Author(s): Rintala, Pauli; Sääkslahti, Arja; Iivonen, Susanna

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

Year: 2017

Version:

Please cite the original version:

Rintala, P., Sääkslahti, A., & 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

August 26, 2016

Reliability Assessment of Scores from Video Recorded TGMD-3 Performances

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. A kappa statistic and a percent agreement calculation were used. The results for intrarater reliability kappa coefficients for locomotor subtest, ball skills subtest and gross motor total score ranged from 0.69 to 0.77, and percent agreement from 87% to 91%. The interrater kappa coefficients for locomotor subtest, ball skills subtest and gross motor total score ranged from 0.57 to 0.64, which can be considered moderate to substantial reliability. Percent agreement for locomotor skills, ball skills, and total skills was 83%. Hop, horizontal jump and two-hand strike were the most differently assessed performance criteria between the assessors. The TGMD-3 showed to be reliable tool 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) defined such motor skills as 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 basis for children to learn more specific skills to participate in games or different sport activities (Gallahue et al., 2012). Children's motor competence becomes visible through children's FMS performances, and is positively associated to their physical activity level (Stodden et al., 2008). Therefore it is important to follow the development and level of children's motor competence through observing children's performances in different FMS. Today, as many children's motor competence and physical activity levels are low (Reilly, 2010; Roth et al. 2010), it is essential to find valid and reliable observational tools to measure children's motor competence. Having 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 and significance of motor skills for overall development, daily wellbeing and health (Robinson et al. 2015). This was well justified in the study by Cools, Martelaer, Samaey and Andriens (2009) who analyzed seven different movement skill measurements. In addition, cultural comparisons also need measurement tools that are not too sensitive to cultural differences (Cools et al., 2009).

When doing research with children, ethical aspects need careful consideration. Observation as a research method is unobtrusive and in that sense much warranted. Unfortunately, reliability of observational tools is questioned. Earlier studies have used either video recordings or live assessments. The TGMD-2 (Ulrich, 2000) was used in the Slotte, Sääkslahti, Metsämuuronen, and Rintala (2015) study. They analyzed children's motor skills through video recordings and reported intrarater reliability for 24 children's motor skills. In their study reliability as intraclass correlation (ICC) was 0.978 for locomotor skills and 0.995 for object-control skills. Another study by Barnett, Minto, Lander and Hardy (2014) also used the TGMD-2 version. They reported reliability based on live observation for interrater reliability in six object control skills. Specifically reliability for object control skills was 0.93 (ICC), varying in individual skills from 0.71 (catch) to 0.94 (dribble). All values reported are in the acceptable range. More reliability studies are needed to provide valuable information for test developers about the characteristics of the test for the future test development. For example, it cannot be assumed that the reliability values found for the TGMD-2 as such using either video recordings or live observations are applicable to the TGMD-3.

The TGMD-3, which was used in this study, is a process-oriented measurement, where children's FMS performances are observed and scored by a rater. The TGMD-3 is a new version of the TGMD-2, but also gathers observations of both locomotor and object control (called ball skills) FMS skills, but differs from TGMD-2 in some individual skill components (Ulrich, 2016). In locomotor skills leaping is replaced with skipping, and in ball skills underhand roll is replaced with underhand throwing. Moreover forehand strike is added which makes altogether six locomotor skills and seven ball skills. Similarly, as in the TGMD-2, the resulting score of each skill is based on the sum of either the presence or absence of the performance criteria (3–5 criteria depending on the skill) of that skill. A more precise description of this tool can be found in another article (see Ulrich, 2013).

The TGMD-3, as its earlier version, will probably be used by different professionals in practical settings such as at schools (Cools et al., 2009). It will also be used for research purposes when data must be as reliable as possible (Ulrich, 2016). Video recordings allow more detailed scrutiny and flexibility when doing assessments. Videos can also be replayed several times if needed, and slow speed replayed when the performance criteria is difficult to observe without slow motion. Finding the most and least challenging skills to score from video reliably also helps practitioners in preparation of their live observations.

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

Methods

Participants and Settings

Participants of this study were randomly selected from the larger study conducted with six elementary schools and eight day care center/kindergarten children (n = 374, 3–10 years) who had performed the TGMD-3 in Central Finland. Forty children's performances were used to study intrarater reliability of the two assessors (A and B). Participants of the assessor A were 10 boys, ranging from 6-9 years ($M = 7.8 \pm 1.2$) and 10 girls, ranging from 5-9 years (M =

7.4 ± 1.2). Participants of the assessor B were eight boys, ranging from 4-7 years (M = 6.6 ± 1.4) and 12 girls, ranging from 3-7 years (M = 6.1 ± 1.6). Another 20 children's (different from the previous 40 children) performances were randomly chosen for interrater reliability. These children were 10 boys, ranging from 4-6 years (M = 5.9 ± 0.7) and 10 girls, ranging from 5-6 years (M = 6.2 ± 0.5). Institutional approval of the research protocol and informed consent from parents were obtained prior to the study that was approved by the university ethics committee. All children had also the right to refuse participation and refrain from testing any time. None of the assessed children had a disability and/or impairment.

Procedure and Data Collection

All trials were conducted in the school gymnasiums or similar locations that were suitable for the administration of the TGMD-3 according to the test instructions. In few cases the space did not allow the full running distance according to the test instructions. Children performed the TGMD-3 administered by a trained physical education professional (one of the authors) and one Master's student in pairs. The professionals were very familiar with administering the TGMD-2 and had used the test before, and the students (five altogether) had had a two-hour training on how to administer the test. One of the two 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 skill performance whenever the 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 3–4, and were given one practice trial to assure 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 a Master's degree (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 had been assessing several hundred children on their motor skills using TGMD-3. These assessors had also participated in a two hour training session organized by the first author for elaborating performance criteria. They had also established 80% reliability in scoring with the TGMD-3 author through electronic videos. In rating performances, the scoring system was the following: 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 20 children's skill performances twice. There was about three months' time interval before their second coding. Secondly, both assessors were analyzed on their own ability to score the performance criteria of the 13 individual skills similarly between the first and second evaluation.

To determine interrater reliability, first, the two assessors (A and B) coded independently, from the videos, same 20 children. Secondly, these two assessors were analyzed on their ability to agree on scoring of the performance criteria of the 13 individual skills.

Statistical Analysis

To determine intrarater and interrater reliability, a kappa statistic (Cohen 1960) and a percent agreement calculation were used. As in a previous study (Barnett et al. 2014) in which reliability of children's gross motor skills measured with TGMD-2 were assessed, we used the magnitudes according to Landis and Koch (1977) for characterizing the resulting statistics: A kappa statistic <0.20 was considered slight; 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 kappa coefficients and corresponding percents of agreement 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. For intrarater reliability assessor A's and B's own kappa coefficients for TS were 0.75 and 0.73, which can be characterized as substantial agreement. Also assessor A's and B's own kappa coefficients were substantial (range from 0.69 to 0.77) in LS and BS. Intrarater percent agreement for LS, BS and TS varied from 87% to 91%. When the individual skills were examined all the kappa values were at least moderate.

Table 1 about here

For interrater reliability kappa coefficients for LS, BS and TS between the two assessors varied from moderate to substantial (range from 0.57 to 0.64). Percent agreement for LS, BS, and TS were all 83% (Table 1).

Based on kappa and/or percent agreement between the assessors, the individual skills most reliably scored were skip (0.87, 93%), two-hand catch (0.84, 94%), and one-hand stationary dribble (0.81, 93%). Denoting slight or fair level of consistency (kappa) three individual skills, (i.e., hop, horizontal jump, and two-hand strike), had the lowest reliability scores (0.19 and 73%; 0.39 and 79%; 0.32 and 72%) (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 amount of 1s and 0s on the same criteria, therefore the Kappa statistic could not be calculated for this criterion. Also, the 4th 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" (κ=0.21, 65%). In the two-hand strike "Preferred hand grips bat above non-preferred hand" indicated slight (κ=0.07, 60%) consistency between assessors (Assessor B scored more "1"). Fair consistency were found in "Non-preferred hip/shoulder faces straight ahead" (κ=0.31, 83%) and in "Steps with non-preferred foot" (κ=0.31, 68%). In both criteria, assessor B scored more "1", but in the first one the assessors agreed 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 showed substantial kappa and excellent percent agreement values for intrarater reliability, and moderate to substantial values for interrater assessment for LS, BS and TS scores. In terms of individual skill reliability, especially the interrater values, there was large variability for three skills (hop, horizontal jump and two-hand strike) with the slight or fair kappa values. It seems that those skills, in particular, have some performance criteria that are challenging to assess.

Reliability values, ranging from 0.62 to 0.75 (TS kappa scores), are considered ‘substantial’ (Landis & Koch, 1977). Moreover, percent agreement ranged from 83 to 91 percent. These high values were expected by assessors A and B who had established reliability with an expert before they began analysis; they coded two children prior to training and established 80% level of agreement with the author of the TGMD-3.

All the children’s performances were on videos. Although the test protocol does not assume videotaping, in this case it allowed assessors to score the same performances twice and to compare their 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. Analysis from the videos has its pros and cons: It allows several viewings to decide whether the criteria were met, but it is time consuming, and does not suit to every day school or daycare life evaluations. However, it is good for research purposes: One can re-analyze the data if necessary.

When looking at the specific individual skill such as ‘two-hand strike on a stationary ball’ (Table 1), we can notice a large difference between assessors’ A and B intrarater kappa values (0.84 vs. 0.47) and percent agreement (94% vs. 80%), but especially in their interrater values (Kappa = 0.32; %Agr = 72). In this case, one challenge will occur if child’s preferred hand is not established: how is the assessor able to determine the score on the first criterion “Child’s preferred hand grips bat above non-preferred hand”. The similar challenges might have been faced in the Barnett et al. (2014) study. Their interrater kappa values for different performance criteria of two-hand strike varied from 0.27 to 0.92 and agreement percentages 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 to be observed and interpreted unambiguously by two different assessors. In the Hop, the kappa value was the lowest (κ =0.19) of all. It was also supported by the low percent agreement (73%). These low values may have originated from the criterion “Foot of non-hopping leg remains behind hopping leg” which may be hard to ‘see’ if the skill is not yet automated. The difference may also become from the fact that one assessor interprets the criterion literally, i.e., another foot cannot pass the other leg at any point during hopping, whereas another assessor may think if it stays behind for the most of the time it will be accepted. 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 fair or slight interrater reliability values, especially in “Preferred hand grips bat above non-preferred hand” (κ =0.07; 60%) that might indicate that it was sometimes difficult to “see” if the criterion was fulfilled. It was not always possible even from the video watching to decide which hand gripped above the other. Sometimes especially younger children’s hands were on top of each other that made the decision difficult. However, there was no indication of similar difficulties in Barnett et al. (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 “Arms extend forcefully forward and upward reaching above the head” -criterion produced the lowest kappa (0.21). In this case the assessors among themselves may have set the different limit for the acceptable performance, i.e., it is acceptable if hands are at the height of a face, or both hands need to reach above head as the criterion says.

Barnett et al. (2014) study revealed that low kappa values may not necessarily mean low values of agreement. In our study, those two values, however, seem to be reflected in one another. Namely, the lowest kappa values as presented above corresponded to same lowest percent agreement values. This distinct phenomenon needs more research to be more fully understood. Differences between the Barnett et al. (2014) and the current study may be explained for example through the scoring protocol and the children’s different skill level. Namely, it is easier to give accurate scores when a child’s skill performance level is high in comparison to those children who are just learning the skill. Similarity of these two values in our study may be caused by the position of video camera. From an ecological validity point of view it is necessary to disturb children as little as possible. In this study it meant that the position of video camera was as constant as possible. This may cause difficulties to see all body movements as precisely as what is seen 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 the two assessors, even with the similar training background, 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 be actually 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 has already established its reputation during the last 30 years. With the development of several national norms of other countries, the test will reach more popularity, and find its way to a practitioners' tool kit.

Ecological validity was the strength of this study. Children's movement skills were able to be measured in their own child care center/ kindergarten or school with familiar educators around them. Children felt comfortable and they did not feel extraordinary stress because of testing situation. Two independent assessors of the study were not aware of the research questions and did their observations based on their understanding of the performance criteria.

In the analysis from the videos, there is a possibility to use slow speed replays of the test performances. When the assessors afterwards discussed the skills that were more challenging to score, they realized they utilized the videos differently in some occasions. Assessor A may have used slow speed replays when assessing especially young children and in unclear situations in specific skills such as hop, horizontal jump, and two-hand strike performances. Assessor B only used the normal video speed. This was a limitation of the study, and might have affected the interrater reliability ratings. For the future video based performance assessments this speed replay option and its use needs to be determined before the beginning of the analysis.

Limited gym sizes in some child care centers can be seen as another limitation of the study. The size of the gym did not allow the full distance for running and galloping. During live observations, assessor 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 angles of the camera in small space, or there is only one optimal location for the camera. In these kind of situations, there will always be hidden spots and not all criteria are visible.

The TGMD-3 showed to be reliable and useful tool to analyze children's gross motor skills. The criteria are well described, and they can be learned through a relatively easy familiarization period. When familiarizing to different observation criteria, special attention needs to be paid on the very quick movements such as in two-hand strike. Moreover, the criteria for hop and horizontal jump need to be recognized as challenging to observe. Additional studies with different kinds of reliability analyses, either based on live observation or video recording, are needed to find 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

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 edition) Dubuque, Iowa: 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.

Ulrich, D. (2016). *Test of Gross Motor Development* (3rd ed.). Austin, TX: Pro-ed.