PRE-PRINT

Inter-rater Reliability of the Test of Gross Motor Development – Third Edition (TGMD-3) Following Raters' Agreement on Measurement Criteria

Abstract

In this study, we calculated inter-rater 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 (FMS) of the same 25 children on the TGMD-3 at two different times, using two different methods: (i) once when simply following the measurement criteria in the TGMD-3 manual; and (ii) again, after a 9-month wash-out period, following the raters' consensus building for the measurement criteria for each skill. After calculating and comparing the inter-rater reliability of these three raters across these two rating times and methods, we found improved inter-rater reliability after the raters' consensus building discussions on ratings of both locomotor skills (moderate-to-good reliability on 2 of 6 skills initially and at least moderate-to-excellent on 4 of 6 skills following criteria consensus building) and ball skills (moderate-to-good reliability on 1 of 7 skills initially and at least moderate-to-excellent reliability on 4 of 7 skills following criteria consensus building). For subtests scores and on overall test scores, raters achieved at least moderate-to-good reliability on their second, post-consensus-building ratings. Based on this improved reliability following consensus building, we recommend that researchers include rater consensus-building before assessing children's FMS or guiding curriculum interventions in Physical Education from TGMD-3 data.

Keywords:

Assessment; child development; motor competence; gross motor skills; TGMD.

Introduction

The assessment of children's fundamental movement skills (FMS) has gained importance over recent decades (Bardid et al., 2019; Scheuer et al., 2019). FMS testing can be valuable for (a) 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 test, those that offer quantitifiable 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 (Stodden et al., 2008). However, scoring an observor'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).

Among process-oriented assessment tools, the Test of Gross Motor Development-Third Edition (TGMD-3; Ulrich, 2019) (TGMD), and its predecesors TGMD and TGMD-2, is based on rater observations (Bardid et al., 2019) that are organized into two subscales – locomotor and ball skills. A recent systematic review suggested that the TGMD, in its various editions, is the most frequently used tool for measuring children's FMS

proficiency in educational, research, and clinical settings (Rey et al., 2020). A prerequisite to the usage of any assessment tool is that it show adequate reliability and validity (DeVellis, 2012; Streiner & Norman, 2008), and, despite touting these psychometric strengths of the TGMD-3, Rey et al. (2020) also found in their review that inter-rater reliability values were generally lower than those observed for intra-rater reliability, 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; Houwen et al., 2010). Inter-rater 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; but these judgments will always be subject to each rater's discretion (Cano-Cappellacci et al., 2015), unless there has been inter-rater consensus-building that is, ideally, widely disseminated and shared.

Barnett et al. (2014) examined the inter-rater 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. 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. **Even after achieving satisfactory inter-rater 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.** Yet, to the best of the authors' knowledge, no study has compared the inter-rater reliability of TGMD after raters have reached agreement among themselves on

measurement criteria. Thus, our aim in this study was to determine inter-rater reliability of the TGMD-3 after raters reached their own consensus on performance criteria.

Method

Participants

Twenty-five healthy primary school children participated in this study as TGMD-3 examinees (15 girls, 10 boys; $M = 9.16$, $SD = 1.31$ years). 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 xxxx).

Study Design

In a first stage, five raters used 13 video-recorded skills performed by 25 participants on the TGMD-13 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. We studied both intra-rater and inter-rater reliability by comparing novel versus expert raters' ratings when viewing the video continuously without pauses and at normal speed to viewing the video in slow-motion as many times as the rater needed (Carballo-Fazanes et al., 2021). Intra-rater reliability was higher than inter-rater reliability (both calculated throughout intraclass correlation coefficient (ICC)), regardless of the evaluator's experience or viewing mode, presumably due to the individual raters' separate viewpoints (Carballo-Fazanes et al., 2021).

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

Assessment Measure - Test of Gross Motor Development—Third Edition [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 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 skills 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 3-6 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.

As noted, following a verbal description and video-based practical demonstration, the participants performed three trials of each skill, with the first trial an unscored practice trial to demonstrate the child's understanding of the skill, and the other two rater-scored trials, with "1" indicating correct performance and "0" indicating incorrect performance. Scores ranged from 0 to 46 points for the locomotor subtest, and 0 to 54 for the ball skills subtest, for an overall maximum score of 100.

Inter-Rater 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 wash-out period of their first ratings, our

three raters discussed TGMD-3 criteria together and reached a consensus for how to best score the participants' performance on the TGDM-3. Their initial ratings (first stage) relied upon the test manual guidelines for rating the 13 skills. In the second stage, **the** raters **made individual revisions** of the **performance criteria for the** 13 skills and then considered all their separate comments, skill by skill, and, through discussion together, established both general and specific agreements (Table 1 and Table 2). They reached general agreements on factors that could apply universally to all skills, and they reached specific agreements that were applicable to only to a particular skill.

[Insert Tables 1 and 2 near here]

Statistical Analyses

We used the intraclass correlation coefficient (ICC) to assess inter-rater reliability. Following Koo and Li (2016), we based ICC values and their 95% confidence intervals on a single-measurement (type), consistency (definition), and 2-way random-effects model. Values less than 0.50 indicated poor reliability, values between 0.50 and 0.75 indicated moderate reliability, values between 0.75 and 0.9 indicated good reliability, and values greater than 0.90 indicated excellent reliability. Interpretation of ICC is based on lower and upper bounds of the 95% confidence intervals.

We performed all analyses using the Statistical Package for the Social Sciences (SPSS, version 23, IBM Corporation, Chicago, IL), and we set the statistical significance level at $p < 0.05$.

Results

Inter-Rater Reliability of Locomotor Skills

The three raters' inter-rater reliabilities (ICC and 95% confident interval) of the participants' locomotor skills are shown in Table 3. Run and Hop skills had poor inter-rater reliability, both initially and after the raters' consensus-building process for performance criteria. For other locomotor skills, inter-rater reliability was higher after the consensus-building process in all pairwise rater comparisons. In the first stage, the inter-rater 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 inter-rater 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 inter-rater consensus building

[Insert Table 3 near here]

Inter-Rater Reliability of Ball Skills

Inter-rater reliabilities (ICC and 95% confident interval) of ball skills are shown in Table 4. Inter-rater reliability increased following the raters' discussions about TGMD-3 performance criteria during consensus building. There was poor inter-rater reliability for kick and underhand throw skills on the initial ratings before consensus building (ICC < 0.5 in all comparisons), but ICC values associated with these two skills overcame the 0.50 threshold in all analyses after the raters' consensus building**.** A particularly pronounced improvement was evident in the case of underhand throw for which inter-rater reliability after consensus building was moderate-to-excellent reliability in all pairwise comparisons. Regarding inter-rater 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 1 skill (overhand throw) after consensus building, while, initially, just 1 of the 7 ball skills (two-hand strike) reached moderate reliability.

[Insert Table 4 near here]

Inter-Rater Reliability on TGMD-3 Subtest Scores and Overall Scores

Inter-rater reliabilities (ICC and 95% confident interval) of subtest and overall scores are shown in Table 5. Raters showed improvements when comparing initial ratings to second ratings nine months later following consensus building. In the case of the locomotor subtest, there was improved reliability in all analysis from initial ratings to post-consensus building ratings (i.e., from "por to-good" reliability initially and "moderate to good" on second ratings. Regarding the ball subtest, inter-rater reliability was only worse between raters A and B (moderate-to-excellent) before the consensus comparing with after consensus (good-to-excellent). Reliability of overall scores remained the same for ratings made before and after the raters' consensus building around performance criteria.

[Insert Table 5 near here]

Discussion

In this study, we aimed to determine inter-rater reliability among raters of children's TGMD-3 motor skills before and after raters' consensus-building about performance skills' performance criteria. Inter-rater reliability improved after raters' consensus building on separate TGMD-3 locomotor skills and ball skills, and on TGMD-3 locomotor subtest scores and ball skills subtest scores.

More specifically, for locomotor skills, raters reached at least moderate inter-rater reliability when rating 4 out of the 6 skills after their consensus building, compared to reaching this level of inter-rater reliability for only 2 out of 6 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 inter-rater reliability improved from poor-to-good (slide) and poor-to-moderate (horizontal jump) initially to good-to-excellent (slide) and moderate-to-excellent (horizontal jump) on the second ratings. This significant improvement might demonstrate the importance, and value of having raters collectively objectify 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 inter-rater 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 meters of running. Also, different prior studies have reported that the hop skill has the lowest inter-rater reliability on the TGMD-3 (Carballo-Fazanes et al., 2021; Rintala et al., 2017; Valentini et al., 2017). However, Maeng et al. (2017) found inter-rater 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 inter-rater reliability than locomotor skills (Carballo-Fazanes et al., 2021). However, in our study, even on these more complex skills, raters improved their inter-rater reliability from their initial ratings to after the consensus building process. In this regard, agreement between the three raters were at least moderate just in 1 ball skill before the consensus building process, reaching at least a moderate-to-excellent inter-rater reliability in the second stage of the study.

In the current study, the most problematic ball skills for understanding inter-rater reliability on the TGMD-3 were the one-hand stationary dribble, two hands catch, and kick. Based on the existing literature, no agreement on the quality of inter-rater reliability can be established for these skills. For instance, Barnet 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 kicking. Again, we consider that this could be due to the subjectivity of the criteria for this "kill: "*Non-kicking 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 inter-rater reliability indices.

The same discordance applies to one-hand stationary dribble, for which our results indicated poor inter-reliability 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 inter-rater reliability index even after raters had a chance to build a performance criteria consensus might come have 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 a 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 inter-rater reliability on these holistic indices than for the individual skills. Most previous studies also obtained good-to-excellent inter-rater reliability on locomotor and ball subtests 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 inter-rater 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 inter-rater reliability.

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 0.75 as good and those above 0.90 as excellent reliability. In contrast, from a qualitative perspective, other studies considered ICC values above 0.6 as good and above 0.75 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 inter-rater 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% confidence interval. 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 inter-rater reliability was lower in the present study despite our raters' improved agreement. Although previous investigations stated the need for inter-rater 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.

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, 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 non-experts in FMS assessment, teachers, and health care 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 wash-out period, we introduced an interval of 9-months in the present study to manage this potential confound.

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 inter-rater reliability is not only improved within a research team but across research teams and across clinical practitioners.

Conclusion

In this study, we showed that rater's agreement on performance criteria might improve inter-rater 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.

References

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

Cano-Cappellacci, M., Leyton, F. A., & Carreño, J. D. (2015). Content validity and reliability of test of gross motor development in Chilean children. *Revista de Saude Publica*, *49*. 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*, 1652.

DeVellis. R.F. (2012). *Scale Development: Theory and Application* (3rd ed.), Sage Publications, Thousand Oaks, CA.

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), 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

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*(7 July), 1–26. 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, Oxford.

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. (2000). *Test of Gross Motor Development - 2* (Second Ed). Pro-Ed Publishers.

Ulrich, D. A. (2019). *Test of Gross Motor Development - 3* (Third 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.

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

6