2. Child takes an elongated stride or leap just prior to ball contact	—
	3. Nonkicking foot placed close to the ball	—
	4. Kicks ball with instep or inside of preferred foot (not the toes)	This performance criterion will be scored with “0” if clearly the ball is kicked with tiptoe or the foot side.
One-hand forehand strike of self-bounced ball	1. Child takes a backswing with the paddle when the ball is bounced	If shoulder starts in abduction, this performance criterion will be scored “0” because it there will not be “backswing.”
	2. Steps with nonpreferred foot	—
	3. Strikes the ball toward the wall	The direction of the ball follows the shoulders line.
	4. Paddle follows through toward nonpreferred shoulder	The movement of the paddle does not stop just after the strike. It continues toward nonpreferred shoulder.

(continued)

Table 2 *(continued)*

Specific agreements regarding ball subtest

Skill	Performance criteria	Consensus
Overhand throw	1. Windup is initiated with a downward movement of hand and arm	In the downward movement, the hand should reach behind the head.
	2. Rotates hip and shoulder to a point where the nonthrowing side faces the wall	If both feet are in a parallel position during the throwing, this performance criterion will be scored “0.”
	3. Steps with the foot opposite the throwing hand toward the wall	If the child starts with the preferred foot behind that opposite foot, this performance criterion will be scored “0.”
	4. Throwing hands follows through after the ball release, across the body toward the hip of the nonthrowing side	—
Underhand throw	1. Preferred hand swings down and back, reaching behind the trunk	—
	2. Steps forward with the foot opposite the throwing hand	If the child starts with the preferred foot behind that opposite foot, this performance criterion will be scored “0.”
	3. Ball tossed forward, hitting the wall without a bounce	—
	4. Hand follows through after ball release to at least chest level	—
Two-hand catch	1. Child’s hands are positioned in front of the body with the elbows flexed	It is not necessary that elbows are in 90°. Arms should surpass the trunk with a slight flexion of the elbows.
	2. Arms extend, reaching for the ball as it arrives	Arms move forward facing the ball.
	3. Ball is caught by hands only	—
One-hand stationary dribble	1. Contacts ball with one hand at about waist level	If contact is over the waist, this performance criterion will be scored “0.”
	2. Pushes the ball with fingertips (not slapping at ball)	The ball should be pushed, not beaten. If ball is slapped, this performance criterion will be scored “0.”
	3. Maintains control of the ball for at least four consecutive bounces without moving the feet to retrieve the ball	If feet move during the reception of the ball, this performance criterion will be scored “1” because it is after the fourth bounce.

subtests, measuring locomotor and ball skills. The locomotor subtest measures skills that require directional coordinated movements (run, gallop, hop, skip, horizontal jump, and slide). The ball skill subtest measures skills related to intercepting and propelling objects (two-hand strike, one-hand stationary dribble, overhand throw, kick, forehand strike, two-hand catch, and underhand throw). Each skill includes three to six performance criteria with each one scored as "0" or "1," depending on the criterion's absence or presence. Thus, a score is obtained for each skill, for each subtest, and the sum of these item skills from both subtests comprises the overall test score. Scores ranged from 0 to 46 points for the locomotor subtest, and 0 to 54 for the ball skill subtest, for an overall maximum score of 100.

Statistical Analyses

Interrater reliability was assessed for all raters (Rater A × Rater B × Rater C). In addition, pairwise analyses were also performed (Rater A × Rater B, Rater A × Rater C, and Rater B × Rater C). We used the intraclass correlation coefficient (ICC) to assess interrater reliability. Following Koo and Li (2016), we based ICC values and their 95% confidence intervals (CIs) on a single measurement (type to use the measurement from a single rater as the basis of the actual measurement), consistency (definition when the same group of subjects is correlated in an additive manner), and two-way random effects (model to generalize results to any raters with the same characteristics). This type of ICC was selected to account for systematic and random variance between and within raters (Maeng et al., 2017). Values <.50 indicated poor reliability, values between .50 and .75 indicated moderate reliability, values between .75 and .90 indicated good reliability, and values >.90 indicated excellent reliability. Interpretation of ICC is based on lower and upper bounds of the 95% CIs. For example, a .678 ICC (95% CI [.454, .798]) would be reported as poor-to-good reliability, since the lower bound is less than .50 (poor) and the upper bound is between .75 and .90 (good).

We performed all analyses using the Statistical Package for the Social Sciences (SPSS, version 23) and set the statistical significance level at $p < .05$.

Results

Interrater Reliability of Locomotor Skills

The three raters' interrater reliabilities (ICC and 95% CI) of the participants' locomotor skills are shown in Table 3. Run and hop skills had poor interrater reliability, both initially and after the raters' consensus-building process for performance criteria. For other locomotor skills, interrater reliability was higher after the consensus-building process in all pairwise rater comparisons. In the first stage, the interrater reliability calculated across the three raters' measurements was at least moderate in two skills (gallop and skip: moderate to good in both skills); however, on the second ratings following interrater consensus building, it was good-to-excellent for slide and skip and moderate-to-excellent for gallop and horizontal jump. Thus, scoring reliability for these three raters improved after the interrater consensus building.

Table 3 Interrater Reliability[a] Among Rater Locomotor Skills

Raters	Test	Run	Slide	Gallop	Hop	Horizontal jump	Skip	ICC and lower bound ≥ .5
Raters A × B	No-c	0.090 [−0.309, 0.462]	0.437 [0.058, 0.705]	*0.687* *[0.407, 0.849]*	0.251 [−0.152, 0.583]	*0.688* *[0.409, 0.849]*	**0.805** **[0.607, 0.909]**	1/6
	Co	0.209 [−0.796, 0.651]	**0.968** **[0.927, 0.986]**	**0.922** **[0.823, 0.966]**	**0.877** **[0.720, 0.946]**	**0.941** **[0.867, 0.974]**	**0.985** **[0.967, 0.994]**	5/6
Raters A × C	No-c	−0.106 [−0.474, 0.295]	**0.749** **[0.509, 0.881]**	**0.771** **[0.546, 0.892]**	0.470 [0.100, 0.726]	0.456 [0.082, 0.717]	*0.735* *[0.485, 0.874]*	2/6
	Co	−0.246 [−0.579, 0.157]	**0.958** **[0.905, 0.982]**	**0.877** **[0.722, 0.946]**	0.393 [−0.378, 0.732]	**0.877** **[0.720, 0.946]**	**0.904** **[0.781, 0.957]**	4/6
Raters B × C	No-c	*0.683* *[0.401, 0.847]*	*0.640* *[0.336, 0.824]*	*0.704* *[0.434, 0.858]*	0.423 [0.042, 0.697]	*0.591* *[0.263, 0.796]*	*0.655* *[0.358, 0.832]*	0/6
	Co	0.376 [−0.416, 0.725]	**0.922** **[0.822, 0.965]**	**0.878** **[0.724, 0.946]**	0.493 [−0.150, 0.777]	**0.913** **[0.801, 0.961]**	**0.939** **[0.861, 0.973]**	4/6
All raters	No-c	0.294 [0.050, 0.556]	*0.647* *[0.439, 0.810]*	**0.721** **[0.539, 0.854]**	0.413 [0.167, 0.650]	0.459 [0.216, 0.684]	**0.774** **[0.616, 0.884]**	2/6
	Co	0.094 [−0.761, 0.571]	**0.906** **[0.827, 0.954]**	**0.807** **[0.665, 0.902]**	0.485 [0.244, 0.703]	**0.836** **[0.711, 0.918]**	**0.894** **[0.807, 0.948]**	4/6

Note. Italic values, skills with ICC ≥ .50; bold values, skills with at least moderate interrater reliability (ICC and lower bound ≥ .50). Co = after raters' consensus; No-c = before raters' consensus; ICC = intraclass correlation coefficient.

[a] ICC: single measurement, consistency, two-way random-effects model, 95% confidence interval.

Interrater Reliability of Ball Skills

Interrater reliabilities (ICC and 95% CI) of ball skills are shown in Table 4. There was poor interrater reliability for kick, forehand strike, and underhand throw skills on the initial ratings before consensus building, but ICC values associated with these three skills overcame the .50 threshold in all analyses after the raters' consensus building. A particularly pronounced improvement was evident in the cases of forehand strike and underhand throw, for which interrater reliability after consensus building was, at least, moderate-to-excellent in all pairwise comparisons. Regarding interrater reliability calculated across the three raters' measurements, raters reached moderate-to-excellent reliability for three skills (two-hand strike, forehand strike, and underhand throw) and good-to-excellent reliability in one skill (overhand throw) after consensus building, while, initially, just one of the seven ball skills (two-hand strike) reached moderate reliability.

Interrater Reliability on TGMD-3 Subtest Scores and Overall Scores

Interrater reliabilities (ICC and 95% CI) of subtest and overall scores are shown in Table 5. Raters showed improvements when comparing initial ratings to second ratings 9 months later following consensus building. In the case of the locomotor subtest, there was improved reliability in all analyses from initial ratings to postconsensus-building ratings. Regarding the ball subtest, interrater reliability was only worse between Raters A and B (moderate-to-excellent) before the consensus compared to after the consensus (good-to-excellent). Reliability of overall scores remained the same for ratings made before and after the raters' consensus building around performance criteria.

Discussion

In this study, we aimed to determine interrater reliability among raters of children's TGMD-3 motor skills before and after raters' consensus building about performance skills' performance criteria. Across the three raters, at least moderate reliability was reached in four of the six locomotor skills after consensus-building process (two of six before consensus): slide, gallop, horizontal jump, and skip. Regarding ball skills, four of the seven reached at least moderate reliability (one of seven before consensus): two-hand strike, overhand throw, forehand strike, and underhand throw. Finally, interrater reliability improved from poor-to-good (before consensus building) to moderate-to-excellent (after consensus building) in the locomotor subscale score. Ball subscale score and overall score remained equal before and after consensus building.

More specifically, for locomotor skills, raters reached at least moderate interrater reliability when rating four out of the six skills after their consensus building, compared to reaching this level of interrater reliability for only two out of six skills on their initial ratings. After rater consensus building, the most improved ratings were on the locomotor skills of slide and horizontal jump skills, for which interrater reliability improved from poor-to-good (slide) and poor-to-moderate

Table 4 Interrater Reliability[a] Among Rater Ball Skills

Raters	Test	Two-hand strike	One-hand stationary dribble	Overhand throw	Kick	Forehand strike	Two-hand catch	Underhand throw	ICC and lower bound ≥ 0.5
Raters A × B	No-c	**0.788** **[0.576, 0.901]**	*0.631* *[0.322, 0.819]*	*0.660* *[0.365, 0.834]*	0.478 [0.110, 0.730]	*0.674* *[0.387, 0.842]*	**0.929** **[0.845, 0.968]**	0.475 [0.106, 0.729]	2/7
	Co	**0.972** **[0.937, 0.987]**	**0.900** **[0.772, 0.956]**	**0.920** **[0.819, 0.965]**	*0.730* *[0.387, 0.881]*	**0.932** **[0.846, 0.970]**	**0.852** **[0.693, 0.932]**	**0.799** **[0.545, 0.912]**	6/7
Raters A × C	No-c	*0.732* *[0.480, 0.872]*	*0.545* *[0.199, 0.770]*	0.476 [0.107, 0.729]	0.412 [0.028, 0.690]	0.378 [–0.012, 0.668]	0.297 [–0.103, 0.614]	0.322 [–0.076, 0.631]	0/7
	Co	**0.860** **[0.683, 0.938]**	0.393 [–0.377, 0.733]	*0.766* *[0.470, 0.897]*	**0.817** **[0.584, 0.919]**	**0.884** **[0.736, 0.949]**	*0.554* *[–0.012, 0.803]*	**0.817** **[0.584, 0.919]**	4/7
Raters B × C	No-c	*0.615* *[0.298, 0.810]*	0.477 [0.109, 0.730]	*0.570* *[0.234, 0.785]*	0.378 [–0.012, 0.668]	0.409 [0.024, 0.688]	0.222 [–0.181, 0.562]	0.377 [–0.014, 0.667]	0/7
	Co	**0.882** **[0.732, 0.948]**	*0.574* *[0.034, 0.812]*	*0.754* *[0.441, 0.891]*	*0.718* *[0.361, 0.876]*	**0.882** **[0.752, 0.946]**	*0.660* *[0.229, 0.850]*	**0.797** **[0.540, 0.911]**	3/7
All raters	No-c	**0.711** **[0.524, 0.848]**	*0.551* *[0.320, 0.748]*	*0.581* *[0.355, 0.767]*	0.413 [0.166, 0.650]	0.483 [0.242, 0.701]	*0.552* *[0.321, 0.749]*	0.397 [0.150, 0.638]	1/7
	Co	**0.848** **[0.731, 0.924]**	0.480 [0.238, 0.699]	**0.874** **[0.756, 0.941]**	*0.612* *[0.394, 0.788]*	**0.851** **[0.735, 0.926]**	*0.602* *[0.381, 0.781]*	**0.860** **[0.727, 0.934]**	4/7

Note. Italic values, skills with ICC ≥ .50; bold values, skills with at least moderate interrater reliability (ICC and lower bound ≥ .50). Co = after raters' consensus; No-c = before raters' consensus; ICC = intraclass correlation coefficient.

[a] ICC: single measurement, consistency, two-way random-effects model, 95% confidence interval.

Table 5 Interrater Reliability * Among Rater Subtests and Overall

Raters	Test	Locomotor	Ball	Overall	ICC and lower bound ≥ 0.5
Raters A × B	No-c	0.499 (0.137–0.743)	**0.787 (0.574–0.900)**	**0.820 (0.634–0.917)**	2/3
	Co	**0.912 (0.811–0.960)**	**0.912 (0.811–0.960)**	**0.945 (0.880–0.976)**	3/3
Raters A × C	No-c	*0.590 (0.261–0.796)*	*0.671 (0.382–0.840)*	**0.812 (0.618–0.912)**	1/3
	Co	**0.791 (0.582–0.902)**	*0.706 (0.438–0.859)*	**0.797 (0.592–0.905)**	2/3
Raters B × C	No-c	*0.621 (0.307–0.813)*	*0.618 (0.303–0.812)*	**0.813 (0.621–0.913)**	1/3
	Co	**0.758 (0.525–0.886)**	*0.693 (0.416–0.852)*	**0.766 (0.538–0.890)**	2/3
All raters	No-c	*0.569 (0.341–0.760)*	**0.695 (0.502–0.838)**	**0.815 (0.678–0.907)**	2/3
	Co	**0.824 (0.692–0.911)**	**0.778 (0.621–0.886)**	**0.842 (0.721–0.921)**	3/3

Note. Co = after raters' consensus; No-c = before raters' consensus; ICC = intraclass correlation coefficient.

*ICC: single measurement, consistency, two-way random-effects model, 95% confidence interval; italic values, skills with ICC ≥ .50; bold values, skills with at least moderate interrater reliability (ICC and lower bound ≥ .50).

(horizontal jump) initially to good-to-excellent (slide) and moderate-to-excellent (horizontal jump) on the second ratings. This improvement might demonstrate the importance, and value of having raters collectively objectifies the assessment criteria. The locomotor skills with the lowest reliability were run and hop in both stages, and other research efforts have also documented lower interrater reliability for judging these skills (Carballo-Fazanes et al., 2021; Maeng et al., 2017; Valentini et al., 2017), possibly due to the extra subjectivity or complexity in their measurement criteria. For instance, one criterion in the run is "*narrow foot placement landing on heel or toes (not flat-footed),*" which is difficult to perceive in each stride in the live assessments or even in video assessments, since the camera should be far enough away to record 20 m of running. Also, different prior studies have reported that the hop skill has the lowest interrater reliability on the TGMD-3 (Carballo-Fazanes et al., 2021; Rintala et al., 2017; Valentini et al., 2017). However, Maeng et al. (2017) found interrater reliability highest on the hop skill. On this skill, two performance criteria might be particularly subjective: "*Non-hopping leg swings forward in pendular fashion to produce force*" and "*Arms flex and swing forward to produce force.*" Especially the part of these criteria describing " … to produce force" requires raters to differentiate between swinging as a natural movement in balance and swinging specifically to produce force.

Previous research has also demonstrated assessment complexity for ball skills (Barnett et al., 2014), and, as a result, ball skills have shown poorer interrater reliability than locomotor skills (Carballo-Fazanes et al., 2021). However, in our study, even on these more complex skills, raters improved their interrater reliability from their initial ratings to after the consensus-building process. In this regard, agreement between the three raters was at least moderate just in one ball skill before the consensus-building process, reaching at least a moderate-to-excellent interrater reliability in four ball skills in the second stage of the study.

In the current study, the most problematic ball skills for understanding interrater reliability on the TGMD-3 were the one-hand stationary dribble, two-hand catch, and kick. Based on the existing literature, no agreement on the quality of interrater reliability can be established for these skills. For instance, Barnett et al. (2014) obtained lower reliability values in catch skill while the rater agreement for the kick was excellent. Consistent with our findings, other studies (Carballo-Fazanes et al., 2021; Y. Kim et al., 2012; Maeng et al., 2017; Rintala et al., 2017) found poor-to-moderate reliability for kick skill. Again, we consider that this could be due to the subjectivity of the criteria for this skill: "*Nonkicking foot placed close to the ball.*" What is considered "close to the ball?" In our study, some raters only considered children "close to the ball" only if their foot was just next to the ball, while others allowed some distance between the ball and the foot. These individualistic interpretations might contribute to rating variance and fluctuations in interrater reliability indices.

The same discordance applies to one-hand stationary dribble, for which our results indicated poor interreliability in both stages. Yet, the rater agreement was excellent in Maeng et al.'s (2017) research. On this skill, our raters expressed the greatest difficulty with the following criteria "*Pushes the ball with fingertips (not slapping at the ball)*" and "*Contacts ball with one hand at about waist level.*" Despite conducting video assessments and being able to stop and slow motion the action, our raters found such short time-lapse movements difficult to code. In this case, our lower interrater reliability index even after raters had a chance to build a performance criteria consensus might come from an incorrect application of one of their general agreements: "*If the rater has doubts in assessing a performance criterion, it will be scored as '1.'*" This general agreement was established considering that FMS is part of the child's gross motor development in which the movement assessment of specific performance criteria should not be as exacting as in the assessment of a technical movement of, for example, a sports performance.

Concerning the subtest scores, our results showed higher interrater reliability on these holistic indices than for the individual skills. Most previous studies also obtained good-to-excellent interrater reliability on locomotor and ball subtest scores and the overall score (Estevan et al., 2017; S. Kim et al., 2014; Lopes et al., 2016; Mohammadi et al., 2019; Simons et al., 2008; Valentini, 2012; Valentini et al., 2017; Wagner et al., 2016), and this level of interrater reliability was higher than in our study. In any case, reliability related to subtests scores, in our study, increased after consensus building from levels demonstrated before consensus building, especially on locomotor subtest scores that improved from poor-to-good to moderate-to-excellent interrater reliability. Interestingly, although interrater reliability increased in several skills and in the locomotor subscale, ball skill subscale and overall score remained with the same interrater reliability after the consensus building. Subscales' scores and overall score result from the sum of the score of each skill, whose scores are the sum of each performance criteria. Thus, the score of a subscale (or overall score) might be similar between raters with different ratings in the individual skills. It is important that physical education teachers, researchers, and all of those that are working in the FMS assessment field are aware about that since it may be considered a "double-edged sword." On one hand, agreement in the subscales' scores and overall scores do not mean necessarily agreement in the score of the skills; on the other hand, subscales' scores and overall scores are not discriminative enough for detecting the need for specific interventions in particular skills.

Our different findings from previous studies may be because we used more demanding descriptors for our ICC values (Koo & Li, 2016). For example, we only considered ICC values above .75 as good and those above .90 as excellent reliability. In contrast, from a qualitative perspective, other studies considered ICC values above .60 as good and above .75 as excellent (Capio et al., 2016; Estevan et al., 2017; Houwen et al., 2010; Mohammadi et al., 2019; Rintala et al., 2017). In effect, if we had interpreted the ICC the same as these earlier studies, our interrater reliability would have also been good-to-excellent on subtests and overall TGMD-3 scores. In addition, although some studies only used ICC values for interpreting reliability (Ayán et al., 2019; Aye et al., 2017; Barnett et al., 2014; Estevan et al., 2017; Houwen et al., 2010; S. Kim et al., 2014; Y. Kim et al., 2012; Maeng et al., 2017; Rintala et al., 2017; Valentini et al., 2017), we followed recommendations from Koo and Li (2016) and used both lower and upper values of the 95% CI, which avoids incomplete or confusing information providing the range in which each ICC lies; according to this, if we had not reported the CIs, assuming ICC values over .50 as moderate reliability, our results would show moderate reliability across the three raters in six of the seven ball skills (not only in four of seven). This, together with the fact that various investigators performed other statistical tests for assessing reliability, such as kappa (Lopes et al., 2016) or

Pearson's coefficient (Simons et al., 2008), might explain why interrater reliability was lower in the present study despite our raters' improved agreement. Although previous investigations stated the need for interrater consensus building before assessment (Barnett et al., 2014; Cano-Cappellacci et al., 2015; Houwen et al., 2010; Maeng et al., 2017), our study is the first to show the benefits of reaching a consensus between raters about performance criteria before they conducted FMS measurements.

Although our study showed better interrater reliability in some skills and locomotor subscale scores after than before the consensus-building process, we are aware that other raters might reach other agreements in other consensus-building process. As was already mentioned, scoring using process-oriented tools might be a complex task in which different factors such as experience and knowledge of the rater and subjective component of the performance criteria are involved. In addition, there are several tools that, assessing the same skills, tested them in different ways (Barnett et al., 2020). Therefore, if authors of tools that have shown reliable results in FMS assessment (experts in the field) understand the assessment of certain skills in different way from each other, the same can occur with the interpretation of the performance criteria by raters. In this sense, our results show the usefulness of reaching agreements before FMS assessment, considering the psychomotor learning assessments as a key component of the quality physical education (Donnelly et al., 2017), and that different physical education teachers might be working together in the same school. Moreover, both the aim of our study and the results found are in line and facilitate some of the principal uses of the TGMD-3, not only in the physical education field, but also in any field in which it is pretended to plan an instructional program in gross motor skill development or to evaluate the success of a gross motor program. Finally, although previous studies described that raters were trained or that they reached agreements before scoring children's motor competence, in the present study the process of consensus building is described and all the agreements are shown, which can be used, or even discussed, by professionals involved in FMS assessment field.

Limitations and Directions for Further Research

Some limitations to this study should be noted. First, we used a small sample of 25 healthy children as examinees (convenience sample), and our results should be replicated with a larger and more diverse sample (e.g., varying socioeconomic backgrounds and laterality preferences) and with a greater number of raters of different background experiences (e.g., experts and nonexperts in FMS assessment, teachers, and healthcare professionals). This would permit a more precise understanding of rater agreement for motor competence measurements with process-oriented test batteries. Also, of possible relevance, our 25 child participants were assessed by the raters twice. Since Ulrich (2019) suggested that raters may have introduced some memory bias after a 14-day washout period, we introduced an interval of 9 months in the present study to manage this potential confound. In addition, three raters participated in the second stage of this study. Although a recent systematic review (Rey et al., 2020) showed that interrater reliability of the different versions of the TGMD was assessed by two raters in the majority of the studies, this small number of raters has to be taken into account when interpreting the results.

Regarding general improvements in methodology that are needed in this line of research, the use of different statistical analyses across studies complicates comparing reliability results, and there is value in standardizing this approach. Of still greater importance, rather than a building rater consensus on performance criteria in separate studies, there is a need for common agreement and common training across all users of the TGMD-3 such that interrater reliability is not only improved within a research team but across research teams and clinical practitioners.

Finally, even after achieving satisfactory interrater agreement, questions arise as to how this consensus is sustained over time and when renewed training and consensus building may be needed. This information is relevant, especially in studies in which measurements of children's FMS span various periods, such as in intervention and longitudinal research.

Conclusions

In this study, we showed that rater's agreement on performance criteria might improve interrater reliability in assessing children's TGMD-3 skills. We recommend that in both research and clinical uses of the TGMD-3, to guide appropriate curriculum interventions in physical education, care should be taken to engage in consensus building to improve rater agreement about complex performance judgments on this subjective qualitative measure. Ideally, there should be a sharing of improved rating criteria across studies to allow for a common agreement and specific, standardized rater training for all users of the TGMD-3.

Acknowledgments

This work was supported by MCIN/AEI/10.13039/501100011033/ and European Regional Development Fund (ERDF), a way to make Europe, under grant PID2021-128640OB-I00. Carballo-Fazanes was supported by the Spanish Ministry of Science, Innovation and Universities under grant FPU19/02017. This study was approved by the Ethical Committee of the Faculty of Education and Sport Sciences of the University of Vigo, Spain.

References

Allen, K.A., Bredero, B., Van Damme, T., Ulrich, D.A., & Simons, J. (2017). Test of Gross Motor Development-3 (TGMD-3) with the use of visual supports for children with autism spectrum disorder: Validity and reliability. *Journal of Autism and Developmental Disorders, 47,* 813–833. https://doi.org/10.1007/s10803-016-3005-0

Ayán, C., Cancela, J.M., Sánchez-Lastra, M.A., Carballo-Roales, A.I., Domínguez-Meis, F., & Redondo-Gutiérrez, L. (2019). Reliability and validity of the TGMD-2 battery in a Spanish population. *Revista Iberoamericana de Diagnostico y Evaluacion Psicologica, 51*(1), 21–33. https://doi.org/10.21865/RIDEP50.1.02

Aye, T., Oo, K.S., Khin, M.T., Kuramoto-Ahuja, T., & Maruyama, H. (2017). Reliability of the test of gross motor development second edition (TGMD-2) for Kindergarten children in Myanmar. *Journal of Physical Therapy Science, 29*(10), 1726–1731. https://doi.org/10.1589/jpts.29.1726

Bardid, F., Vannozzi, G., Logan, S.W., Hardy, L.L., & Barnett, L.M. (2019). A hitchhiker's guide to assessing young people's motor competence: Deciding what method to use. *Journal of Science and Medicine in Sport, 22*(3), 311–318. https://doi.org/10.1016/j.jsams.2018.08.007

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*(6), 667–670. https://doi.org/10.1016/j.jsams.2013.09.013

Barnett, L.M., Stodden, D.F., Hulteen, R.M., & Sacko, R.S. (2020). Motor competence assessment. In T.A. Brusseau, S.J. Fairchlough, & D.R. Lubans (Eds.), *The Routledge handbook of youth physical activity* (pp. 384–408). Routledge (Taylor & Francis Group). https://doi.org/10.4324/9781003026426

Cano-Cappellacci, M., Leyton, F.A., & Carreño, J.D. (2015). Content validity and reliability of test of gross motor development in children. *Revista de Saude Publica, 49*, 97. https://doi.org/10.1590/S0034-8910.2015049005724

Capio, C.M., Eguia, K.F., & Simons, J. (2016). Test of gross motor development-2 for Filipino children with intellectual disability: Validity and reliability. *Journal of Sports Sciences, 34*(1), 10–17. https://doi.org/10.1080/02640414.2015.1033643

Carballo-Fazanes, A., Rey, E., Valentini, N.C., Rodríguez-Fernández, J.E., Varela-casal, C., Rico-Díaz, J., Barcala-furelos, R., & Abelairas-Gómez, C. (2021). Intra-rater (live vs. video assessment) and inter-rater (expert vs. novice) reliability of the test of gross motor development—Third edition. *International Journal of Environmental Research and Public Health, 18,* Article 1652. https://doi.org/10.3390/ijerph18041652

DeVellis, R.F. (2012). *Scale development: Theory and application* (3rd ed.). Sage Publications.

Donnelly, F.C., Mueller, S.S., & Gallahue, D.L. (2017). *Developmental physical education for all children*. Human Kinetics.

Eddy, L.H., Bingham, D.D., Crossley, K.L., Shahid, N.F., Ellingham-Khan, M., Otteslev, A., Figueredo, N.S., Mon-Williams, M., & Hill, L.J.B. (2020). The validity and reliability of observational assessment tools available to measure fundamental movement skills in school-age children: A systematic review. *PLoS One, 15*(8), Article e0237919. https://doi.org/10.1371/journal.pone.0237919

Estevan, I., Molina-García, J., Queralt, A., álvarez, O., Castillo, I., & Barnett, L. (2017). Validity and reliability of the Spanish version of the test of gross motor development-3. *Journal of Motor Learning and Development, 5*(1), 69–81. https://doi.org/10.1123/jmld.2016-0045

Goodway, J.D., Ozmun, J.C., & Gallahue, D.L. (2021). *Understanding motor development: Infants, Children, Adolescents, Adults* (8th ed.). Jones & Bartlett Learning.

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. https://doi.org/10.1123/apaq.27.2.143

Kim, S., Kim, M.J., Valentini, N.C., & Clark, J.E. (2014). Validity and reliability of the TGMD-2 for South Korean children. *Journal of Motor Behavior, 46*(5), 351–356. https://doi.org/10.1080/00222895.2014.914886

Kim, Y., Park, I., & Kang, M. (2012). Examining rater effects of the TGMD-2 on children with intellectual disability. *Adapted Physical Activity Quarterly, 29*(4), 346–365. https://doi.org/10.1123/apaq.29.4.346

Klingberg, B., Schranz, N., Barnett, L.M., Booth, V., & Ferrar, K. (2019). The feasibility of fundamental movement skill assessments for pre-school aged children. *Journal of Sports Sciences, 37*(4), 378–386. https://doi.org/10.1080/02640414.2018.1504603

Koo, T.K., & Li, M.Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lopes, V.P., Saraiva, L., & Rodrigues, L.P. (2016). Reliability and construct validity of the test of gross motor development-2 in Portuguese children. *International Journal of Sport and Exercise Psychology, 16*(3), 250–260. https://doi.org/10.1080/1612197X.2016.1226923

Maeng, H., Webster, E.K., Pitchford, E.A., & Ulrich, D.A. (2017). Inter- and intrarater reliabilities of the Test of Gross Motor Development—Third edition among experienced TGMD-2 raters. *Adapted Physical Activity Quarterly, 34*(4), 442–455. https://doi.org/10.1123/apaq.2016-0026

Mohammadi, F., Bahram, A., Khalaji, H., Ulrich, D.A., & Ghadiri, F. (2019). Evaluation of the psychometric properties of the Persian version of the test of gross motor development-3rd edition. *Journal of Motor Learning and Development, 7*(1), 106–121. https://doi.org/10.1123/JMLD.2017-0045

Rey, E., Carballo-Fazanes, A., Varela-Casal, C., & Abelairas-Gómez, C. (2020). Reliability of the test of gross motor development: A systematic review. *PLoS One, 15,* Article e0236070. https://doi.org/10.1371/journal.pone.0236070

Rintala, P.O., Sääkslahti, A.K., & Iivonen, S. (2017). Reliability assessment of scores from video-recorded TGMD-3 performances. *Journal of Motor Learning and Development, 5*(1), 59–68. https://doi.org/10.1123/jmld.2016-0007

Scheuer, C., Herrmann, C., & Bund, A. (2019). Motor tests for primary school aged children: A systematic review. *Journal of Sports Sciences, 37*(10), 1097–1112. https://doi.org/10.1080/02640414.2018.1544535

Simons, J., Daly, D., Theodorou, F., Caron, C., Simons, J., & Andoniadou, E. (2008). Validity and reliability of the TGMD-2 in 7-10-year-old Flemish children with intellectual disability. *Adapted Physical Activity Quarterly, 25*(1), 71–82. https://doi.org/10.1123/apaq.25.1.71

Stodden, D.F., Goodway, J.D., Langendorfer, S.J., Roberton, M.A., Rudisill, M.E., Garcia, C., & Garcia, L.E. (2008). A developmental perspective on the role of motor skill competence in physical activity: An emergen relationship. *Quest, 60*(2), 290–306. https://doi.org/10.1080/00336297.2008.10483582

Streiner, D.L., & Norman, G.R. (2008). *Health Measurement Scales: A practical guide to their development and use* (4th ed.). Oxford University Press.

Tamplain, P., Webster, E.K., Brian, A., & Valentini, N.C. (2020). Assessment of motor development in childhood: Contemporary issues, considerations, and future directions. *Journal of Motor Learning and Development, 8*(2), 391–409. https://journals.humankinetics.com/view/journals/jmld/8/2/article-p391.xml

Ulrich, D.A. (2019). *Test of gross motor development—3* (3rd ed.). Pro-Ed Publishers.

Ulrich, D.A., & Webster, E.K. (2014). *TGMD-3 administration*. https://www.youtube.com/watch?v=9WggHyZpXl0&ab_channel=KipWebster

Valentini, N.C. (2012). Validity and reliability of the TGMD-2 for Brazilian children. *Journal of Motor Behavior, 44,* 275–280. https://doi.org/10.1080/00222895.2012.700967

Valentini, N.C., Zanell, L.W., & Webster, E.K. (2017). Test of Gross Motor Development—Third edition: Establishing content and construct validity for Brazilian children. *Journal of Motor Learning and Development, 5*(1), 15–28. https://doi.org/10.1123/jmld.2016-0002

Wagner, M.O., Webster, E.K., & Ulrich, D.A. (2016). Psychometric properties of the test of gross motor development 3rd edition (German translation)—Results of a pilot-study. *Journal of Motor Learning and Development, 5*(1), 29–44. https://dx.doi.org/10.1123/jmld.2016-0006