# 

# Investigating Expert-Rater Agreement and Inter/Intra-Rater Reliability of Two Fundamental Movement Skills for the Locomotor Subscale of the FG-COMPASS

Lino Perez and Ovande Furtado Jr

Department of Kinesiology, California State University, Northridge

# Author Note

Lino Perez  https://orcid.org/0009-0008-3616-2964

Ovande Furtado Jr  https://orcid.org/0000-0003-3847-6314

Correspondence concerning this article should be addressed to Ovande Furtado Jr, Department of Kinesiology, California State University, Northridge, 18111 Nordhoff St, Northridge, CA 91330-8287, Email: ovande@gmail.com (818-564-7507)

# Abstract

The Furtado-Gallagher Child Observational Movement Pattern Assessment System (FG-COMPASS) is an observational tool using sequential decisions to assess fundamental movement skill proficiency. The current version of the test has three locomotor and five manipulative skills. This study assessed expert-rater agreement and inter- and intra-rater reliability of two new scales to be added to the locomotor subscale. This study was divided into two phases. In Phase I, 60 children between the ages of 5 and 10 were filmed performing the gallop and vertical jump skills. An expert in motor behavior classified the videotapes using the newly created rating scales. Next, eight videos were selected for training purposes and 24 videos for testing purposes. In Phase II, 30 undergraduate students served as raters and underwent training before the testing session. The data were analyzed using weighted kappa (Kw) and the intra-class correlation coefficient (ICC). The results suggested a ‘very good’ agreement between the expert and raters for vertical jump (Kw = .96) and gallop (Kw =.89). The ICC values for vertical jump and gallop were .98 and .94, respectively. The mean kappa values for inter-rater reliability were considered ‘very good’ for vertical jump (MKw=.92) and ‘good’ for gallop (MKw=.78). The ICC values were .98 and .95 for vertical jump and gallop, respectively - which is considered ‘excellent’. The kappa values for intra-rater reliability were .96 and .85 for vertical jump and gallop, respectively. The ICC values were .98 and .92 for vertical jump and gallop, respectively. These results show that the proposed rating scales are reliable in assessing vertical jump and gallop. Future studies should focus on criterion-related validity and reliability evidence from live performances.

*Keywords*: fundamental movement skills, rating scales, assessment, reliability, agreement

# Investigating Expert-Rater Agreement and Inter/Intra-Rater Reliability of Two Fundamental Movement Skills for the Locomotor Subscale of the FG-COMPASS

Fundamental movement skills, such as running, jumping, and throwing, are the foundational building blocks for acquiring more complex skills essential for physical activity throughout childhood and adolescence (Gallahue, 1982). These motor skills enable children to explore and interact with their environment (Haywood & Getchell, 2014). However, it is often erroneously assumed that children will naturally attain proficiency in these fundamental movement skills (FMS) through maturation alone. Instructing children on optimizing these movement patterns to achieve proficiency, particularly during elementary education. This developmental stage presents an unparalleled opportunity for children to understand and master these fundamental movements compared to other stages (Clark, 2007; Colvin et al., 2000).

Moreover, early mastery of FMS has been linked to increased physical activity participation during adolescence and adulthood (Barnett et al., 2009; Kuh & Cooper, 1992; Okely et al., 2001). Alarmingly, children exhibiting lower FMS competence are less inclined towards physical activity (Stodden et al., 2008) and be correlated with higher instances of childhood obesity (Cliff et al., 2012). Consequently, evaluating the development of FMS in children is important. Assessment tools can equip practitioners with insights into a child’s motor skill proficiency and potentially forecast whether a child will likely lead an active or sedentary lifestyle (Stodden et al., 2008).

The assessment in consideration in this study (Furtado-Gallagher Child Observational Movement Pattern Assessment System - FG-COMPASS) aims to assess children’s qualitative aspects of FMS (locomotor and manipulative skills) proficiency. When developing the FG-COMPASS, Furtado and Gallagher (2012) proposed a new method for creating FMS rating scales, which uses key performance criteria in a decision-tree format. A decision tree created for the skill of overhand throw using the method proposed by the authors is depicted in Figure 1. The flowchart is comprised of a) decision nodes (questions), b) chance nodes (Yes or No), and c) end nodes (levels). Although several performance criteria can be proposed for a skill, only three are selected to form the decision tree of a skill based on the suggested method. These performance criteria are then turned into questions which are the decision nodes of the decision tree. In Figure 1, the top decision node contains a discriminatory question. It aims to differentiate between levels 1 and 4 (end nodes). If the chance node is YES, the decision takes the path on the right, and the observer is presented with a confirmatory question, which is meant to verify if a performer is at level 4 (chance node is YES) regarding proficiency of the skill of overhand throw. If the decision tree fails to confirm a level 4 (chance node is NO), the performer is assessed as level 3. The same applies to the left side of the decision tree, except that the purpose is to confirm whether the performer is at level 1. The decision-tree approach simplifies assessing fundamental movement skill live performances as only two performance criteria are used when deciding the proficiency levels of performers.

[Insert Figure 1]

The manipulative subscale of the FG-COMPASS currently comprises strike, hand dribble, kick, overhand throw, and catch, while the locomotor subscale includes skip, hop, and horizontal jump. Although the instrument is functional with three locomotor and five manipulative skills (Furtado & Gallagher, 2012), adding two additional locomotor skills would create a more balanced instrument, thus improving its efficacy in evaluating gross motor skill proficiency. Therefore, this study aimed to develop two rating scales for assessing vertical jump and gallop using the Furtado and Gallagher (2012) method. Additionally, we aimed to examine the reliability of classification decisions of the scales by evaluating rater-expert agreement and rater consistency. We anticipated ratings to be “excellent” (.75 or above) and “very good” (.81 or above) when analyzed with Intraclass Correlation Coefficient and Weighted Kappa, respectively.

# Method

This study was divided into two phases. In Phase I, we videotaped children performing the locomotor skills of vertical jump and gallop. In Phase II, we investigated initial psychometric characteristics for the proposed scales by assessing expert-rater agreement and inter/intra-rater reliability for the two new proposed skills.

## Phase I

### Scale Development

We created the rating scales for gallop (Figure 2) and vertical jump (Figure 3) following the procedures outlined in Furtado and Gallagher (2012).

[Insert Figure 2]

[Insert Figure 3]

Following IRB approval, recruitment packets were distributed to K-5 Los Angeles Unified School District elementary school classrooms. Upon parental/guardian consent, 5-10-year-old children were filmed performing the skills under investigation. We instructed children to perform multiple performance trials. Later, videotapes comprising three trials were created and rated from 1 to 4 using the previously created scales. The videos were first rated independently by the PI (principal investigator) and an expert in motor behavior. Then, their scores were compared, and a consensus was reached for each video on their developmental level.

### Scale Evaluation

We returned to the same school to recruit different participants to help us assess the effectiveness of the previously developed scales. An information packet was sent to the students' households through their classroom teacher, and 60 children were recruited. The packet included a flyer with the study's description, a consent form for the parent/guardian, and an adapted version of the original Physical Activity Readiness Questionnaire - PAR-Q [@shephard1988]. The survey contains 15 questions to identify potential risks while engaging in moderate physical activity. It also assesses the severity of diseases or conditions in the individual's family history. Upon receiving the packets back, the PI randomly selected 12 children from each grade level (kindergarten through 4th grade) between the ages of 5-10. Parents were informed that not all children would be selected to participate. Children were excluded from the study if they 1) had disabilities that could affect the performance of motor skills, 2) had a recent surgery or had undergone surgery in the past six months, 3) had been deemed ineligible based on PAR-Q responses, 4) did not want to participate or be filmed.

#### Filming Children.

Children were assigned a code ranging from 001 to 060 to ensure confidentiality. They were called individually and asked to perform five to six trials for each skill. A GoPro camera (Hero 7) was used to film children with a 1080/50/wide recording setting. Children were filmed galloping twenty feet with the camera placed nine feet away. Cones were placed to indicate where children should begin and stop each trial. To determine the lead foot, children were asked to alternate between right and left feet during practice trials and to use their preferred foot during the remaining trials. Children were asked not to switch feet during filming trials. For the vertical jump skill, tape markers were placed on a wall at 3 ½ feet, with a marker set at every half foot leading up to 7 feet. Children were asked to touch the highest marker with their dominant hand during each trial.

Once the PI recorded all 60 videos, a consensus was reached on the outcome levels of all videos and selected 16 video clips for training purposes (2 videos X 4 levels X 2 skills), and 24 video clips for testing purposes (3 videos X 4 levels X 2 skills); 12 videos for gallop and 12 videos for vertical jump. In case of disagreement between the researchers’ classifications of a video, the video was replaced with one in which the researchers reached a consensus. In Phase II, we recruited and trained undergraduate kinesiology students to participate as raters as part of the evaluation of the proposed rating scales.

## Phase II

### Recruiting Participant

Thirty participants were recruited from undergraduate kinesiology classes via verbal announcements and flyers. Participants were excluded from the study if 1) they were not enrolled in six or more undergraduate units and 2) they were not enrolled in a motor development course. This was done to ensure minimal exposure to movement skill assessment.

### Procedures

#### Training Sessions.

As participants arrived at the lab, they were given general information about the project. Next, participants were asked to read and sign the research informed consent and to carefully study the rating scale for either gallop or vertical jump. The PI physically demonstrated the skill while emphasizing key points related to the skill performance for each level of the rating scale. Participants were then asked to sit before a portable (80 inches across) wide projection screen. Participants then used a pencil and paper version of the scale to classify eight video clips (two random videos of each stage level) depicting different children performing the evaluated skill. Participants received feedback on whether their classifications were correct and were encouraged to ask questions during the training session. The training was done with no more than three participants at a time and lasted about one hour.

#### Testing Sessions.

The testing session took place 48 hours after the training session and recreated the same setup as the training session. Participants were shown a practice video, then classified 12 (two for each skill level) videos presented randomly. Furthermore, participants were instructed not to classify the videos based on the apparent age of the children, as skill levels were distributed across all age levels. Unlike the training session, participants did not receive feedback in case of disagreement with the PI’s classifications. To avoid guessing, participants were not informed of the number of videos they were rating for each level. Each participant underwent individual testing, which took approximately an hour. A week later, they returned to the lab to classify the same videos but in a different order than the initial testing session. The retest was utilized to assess the consistency of measurements by the same rater.

## Data Analysis

Because a certain amount of agreement is expected to occur by chance (Fleiss, 1981), the weighted kappa (Kw) was calculated. Although kappa was originally established for nominal data, weighted kappa was created for categorical data with an ordinal structure. Both scales under investigation range from 1 to 4, with 1 being the lowest assessment score and 4 being the highest. Weighted kappa allows different penalties to be assigned to different mismatches based on the magnitude of disagreement (Hallgren, 2012). If Rater A were to rate a child at a level of 1, and Rater B rated the same child at a level of 2, that would be less severe than a rating of 1 and 4. The values of weighted kappa were compared with the criteria suggested by Altman (1991), with values equal or less than .20 = “poor”; .21 - .40 = “fair”; .41-.60 = moderate; .61-.80 = “good”; and .81-1.00 = “very good.” In addition, the intraclass correlation coefficient (ICC) was used to test reliability. A two-way random (consistency), average-measures was used and interpreted as follows: less than .40 = “poor”; .40-.75 = “fair to good”; >.75 = “excellent” (Fleiss, 2011). To be included in the test, the ratings needed to meet a minimum of .75 for ICC and .81 for Kappa.

# Results and Discussion

This study aimed to propose and assess initial psychometric characteristics for two rating scales to be added to the locomotor subscale of the FG-COMPASS. All expected hypotheses were confirmed except for the Kappa value of gallop skill’s inter-rater reliability. The results are detailed in the following section. First, we present the agreement between the expert and participants’ ratings. Then, inter-rater rater reliability data are presented, followed by intra-rater reliability.

## Expert-rater agreement

The expert-rater agreement was assessed by comparing the ratings of all 30 participants with those of an expert. The agreement was considered “very good” for both vertical jump (Kw = .96) and gallop (Kw = .89;). In addition, ICC scores were considered “excellent” for both vertical jump (ICC = .98) and gallop (ICC = .94). Table 1 presents the results for both skills. The high agreement observed could be due to the specific performance criteria selected for the gallop and vertical jump rating scales. We followed Painter (1994) suggestion of selecting only the critical performance criteria for the skill when creating the scales, which may have aided the raters in the study. It is important to note that the individuals involved in this research received thorough training before evaluating the videos. Moreover, this study solely focused on pre-recorded performances, which are less challenging to evaluate than live performances.

Table 1

Weighted Kappa and ICC statistics for the expert-rater agreement

|  | Kw | ICC |
| --- | --- | --- |
| Vertical Jump | .96(.939, .978) | .98(.974, .982) |
| Gallop | .89(.847, .926) | .94(.927, .952) |

*Note*. Sample size = 30. Values in parentheses refer to the confidence interval; Kw= Weighted Kappa; ICC = Intraclass Correlation Coefficient.

## Inter-rater reliability

Inter-rater reliability was assessed by comparing the ratings of five raters chosen at random. Weighted kappa scores were computed for each rater pair and averaged to provide a single reliability index. The results are presented in Tables 2 and 3. The weighted kappa values for the vertical jump ranged from .82 to 1.0 (M = .92). Weighted kappa values for the skill of gallop ranged from .57 to 1.0 (MKw = .78). These scores were considered “very good” for the vertical jump and “good” for the skill of gallop (Kw < .81). Note that for the skill of gallop, rater #2 fell below the expected agreement when compared with all raters. ICC scores were interpreted as “excellent” for both vertical jump (ICC = .98) and gallop (ICC = .95). It is possible that the discrepancy in ratings for the skill of gallop between rater #1 and three other raters was caused by rater #2’s lack of understanding of the protocol, despite the training provided. This is supported by the fact that the ratings of all other raters fell within a range of (Kw = .81-1.00). If rater #2 is removed, the weighted kappa scores increase from .78 to .88, suggesting a “very good” agreement for Kappa.

Table 2

Weighted Kappa for inter-rater-rater analyses

|  | Raters |  |  |  |  |  |  |  |  |  | Mkw |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1x2 | 1x3 | 1x4 | 1x5 | 2x3 | 2x4 | 2x5 | 3x4 | 3x5 | 4x5 |  |
| Vertical Jump | 0.96 | 0.93 | 0.9 | 0.93 | 0.96 | 0.82 | 0.96 | 0.85 | 1.00 | 0.85 | 0.92 |
| Gallop | 0.73 | 0.82 | 0.82 | 0.81 | .63\* | .63\* | .57\* | 1.00 | 0.91 | 0.91 | 0.78 |

*Note*. Kw= weighted kappa; 1x2, 1x3, …= rater pairs agreement; Mk= arithmetic mean of rater pairs agreement; Asterisks indicate disagreement below expected.

Table 3

Intraclass Correlation Coefficient for the inter-rater analyses

|  | ICC | 95% CI |
| --- | --- | --- |
| Vertical Jump | .98 | .963, .995 |
| Gallop | .95 | .890, .984 |

*Note*. Values computed using two-way random, average measures ICC (2,1).

## Intra-rater reliability

To determine intra-rater reliability, raters randomly reassessed the same videos one week after their initial assessment. Intra-rater reliability was considered “very good” for both vertical jump (Kw = .96) and gallop (Kw = .85). ICC scores were interpreted as “excellent” for vertical jump (ICC = .98) and gallop (ICC = .92). The results are presented in Table 4. Although the agreement for both skills was considered “very good,” the kappa score was higher for the vertical jump. The difference may be due to the skill of gallop having more difficult performance criteria to assess than that of vertical jump. For example, the discriminating criterion for the scale of gallop asked, “Was the action smooth (not choppy) and performed at a moderate speed?” Raters might have had difficulties determining a “smooth” and “moderate speed” action from a “choppy” and “fast tempo”. Although training provided examples of various proficiency levels, performances varied across children. That is, no two children displayed each criterion in the same manner. Although the same holds for the skill of vertical jump, it may be that raters found it easier to assess the skill of vertical jump because the criteria might have been more objective than that of the skill of gallop. The criterion for discrimination in the vertical jump scale was whether the arms moved backward during the preparation phase. This criterion may have been easier for raters to assess compared to the gallop scale. There may have been more variability in skill when determining this criterion for gallop, while detecting variability in the vertical jump may have been simpler.

Table 4

Weighted Kappa and ICC statistics for the intra-rater analyses

|  | Kw | ICC |
| --- | --- | --- |
| Vertical Jump | .96(.940, .976) | .98(.974, .983) |
| Gallop | .85(.795, .895) | .92(.897, .932) |

*Note*. Sample size = 30; Values in parentheses refer to confidence interval; Kw= Weighted Kappa; ICC = Intraclass Correlation Coefficient.

# Conclusion

Fundamental movement skills have been proposed to be the foundation of an active lifestyle (Clark, 2007; Gallahue, 1982). Early detection of delays in FMS performance may lead to earlier and more appropriate interventions. Having a practical FMS assessment tool such as the FG-COMPASS may facilitate the detection of motor skill development of children in school settings where time may be of conflict. Existing assessment tools often comprise complex criteria, requiring extensive training (Lander et al., 2016). By employing a method that uses only three performance criteria as part of the rating scales, the FG-COMPASS may be a more practical tool for the average K-5 educator who might only get a limited amount of time with a surplus of students (Olrich, 2002). The development of such tools may lead to enhancing physical education programs, allowing educators to tailor plans according to children’s developmental levels. Thus, children with higher FMS are more likely to participate in physical activity, which may allow for continued active lifestyles later in life (Cliff et al., 2012). The results of this study provide evidence of rater-expert agreement and inter/intra-rater reliability for two rating scales to be added to the locomotor subscale of the FG-COMPASS. These results show that the proposed rating scales are reliable in assessing vertical jump and gallon. Future studies should focus on criterion-related validity and reliability evidence from live performances.

## Declaration of Conflicting Interest

The authors state that they have no conflicts of interest.

# References

Adams, R. (1999). Revised Physical Activity Readiness Questionnaire. *Canadian Family Physician*, *45*, 992–1005. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2328306/>

Altman, D. (1991). *Practical statistics for medical research* (1st ed.). Chapman; Hall.

Barnett, L. M., van Beurden, E., Morgan, P. J., Brooks, L. O., & Beard, J. R. (2009). Childhood Motor Skill Proficiency as a Predictor of Adolescent Physical Activity. *Journal of Adolescent Health*, *44*(3), 252–259. <https://doi.org/10.1016/j.jadohealth.2008.07.004>

Clark, J. (2007). On the problem of motor skill development. *Journal of Physical Education, Recreation & Dance*, *78*(5), 39–44. <https://doi.org/10.1080/07303084.2007.10598023>

Cliff, D. P., Okely, A. D., Morgan, P. J., Jones, R. A., Steele, J. R., & Baur, L. A. (2012). Proficiency deficiency: Mastery of fundamental movement skills and skill components in overweight and obese children. *Obesity*, *20*(5), 1024–1033. <https://doi.org/10.1038/oby.2011.241>

Colvin, A., Markos, N., & Walker, P. (2000). *Teaching the nuts and bolts of physical education : Building basic movement skills*. Champaign, IL: Human Kinetics.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). Wiley.

Fleiss, J. L. (2011). *Design and analysis of clinical experiments*. John Wiley & Sons.

Furtado, O., Jr., & Gallagher, J. D. (2012). The reliability of classification decisions for the Furtado-Gallagher Computerized Observational Movement Pattern Assessment System - FG-COMPASS. *Research Quarterly for Exercise and Sport*, *83*(3), 383–390. <https://doi.org/10.5641/027013612802573021>

Gallahue, D. L. (1982). Assessing motor development in young children. *Studies in Educational Evaluation*, *8*(3), 247–252.

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3402032/>

Haywood, K., & Getchell, N. (2014). *Life span motor development*. Champaign, IL: Human Kinetics.

Kuh, D. J., & Cooper, C. (1992). Physical activity at 36 years: Patterns and childhood predictors in a longitudinal study. *Journal of Epidemiology and Community Health*, *46*(2), 114–119. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1059517/>

Lander, N., Morgan, P. J., Salmon, J., & Barnett, L. M. (2016). Teachers’ perceptions of a fundamental movement skill (FMS) assessment battery in a school setting. *Measurement in Physical Education and Exercise Science*, *20*(1), 50–62. <https://doi.org/10.1080/1091367X.2015.1095758>

Okely, A. D., Booth, M. L., & Patterson, J. W. (2001). Relationship of physical activity to fundamental movement skills among adolescents: *Medicine and Science in Sports and Exercise*, *33*(11), 1899–1904. <https://doi.org/10.1097/00005768-200111000-00015>

Olrich, T. W. (2002). Assessing Fundamental Motor Skills in the Elementary School Setting: Issues and Solutions. *Journal of Physical Education, Recreation & Dance*, *73*(7), 26–28. <https://doi.org/10.1080/07303084.2002.10607843>

Painter, M. A. (1994). Developmental sequences for hopping as assessment instruments: A generalizability analysis. *Research Quarterly for Exercise & Sport*, *65*(1), 1–10.

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 emergent relationship. *Quest*, *60*(2), 290–306. <https://doi.org/10.1080/00336297.2008.10483582>

# Appendix

## Figure captions

Figure 1

Rating scale for the skill of overhand throw.

Figure 2

Rating scale for the skill of gallop.

Figure 3

Rating scale for the skill of vertical jump.