

**This is an electronic reprint of the original article.
This reprint *may differ* from the original in pagination and typographic detail.**

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/jml.2016-0007>

All material supplied via JYX is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.

1 August 26, 2016

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

3
4 **Abstract**

5 This study examined the intrarater and interrater reliability of the *Test of Gross Motor*
6 *Development—Third Edition* (TGMD-3). Participants were 60 Finnish children aged between
7 3 and 9 years divided into three separate samples of 20. Two samples of 20 were used to
8 examine the intrarater reliability of two different assessors, and the third sample of 20 was
9 used to establish interrater reliability. Children’s TGMD-3 performances were video recorded
10 and later assessed. A kappa statistic and a percent agreement calculation were used. The
11 results for intrarater reliability kappa coefficients for locomotor subtest, ball skills subtest and
12 gross motor total score ranged from 0.69 to 0.77, and percent agreement from 87% to 91%.
13 The interrater kappa coefficients for locomotor subtest, ball skills subtest and gross motor
14 total score ranged from 0.57 to 0.64, which can be considered moderate to substantial
15 reliability. Percent agreement for locomotor skills, ball skills, and total skills was 83%. Hop,
16 horizontal jump and two-hand strike were the most differently assessed performance criteria
17 between the assessors. The TGMD-3 showed to be reliable tool to analyze children’s gross
18 motor skills.

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

20

22 Fundamental motor/movement skills (FMS) are needed to manage motor challenges
23 generated by everyday life (Gallahue, Ozmun, & Goodway, 2012). Gallahue et al. (2012)
24 defined such motor skills as balance skills (e.g., balancing on one foot), locomotor skills
25 (e.g., walking, running and hopping) and manipulative skills (e.g., ball handling skills). These
26 FMS create a basis for children to learn more specific skills to participate in games or
27 different sport activities (Gallahue et al., 2012). Children's motor competence becomes
28 visible through children's FMS performances, and is positively associated to their physical
29 activity level (Stodden et al., 2008). Therefore it is important to follow the development and
30 level of children's motor competence through observing children's performances in different
31 FMS. Today, as many children's motor competence and physical activity levels are low
32 (Reilly, 2010; Roth et al. 2010), it is essential to find valid and reliable observational tools to
33 measure children's motor competence. Having psychometrically valid tools will help
34 researchers and teachers monitor change, the impact of interventions, and the impact of
35 policies. Moreover, measurement tools are needed not only for diagnostic purposes but also
36 to find associations and significance of motor skills for overall development, daily wellbeing
37 and health (Robinson et al. 2015). This was well justified in the study by Cools, Martelaer,
38 Samaey and Andriens (2009) who analyzed seven different movement skill measurements. In
39 addition, cultural comparisons also need measurement tools that are not too sensitive to
40 cultural differences (Cools et al., 2009).

41 When doing research with children, ethical aspects need careful consideration. Observation
42 as a research method is unobtrusive and in that sense much warranted. Unfortunately,
43 reliability of observational tools is questioned. Earlier studies have used either video
44 recordings or live assessments. The TGMD-2 (Ulrich, 2000) was used in the Slotte,
45 Sääkslahti, Metsämuuronen, and Rintala (2015) study. They analyzed children's motor skills
46 through video recordings and reported intrarater reliability for 24 children's motor skills. In
47 their study reliability as intraclass correlation (ICC) was 0.978 for locomotor skills and 0.995
48 for object-control skills. Another study by Barnett, Minto, Lander and Hardy (2014) also used
49 the TGMD-2 version. They reported reliability based on live observation for interrater
50 reliability in six object control skills. Specifically reliability for object control skills was 0.93
51 (ICC), varying in individual skills from 0.71 (catch) to 0.94 (dribble). All values reported are
52 in the acceptable range. More reliability studies are needed to provide valuable information
53 for test developers about the characteristics of the test for the future test development. For

54 example, it cannot be assumed that the reliability values found for the TGMD-2 as such using
55 either video recordings or live observations are applicable to the TGMD-3.

56 The TGMD-3, which was used in this study, is a process-oriented measurement, where
57 children's FMS performances are observed and scored by a rater. The TGMD-3 is a new
58 version of the TGMD-2, but also gathers observations of both locomotor and object control
59 (called ball skills) FMS skills, but differs from TGMD-2 in some individual skill components
60 (Ulrich, 2016). In locomotor skills leaping is replaced with skipping, and in ball skills
61 underhand roll is replaced with underhand throwing. Moreover forehand strike is added
62 which makes altogether six locomotor skills and seven ball skills. Similarly, as in the TGMD-
63 2, the resulting score of each skill is based on the sum of either the presence or absence of the
64 performance criteria (3–5 criteria depending on the skill) of that skill. A more precise
65 description of this tool can be found in another article (see Ulrich, 2013).

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

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

78 **Methods**

79 **Participants and Settings**

80 Participants of this study were randomly selected from the larger study conducted with six
81 elementary schools and eight day care center/kindergarten children ($n = 374$, 3–10 years) who
82 had performed the TGMD-3 in Central Finland. Forty children's performances were used to
83 study intrarater reliability of the two assessors (A and B). Participants of the assessor A were
84 10 boys, ranging from 6-9 years ($M = 7.8 \pm 1.2$) and 10 girls, ranging from 5-9 years ($M =$

85 7.4 ± 1.2). Participants of the assessor B were eight boys, ranging from 4-7 years (M = 6.6 ±
86 1.4) and 12 girls, ranging from 3-7 years (M = 6.1 ± 1.6). Another 20 children's (different
87 from the previous 40 children) performances were randomly chosen for interrater reliability.
88 These children were 10 boys, ranging from 4-6 years (M = 5.9 ± 0.7) and 10 girls, ranging
89 from 5-6 years (M = 6.2 ± 0.5). Institutional approval of the research protocol and informed
90 consent from parents were obtained prior to the study that was approved by the university
91 ethics committee. All children had also the right to refuse participation and refrain from
92 testing any time. None of the assessed children had a disability and/or impairment.

93

94 **Procedure and Data Collection**

95 All trials were conducted in the school gymnasiums or similar locations that were suitable for
96 the administration of the TGMD-3 according to the test instructions. In few cases the space
97 did not allow the full running distance according to the test instructions. Children performed
98 the TGMD-3 administered by a trained physical education professional (one of the authors)
99 and one Master's student in pairs. The professionals were very familiar with administering
100 the TGMD-2 and had used the test before, and the students (five altogether) had had a two-
101 hour training on how to administer the test. One of the two instructed the performer and the
102 other video recorded the performance. The camera was placed optimally (i.e., side view,
103 frontal view or rear view) to best detect skill performance whenever the circumstances
104 permitted. The skills were administered in the order of the scoring sheet as depicted in Table
105 1. Preceding assessment, an accurate demonstration of the skill was performed by the test
106 administrator. Participants were tested in groups of 3-4, and were given one practice trial to
107 assure that the child understood what to do. One additional demonstration was given if a child
108 did not seem to understand the task. Each participant performed two trials individually for
109 each gross motor skill.

110 Two physical education teachers with a Master's degree (different from the test
111 administrators) assessed the test performances from the videos. Both teachers had a good
112 knowledge base about children's motor skills and had been assessing several hundred
113 children on their motor skills using TGMD-3. These assessors had also participated in a two
114 hour training session organized by the first author for elaborating performance criteria. They
115 had also established 80% reliability in scoring with the TGMD-3 author through electronic
116 videos. In rating performances, the scoring system was the following: a score of 1 meant the

117 criterion was performed accurately, and a score 0 meant the criterion was not performed
118 accurately or not performed at all.

119 To determine intrarater reliability, first, the two assessors both coded 20 children's skill
120 performances twice. There was about three months' time interval before their second coding.
121 Secondly, both assessors were analyzed on their own ability to score the performance criteria
122 of the 13 individual skills similarly between the first and second evaluation.

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

126 **Statistical Analysis**

127 To determine intrarater and interrater reliability, a kappa statistic (Cohen 1960) and a percent
128 agreement calculation were used. As in a previous study (Barnett et al. 2014) in which
129 reliability of children's gross motor skills measured with TGMD-2 were assessed, we used
130 the magnitudes according to Landis and Koch (1977) for characterizing the resulting
131 statistics: A kappa statistic <0.20 was considered slight; between 0.21 and 0.40 fair; between
132 0.41 and 0.60 moderate, and 0.61 and above was considered substantial agreement. Percent
133 agreement was also calculated for each sub skill. Significance level was set at 0.05. Data
134 were analyzed using SPSS (version 22 for Windows).

135

136

Results

137 Intra- and interrater kappa coefficients and corresponding percents of agreement of the
138 assessments for individual skills, subtests of locomotor skills (LS), ball skills (BS) and gross
139 motor test total score (TS) are provided in Table 1. For intrarater reliability assessor A's and
140 B's own kappa coefficients for TS were 0.75 and 0.73, which can be characterized as
141 substantial agreement. Also assessor A's and B's own kappa coefficients were substantial
142 (range from 0.69 to 0.77) in LS and BS. Intrarater percent agreement for LS, BS and TS
143 varied from 87% to 91%. When the individual skills were examined all the kappa values were
144 at least moderate.

145

146 Table 1 about here

147

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

151 Based on kappa and/or percent agreement between the assessors, the individual skills most
152 reliably scored were skip (0.87, 93%), two-hand catch (0.84, 94%), and one-hand stationary
153 dribble (0.81, 93%). Denoting slight or fair level of consistency (kappa) three individual
154 skills, (i.e., hop, horizontal jump, and two-hand strike), had the lowest reliability scores (0.19
155 and 73%; 0.39 and 79%; 0.32 and 72%) (Table 1).

156 A more detailed examination of these three skills with the lowest reliability scores was
157 performed (Table 2). For the hop, these criteria were “Arms flex and swing forward to
158 produce force” ($\kappa=0.13$, 63%) and “Foot of non-hopping leg remains behind hopping leg”
159 (43%). In the latter criterion both raters scored the same amount of 1s and 0s on the same
160 criteria, therefore the Kappa statistic could not be calculated for this criterion. Also, the 4th
161 criterion “Hops four consecutive...” assessor A scored all cases “1” in both trials and
162 assessor B scored similarly except for one case, which again did not allow the kappa statistic
163 to be calculated. However, the percent agreement in this criterion was high (98%).

164 In assessing the Horizontal Jump the most inconsistent performance criterion was “Arms
165 extend forcefully forward and upward reaching above the head” ($\kappa=0.21$, 65%). In the two-
166 hand strike “Preferred hand grips bat above non-preferred hand” indicated slight ($\kappa=0.07$,
167 60%) consistency between assessors (Assessor B scored more “1”). Fair consistency were
168 found in “Non-preferred hip/shoulder faces straight ahead” ($\kappa=0.31$, 83%) and in “Steps with
169 non-preferred foot” ($\kappa=0.31$, 68%). In both criteria, assessor B scored more “1”, but in the
170 first one the assessors agreed 83% of the cases.

171

172 Table 2 about here

173

174

175

177 The main purpose of this study was to assess the intra- and interrater reliability of the
178 TGMD-3 video performances of children from 3 to 9 years of age. The results showed
179 substantial kappa and excellent percent agreement values for intrarater reliability, and
180 moderate to substantial values for interrater assessment for LS, BS and TS scores. In terms of
181 individual skill reliability, especially the interrater values, there was large variability for three
182 skills (hop, horizontal jump and two-hand strike) with the slight or fair kappa values. It seems
183 that those skills, in particular, have some performance criteria that are challenging to assess.

184 Reliability values, ranging from 0.62 to 0.75 (TS kappa scores), are considered ‘substantial’
185 (Landis & Koch, 1977). Moreover, percent agreement ranged from 83 to 91 percent. These
186 high values were expected by assessors A and B who had established reliability with an
187 expert before they began analysis; they coded two children prior to training and established
188 80% level of agreement with the author of the TGMD-3.

189 All the children’s performances were on videos. Although the test protocol does not assume
190 videotaping, in this case it allowed assessors to score the same performances twice and to
191 compare their scoring of the same children. Similarly, videotaping has been successfully used
192 in earlier studies (Rintala & Linjala, 2003; Parkkinen & Rintala, 2004; Rintala & Loovis,
193 2013) with earlier TGMD-versions. Analysis from the videos has its pros and cons: It allows
194 several viewings to decide whether the criteria were met, but it is time consuming, and does
195 not suit to every day school or daycare life evaluations. However, it is good for research
196 purposes: One can re-analyze the data if necessary.

197 When looking at the specific individual skill such as ‘two-hand strike on a stationary ball’
198 (Table 1), we can notice a large difference between assessors’ A and B intrarater kappa
199 values (0.84 vs. 0.47) and percent agreement (94% vs. 80%), but especially in their interrater
200 values (Kappa = 0.32; %Agr = 72). In this case, one challenge will occur if child’s preferred
201 hand is not established: how is the assessor able to determine the score on the first criterion
202 “Child’s preferred hand grips bat above non-preferred hand”. The similar challenges might
203 have been faced in the Barnett et al. (2014) study. Their interrater kappa values for different
204 performance criteria of two-hand strike varied from 0.27 to 0.92 and agreement percentages
205 from 78 to 97.

206 The interrater reliability scores of this study showed that hop, two-hand strike and horizontal
207 jump were the most challenging skill performances to be observed and interpreted
208 unambiguously by two different assessors. In the Hop, the kappa value was the lowest (κ
209 =0.19) of all. It was also supported by the low percent agreement (73%). These low values
210 may have originated from the criterion “Foot of non-hopping leg remains behind hopping
211 leg” which may be hard to ‘see’ if the skill is not yet automated. The difference may also
212 become from the fact that one assessor interprets the criterion literally, i.e., another foot
213 cannot pass the other leg at any point during hopping, whereas another assessor may think if
214 it stays behind for the most of the time it will be accepted. Similarly low values were found
215 for “Arms flex and swing forward to produce force”, when there are different kinds of ‘flexed
216 arms’ and the pendulum movement varies in length.

217 The Two-hand strike also had some performance criteria with fair or slight interrater
218 reliability values, especially in “Preferred hand grips bat above non-preferred hand” (κ =0.07;
219 60%) that might indicate that it was sometimes difficult to “see” if the criterion was fulfilled.
220 It was not always possible even from the video watching to decide which hand gripped above
221 the other. Sometimes especially younger children’s hands were on top of each other that
222 made the decision difficult. However, there was no indication of similar difficulties in Barnett
223 et al. (2014) study in which “Hip and shoulder rotation during swing” had the lowest kappa
224 values (0.27 and 0.32). It is notable that they used live observation.

225 In the Horizontal jump the “Arms extend forcefully forward and upward reaching above the
226 head” -criterion produced the lowest kappa (0.21). In this case the assessors among
227 themselves may have set the different limit for the acceptable performance, i.e., it is
228 acceptable if hands are at the height of a face, or both hands need to reach above head as the
229 criterion says.

230 Barnett et al. (2014) study revealed that low kappa values may not necessarily mean low
231 values of agreement. In our study, those two values, however, seem to be reflected in one
232 another. Namely, the lowest kappa values as presented above corresponded to same lowest
233 percent agreement values. This distinct phenomenon needs more research to be more fully
234 understood. Differences between the Barnett et al. (2014) and the current study may be
235 explained for example through the scoring protocol and the children’s different skill level.
236 Namely, it is easier to give accurate scores when a child’s skill performance level is high in
237 comparison to those children who are just learning the skill. Similarity of these two values in

238 our study may be caused by the position of video camera. From an ecological validity point
239 of view it is necessary to disturb children as little as possible. In this study it meant that the
240 position of video camera was as constant as possible. This may cause difficulties to see all
241 body movements as precisely as what is seen in a live observation situation. In live
242 observation the observer may change his/her visual angle naturally, without disturbing
243 children's performance. In general, it can be assumed that the two assessors, even with the
244 similar training background, will always have slightly different views, experience, and
245 potential to assess motor skills.

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

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

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

266 Limited gym sizes in some child care centers can be seen as another limitation of the study.
267 The size of the gym did not allow the full distance for running and galloping. During live
268 observations, assessor may need the full distance to observe all criteria. On one hand this
269 problem can be minimized by videotaping, because the performance can be observed as many

270 times as needed. On the other hand, it is difficult to change the angles of the camera in small
271 space, or there is only one optimal location for the camera. In these kind of situations, there
272 will always be hidden spots and not all criteria are visible.

273 The TGMD-3 showed to be reliable and useful tool to analyze children's gross motor skills.
274 The criteria are well described, and they can be learned through a relatively easy
275 familiarization period. When familiarizing to different observation criteria, special attention
276 needs to be paid on the very quick movements such as in two-hand strike. Moreover, the
277 criteria for hop and horizontal jump need to be recognized as challenging to observe.
278 Additional studies with different kinds of reliability analyses, either based on live observation
279 or video recording, are needed to find the most reliable gross motor skill measurement
280 practices. In addition, studies addressing cultural differences in interpreting different
281 performance criteria are warranted.

282

283

References

284 Barnett, L.M., Minto, C., Lander, N., & Hardy, L.L. (2014). Interrater reliability assessment
285 using the Test of Gross Motor Development-2. *Journal of Science and Medicine in Sport*, 17,
286 667-670. doi: 10.1016/jsams.2013.09.013

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

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

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

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

296

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

300

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

303

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

306
307 Rintala, P., & Loovis, E.M. (2013). Measuring motor skills in Finnish children with
308 intellectual disabilities. *Perceptual and Motor Skills*, *116*, 294-303.
309
310 Robinson, L.E, Stodden, D.F., Barnett, L.M., Lopes, V.P., Logan, S.W., Rodrigues, L.P., &
311 D'Hondt, E. (2015). Motor competence and its effect on positive developmental trajectories
312 of health. *Sports Medicine*, *45*, 1273-1284. doi: 10.1007/s40279-015-0351-6
313
314 Roth, K., Ruf, K., Obinger, M., Mauer, S., Ahnert, J., Schneider, W., ...Hebestreit, H. (2010).
315 Is there a secular decline in motor skills in preschool children? *Scandinavian Journal of*
316 *Medicine and Science in Sports*, *20*, 670–678. doi:10.1111/j.1600-0838.2009.00982.x
317
318 Slotte, S., Sääkslahti, A., Metsämuuronen, J., & Rintala, P. (2015). Fundamental movement
319 skills proficiency and body composition measured by dual energy X-ray absorptiometry in
320 eight-year-old children. *Early Child Development and Care*, *185*, 475-485.

321 Stodden, D., Goodway, J., Langendorfer, S., Robertson, M., Rudisill, M., & Garcia, C.
322 (2008). A developmental perspective on the role of motor skill competence in physical
323 activity: An emergent relationship. *Quest*, *60*, 290–306.
324
325 Ulrich, D. (2000). *Test of Gross Motor Development* (2nd ed.). Austin, TX: Pro-ed.
326
327 Ulrich, D. (2013). The Test of Gross Motor Development-3 (TGMD-3): Administration,
328 scoring, & international norms. *Hacettepe Journal of Sport Sciences*, *24*(2), 27-33.
329
330 Ulrich, D. (2016). *Test of Gross Motor Development* (3rd ed.). Austin, TX: Pro-ed.