# Collecting Concurrent Validity, Expert-Rater Agreement, and Inter/Intra-Rater Reliability for the FG-COMPASS

Mackinsey Woolever1, Ovande Furtado Jr1, and Jere D. Gallagher2

1Department of Kinesiology, California State University, Northridge

2University of Pittsburgh

# Author Note

Mackinsey Woolever https://orcid.org/0000-0000-0000-0000

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

Jere D. Gallagher is deceased.

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

Fundamental movement skills (FMS) are the building blocks for developing specialized sports skills. In addition, fundamental movement skill competency has been linked to decreased levels of obesity and increased levels of physical activity/sports participation. This study aimed to collect evidence for concurrent validity, inter-intra-, and expert-rater reliability for the FG-COMPASS. Participants were 34 children ages 5-10 years. The agreement between the FG-COMPASS and TGMD-2 was evaluated using the intraclass correlation coefficient (ICC) and Bland-Altman analysis for LFMS, MFMS, and TFMS. The ICC for the LFMS subtest was 0.68, indicating a ‘good’ agreement between the two tests. The Bland-Altman analysis revealed a mean bias close to zero. ‘Excellent’ agreement was observed for the MFMS subtest, with an ICC of 0.89 and Bland-Altman bias close to zero. The TFMS also demonstrated ‘excellent’ agreement, with an ICC of 0.89 and Bland-Altman bias close to zero. The inter-rater reliability of the FG-COMPASS was assessed by comparing the live scores of all five raters involved in the study. Weighted kappa values ranged from 0.42 to 0.93 (M = 0.66, SD = .15) for the LFMS subtest and from 0.31 to 0.93 (M = 0.77, SD = 0.15) for the MFMS subtest. The test’s combined score (TFMS) was 0.72 (SD = 0.15). The resulting weighted kappa indicates ‘good’ agreement for the LFMS, MFMS subtests, and the total test (TFMS). The weighted kappa results for the intra-rater reliability indicate a ‘good’ agreement for LFMS (0.69), MFMS (0.70), and TFMS (0.70). The agreement between raters and the expert indicates a ‘good’ agreement for LFMS (0.64), MFMS (0.75), and TFMS (0.70). This study shows that the FG-COMPASS is a reliable and effective tool for assessing motor skills. Future research should focus on exploring any minor differences in the locomotor subtest and understanding the variations in rater interpretations. Additionally, adding two more skills to the locomotor subtest would be beneficial.

*Keywords*: Assessment, Fundamental Movement Skills, Children, Movement Competence, Rating Scales, Assessment, Reliability, Agreement, Concurrent Validity

# Collecting Concurrent Validity, Expert-Rater Agreement, and Inter/Intra-Rater Reliability for the FG-COMPASS

The acquisition of basic movement skills (FMS) holds significant importance for young children, as it has the potential to impact their development and lifestyle in both positive and negative ways. The level of motor competence is crucial to an individual’s engagement in physical activities (Castelli & Valley, 2007). Previous research has demonstrated a favorable correlation between perceived motor competence and proficiency in fundamental motor skills (FMS) among children and adolescents (Woods et al., 2007). There is a favorable correlation between engagement in physical activity and the development of fundamental movement skills (FMS), particularly when the activities are of moderate to vigorous intensity (Bellows et al., 2013; Fisher et al., 2005; Lemos, 2012; McKenzie et al., 2002). Additionally, there is an inverse relationship between physical activity involvement and the prevalence of obesity (Bayer et al., 2009; Graf et al., 2004; Lopes et al., 2012). Research has indicated a correlation between inadequate motor skill competence in children and reduced physical activity levels (Fisher et al., 2005; McKenzie et al., 2002). Furthermore, this lack of motor skill proficiency has been associated with an increased likelihood of being overweight or obese (Cliff et al., 2012).

Assessing the development of fundamental movement skills (FMS) is crucial for preschool and primary school children, as these skills significantly influence their overall development and well-being throughout their lifespan. According to Ulrich (2000), early childhood education tends to neglect the development of gross motor abilities. Assessment instruments are specifically developed to identify variations in motor development that can span a spectrum of severity, encompassing slight and significant delays. Failure to promptly address delays in motor skill development may impede the subsequent development of gross motor abilities (Ulrich, 2000) and may affect the acquisition of specialized skills (Gabbard, 2021). According to Provost et al. (2000), a developmental delay is a discrepancy of 25% or greater between a child’s chronological age and developmental age. Early detection of motor delay enables practitioners, parents, and educators to employ intervention measures to assist the child. A range of instruments exists for assessing fundamental movement skills, as highlighted by Cools et al. (2009). Nevertheless, it is important to note that these assessment tools are often not intended for real-time assessment by a single practitioner. Typically, test administrators are required to record the performances and then evaluate them. Developing an assessment tool that enables professionals to assess children’s FMS proficiency without needing performance recording will enhance convenience.

It is essential to have a practical assessment instrument to assess gross motor development in school-age children. This helps identify any potential motor delays and ensures typical development. The Furtado-Gallagher Child Observational Movement Pattern Assessment System (FG-COMPASS) was developed to fill this gap, relying on only three performance criteria to evaluate FMS performance (Furtado & Gallagher, 2012). To date, there have been no attempts to compare the results of the FG-COMPASS with those of a criterion measure. In addition, while the FG-COMPASS has demonstrated evidence of both inter- and intra-rater reliability, the agreement data utilized in these studies were obtained through video analysis rather than live performances (Furtado & Gallagher, 2012, 2018). Therefore, this study aimed to collect criterion-related (concurrent) validity for the FG-COMPASS by comparing its results with the TGMD-2 (Ulrich, 2000), a popular instrument for assessing FMS in children. In addition, this study sought to collect further inter-, intra-, and expert-rater reliability evidence for the FG-COMPASS from live assessments. Several hypotheses were proposed for this study. We anticipated that there would be at least a ‘good’ agreement (ICC/kappa scores above 0.74) for the locomotor (LFMS), manipulative (MFMS), and total test (TFMS) when investigating concurrent validity, inter- and intra-rater reliability, and expert-rater reliability for the FG-COMPASS.

# Materials and Methods

## Participants

A convenient sampling method was used to recruit participants for this study. After Institutional Review Board approval, 41 children between the ages of 5 and 10 were recruited. However, only 34 children, 22 girls (M = 8.14, SD = 1.78) and 12 boys (M = 8.44, SD = 1.49), participated fully. Three participants dropped out independently; two never attended the assessment sessions; one was injured before the assessment, and one child moved out of the state during the data collection. Informed consent was obtained from each participant’s parents or legal guardians before their involvement in the study. One randomly selected kindergarten through fourth-grade classroom received a recruitment packet to control for an even distribution of age ranges. Recruitment packets were given to additional classrooms if no consent forms were returned. A participant was excluded from the study if he/she: 1) was younger than five years or older than 10 years and 11 months; 2) had developmental delays or disabilities that may have affected their motor performance; 3) had no parental or verbal consent; or 4) had a “yes” response on any of the first five Physical Activity Readiness-Questionnaire (PAR-Q) questions (Adams, 1999).

## Measures

### TGMD-2

The TGMD-2 (Ulrich, 2000), served as the gold standard for the concurrent validity analysis. This tool evaluates twelve gross motor skills, divided into two categories: six concerning locomotion and six related to object manipulation, suitable for children aged between 3 and 10 years. The TGMD-2 assesses various performance criteria focusing on different body components, including the arms, legs, and trunk. For instance, one criterion for hopping is that the “arms be flexed and swing forward to generate force.” Each child is given two attempts to demonstrate each skill; a successful attempt awards a score of 1, while unsuccessful attempts receive a score of 0. After two trials for each skill, the scores are totaled to find the overall raw score for each subtest. These raw scores are then converted to standard scores, enabling comparison to normative data. Researchers can determine age equivalents and percentiles by analyzing the standard scores from both subtests. These standard scores, when combined, form a gross motor quotient value.

Several studies have addressed the validity and reliability of the TGMD-2. The instrument demonstrates strong construct validity, effectively measuring the fundamental theoretical construct of gross motor development (Ulrich, 2000). Its concurrent validity is supported through correlations with other established motor skill assessments, such as the Movement Assessment Battery for Children (MABC) (Cools et al., 2009). High test-retest reliability of the TGMD-2 is reported, with coefficients frequently exceeding 0.90 (Valentini & Rudisill, 2004). The test maintains good internal consistency, evidenced by Cronbach’s alpha values generally above 0.80 for both the locomotion and object manipulation subtests (Wiart & Darrah, 2001). The TGMD-2 has also been adapted and validated in different cultural contexts, indicating its utility across various populations (Barnett et al., 2009). The standardization process involved a representative sample of children in the United States, ensuring the relevance of the normative data for the target age group.

### FG-COMPASS

The FG-COMPASS is a tool created to assess the gross motor skills of children aged 5 to 10 years. It evaluates eight skills, including three pertaining to locomotion and five to object manipulation. In the development of FG-COMPASS, Furtado and Gallagher (2012) introduced a novel approach to constructing FMS rating scales. This method, detailed further in works by Perez (2018) and Furtado and Gallagher (2012), utilizes only key performance criteria arranged in a decision tree. Figure 1 depicts the decision tree for the ‘Overhand Throw’ skill, demonstrating three types of nodes: decision nodes, which pose questions; chance nodes, offering ‘yes or no’ options; and end nodes, which indicate levels.

Despite the availability of many performance criteria for each skill, only three are selected to create the decision tree, which is then formed through questions derived from these criteria. A particular instance of this can be observed in Figure 1, where a question at the top decision node aims to differentiate between levels 1 and 4. A ‘yes’ at the chance node directs to the right pathway, prompting a follow-up question to confirm if the individual exhibits a level 4 proficiency in overhand throws. If the answer is ‘no’, the rating settles at level 3.

A similar process occurs on the left side of the tree, this time to assess for level 1 proficiency. This approach allows for a simplified evaluation of live performances of fundamental movement skills, restricting the assessment to just two performance criteria when determining skill proficiency levels. It enjoys support from several studies validating its efficacy and reliability, including evidence for content-related validity (Furtado, 2004), expert-rater agreement (Furtado & Gallagher, 2012, 2018), and intra- and inter-rater reliability (Furtado & Gallagher, 2018).

[figure 1 here]

## Research Assistants

Sixteen undergraduate students studying kinesiology were selected as research assistants for the study. Ten students were chosen to work as raters in the assessment of gross motor skills, with a random assignment of five raters per instrument: TGMD-2 and FG-COMPASS. The remaining students were assigned roles as test administrators- three for TGMD-2 and two for FG-COMPASS- except for one individual assigned to edit videos remotely, hence not requiring presence at the research site.

Before initiating data collection, all participants underwent training tailored to their designated roles and respective instruments. This involved multiple sessions aimed at acquainting them with the necessary skills, including instruction on the setup for skills assessment, demonstrating tasks accurately, and the appropriate use and positioning of cameras for recording. The training ensured that both raters and test administrators were well-prepared to carry out their tasks effectively during the data collection phase.

## Procedures

### Data collection

The TGMD-2 and the FG-COMPASS, although assessing the same construct, have distinct differences in their evaluation methods. Despite these differences, there were common procedures followed in administering both tests. The assessments were conducted during the morning or lunch recess periods, ensuring all student participants could complete both tests.

Before the onset of data collection, the primary investigator assigned identification numbers to the participants to maintain order and organization. Upon reaching the test location, participants were guided to either the TGMD-2 or the FG-COMPASS stations as determined by a randomized list. This setup was revisited weekly to negate potential learning effects from repeated measurements of the same skill. After concluding the initial test at one of the stations, participants moved to the alternate station for further assessment. In certain instances, participants were asked to demonstrate multiple skills across both stations, alternating between them as needed.

The administrators overseeing the TGMD-2 station visited the site weekly to conduct assessments per the directives detailed in the user manual by Ulrich (2000). All skill performances were videotaped. Subsequently, the TGMD-2 raters were called to the lab to assess the performances through the recorded videos. The FG-COMPASS raters evaluated the participants directly on-site, though these performances were also recorded to allow for intra-rater reliability analysis. The participants were accorded one practice trial and three test trials for each skill.

## Data Analysis

Concurrent validity was assessed using the intraclass correlation coefficient (ICC, 2, k) and Bland-Altman plots. The reliability of the FG-COMPASS, both in terms of inter- and intra-rater consistency and the agreement between the expert and raters, was examined using weighted kappa statistics.

The ICC is designed to evaluate the reliability of ratings by comparing the variation of different ratings for the same participant to the total variation across all ratings and participants (McGraw & Wong, 1996). The Bland-Altman method involves determining the agreement between two quantitative measurements by studying the average difference and setting agreement limits, a strategy preferred over correlation techniques (Bland & Altman, 1986; Giavarina, 2015).

The weighted kappa index is recommended when more than two coders independently classify an entity into three or more categories at an ordinal level, assigning different weights based on the degree of disagreement (Fleiss & Cohen, 1973). Unlike kappa, the weighted version considers the degree of disagreement in its calculation. The interpretation of weighted kappa and ICC values followed this scheme: above 0.75 indicated excellent agreement; values between 0.74 and 0.60 were good; values between 0.59 and 0.40 were fair, and values below 0.40 were poor (Cicchetti, 1994).

A cross-sectional design was utilized in this study, and all statistical analyses were conducted using the statistical package jamovi (The jamovi project, 2022).

# Results

## Concurrent Validity

The agreement between the FG-COMPASS and TGMD-2 (see Table 1) was evaluated using the intraclass correlation coefficient and Bland-Altman analysis for LFMS, MFMS, and TFMS. The ICC for the LFMS subtest was 0.68, indicating a ‘good’ agreement between the two tests. The Bland-Altman analysis revealed a mean bias close to zero. ‘Excellent’ agreement was observed for the MFMS subtest, with an ICC of 0.89 and Bland-Altman bias close to zero. The TFMS also demonstrated ‘excellent’ agreement, with an ICC of 0.89 and Bland-Altman bias close to zero.

Table 1

Concurrent Validity Analysis for FG-COMPASS

|  | ICC | F-Statistic (p-value) | Bias | LLA | ULA |
| --- | --- | --- | --- | --- | --- |
| **Locomotor** | 0.68 | 3.05 (\*) | 1.73e−16 | −1.95 | 1.95 |
| **Manipulative** | 0.89 | 8.85 (\*\*) | 2.60e−16 | −1.25 | 1.25 |
| **Total test** | 0.89 | 9.00 (\*\*) | −2.53e−16 | −1.24 | 1.24 |

*Note*. ICC: Intraclass Correlation Coefficient (ICC, 2, k). Bias: Mean difference between the two methods. LLA & ULA: 95% lower and upper limits of agreement in the Bland-Altman analysis. (\*): p < 0.05. (\*\*): p < 0.001.

## Inter-Rater Reliability

The inter-rater reliability of the FG-COMPASS was assessed by comparing the live scores of all five raters involved in the study (see Table 2). Weighted kappa values were computed for each pair of raters, and then the mean was calculated to provide a single index of reliability (Light, 1971). Weighted kappa values ranged from 0.42 to 0.93 (M = 0.66 SD = .15) for the LFMS subtest and from 0.31 to 0.93 (M = 0.77 SD = 0.15) for the MFMS subtest. The test’s combined score (TFMS) was 0.72 (SD = 0.15). The resulting weighted kappa indicates ‘good’ agreement for the LFMS, MFMS subtests, and the total test (TFMS).

Table 2

Weighted Kappa Statistics for the Inter-Rater Analysis

|  | Skill | 1x2 | 1x3 | 1x4 | 1x5 | 2x3 | 2x4 | 2x5 | 3x4 | 3x5 | 4x5 | Mean | SD |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **LFMS** |  |  |  |  |  |  |  |  |  |  |  | **.66** | .15 |
|  | Hop | .56 | .43 | .43 | .50 | .42 | .49 | .46 | .53 | .64 | .52 | .50 | .07 |
|  | Jump | .93 | .91 | .70 | .65 | .84 | .73 | .62 | .61 | .61 | .69 | .73 | .12 |
|  | Skip | .78 | .74 | .60 | .63 | .87 | .77 | .76 | .69 | .71 | .83 | .74 | .08 |
| **MFMS** |  |  |  |  |  |  |  |  |  |  |  | **.77** | .15 |
|  | Throw | .85 | .79 | .85 | .45\* | .80 | .73 | .35\* | .71 | .31\* | .40\* | .62 | .22 |
|  | Kick | .83 | .68 | .64 | .82 | .63 | .59 | .73 | .50 | .72 | .64 | .68 | .10 |
|  | Dribble | .86 | .87 | .83 | .92 | .88 | .84 | .91 | .88 | .93 | .93 | .89 | .04 |
|  | Catch | .87 | .91 | .87 | .90 | .83 | .84 | .90 | .80 | .80 | .89 | .86 | .04 |
|  | Strike | .81 | .80 | .72 | .72 | .87 | .80 | .80 | .69 | .78 | .79 | .78 | .05 |
| **TFMS** |  |  |  |  |  |  |  |  |  |  |  | **.72** | .15 |

*Note*. LFMS = Locomotor Subtest. MFMS = Manipulative Subtest. TFMS = Locomotor and Manipulative scores combined. Asterisks indicate major disagreement between rater 5 and all other raters.

## Intra-rater reliability for the FG-COMPASS

We evaluated the consistency of ratings by comparing live assessments with those obtained through video assessments (see Table 3). Of particular interest is ‘IR1’, which compares the ratings obtained from live assessments with those from recorded performances (time 1), which took place one week after the live assessment at the site. Additionally, we included ratings from a second video assessment and compared it with the live assessment (IR2) and the first video assessment (IR3). In Table 4, we provide the mean and standard deviation scores for each intra-rater comparison across all five raters. Scores ranged from 0.50 to 0.89, indicating agreement from ‘fair’ to ‘excellent’ (Cicchetti, 1994).

Table 3

Intra-Rater Agreement Measured by Weighted Kappa Coefficients Across Different Skills

|  | Rater 1 | Rater 2 | Rater 3 | Rater 4 | Rater 5 |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | IR1 | IR2 | IR3 | IR1 | IR2 | IR3 | IR1 | IR2 | IR3 | IR1 | IR2 | IR3 | IR1 | IR2 | IR3 |
| Hop (17) | 0.72 | 0.83 | 0.89 | 0.44 | 0.55 | 0.75 | 0.64 | 0.62 | 0.88 | 0.69 | 0.40 | 0.62 | 0.84 | 0.62 | 0.80 |
| Jump (18) | 0.56 | 0.49 | 0.83 | 0.46 | 0.68 | 0.39 | 0.74 | 0.38 | 0.49 | 0.56 | 0.28 | 0.56 | 0.67 | 0.67 | 1.00 |
| Skip (18) | 0.96 | 0.96 | 1.00 | 0.94 | 0.78 | 0.86 | 0.92 | 0.86 | 0.89 | 0.51 | 0.68 | 0.86 | 0.66 | 0.52 | 0.83 |
| Throw (18) | 0.80 | 0.52 | 0.53 | 0.63 | 0.62 | 0.97 | 0.52 | 0.54 | 0.60 | 0.69 | 0.76 | 0.70 | 0.54 | 0.66 | 0.86 |
| Kick (19) | 0.72 | 0.68 | 0.78 | 0.66 | 0.60 | 0.76 | 0.80 | 0.70 | 0.74 | 0.74 | 0.64 | 0.72 | 0.64 | 0.62 | 0.80 |
| Dribble (17) | 0.86 | 0.80 | 0.82 | 0.70 | 0.78 | 0.86 | 0.66 | 0.60 | 0.82 | 0.76 | 0.68 | 0.74 | 0.60 | 0.64 | 0.68 |
| Catch (20) | 0.68 | 0.62 | 0.86 | 0.86 | 0.66 | 0.80 | 0.76 | 0.78 | 0.82 | 0.66 | 0.60 | 0.78 | 0.84 | 0.88 | 0.88 |
| Strike (20) | 0.62 | 0.56 | 0.60 | 0.74 | 0.70 | 0.70 | 0.64 | 0.64 | 0.64 | 0.74 | 0.68 | 0.68 | 0.70 | 0.62 | 0.78 |

*Note*. Quadratic weighted kappa values range from -1 to 1. IR1, IR2, and IR3 represent three different comparison pairs: IR1 compares the ratings from the live performances with the recorded video (time 1), IR2 compares ratings from the live performances with the recorded video (time 2), and IR3 compares the ratings from the recorded performances from time 1 and time 2. Sample sizes for each skill are provided in parentheses next to the skill name. Rater 1 to Rater 5 represents five different individuals who rated the performances.

Table 4

Summary of Quadratic Weighted Kappa for Intra-Rater Comparisons Across Motor Skills

| Skill | IR1 Mean (SD) | IR2 Mean (SD) | IR3 Mean (SD) |
| --- | --- | --- | --- |
| LFMS | **0.69** | 0.62 | 0.78 |
| Hop | 0.67 (0.13) | 0.60 (0.14) | 0.79 (0.10) |
| Jump | 0.60 (0.10) | 0.50 (0.16) | 0.65 (0.23) |
| Skip | 0.80 (0.18) | 0.76 (0.15) | 0.89 (0.06) |
| MFMS | **0.70** | 0.66 | 0.76 |
| Throw | 0.64 (0.10) | 0.62 (0.09) | 0.73 (0.16) |
| Kick | 0.71 (0.06) | 0.65 (0.04) | 0.76 (0.03) |
| Dribble | 0.72 (0.09) | 0.70 (0.08) | 0.78 (0.06) |
| Catch | 0.76 (0.08) | 0.71 (0.11) | 0.83 (0.04) |
| Strike | 0.69 (0.05) | 0.64 (0.05) | 0.68 (0.06) |
| TFMS | **0.70** | 0.64 | 0.77 |

*Note*. IR1, IR2, IR3: Represent different inter-rater comparisons (e.g., live vs. video 1, live vs. video 2, video 1 vs. video 2). Mean: The average of the quadratic weighted kappa values for each scenario across all raters and skills. SD: The standard deviation of the quadratic weighted kappa values for each scenario across all raters and skills.

## Expert-Rater Agreement for the FG-COMPASS

The study aimed to assess the level of agreement between raters and expert ratings for the FG-COMPASS. The live ratings of five raters were compared with the expert’s video scores. The results showed that the level of agreement varied across the skills assessed. For ‘Hop’, the agreement ranged from ‘fair’ to ‘excellent’, with a mean kappa value of 0.67. The agreement was more varied for’ Skip’, with kappa values ranging from ‘poor’ to ‘excellent’. In the MFMS subtest, the agreement was considered ‘excellent’. For ‘Throw’, the agreement ranged from good to excellent, with a mean kappa value of 0.75. For ‘Dribble’, all raters showed ‘excellent’ agreement, with kappa values ranging from 0.77 to 0.94 and a mean kappa of 0.86. For ‘Catch’, the agreement was ‘excellent’, with a mean kappa value of 0.83. For ‘Strike’, the agreement was ‘fair’ to ‘good’ with kappa values ranging from 0.40 to 0.61 and a mean kappa of 0.51. Considering all skills (TFMS), the mean kappa value was 0.70, indicating ‘good’ agreement across all skills when considering all raters.

Table 5

Weighted Kappa Scores between Raters and an Expert

|  | Skill | Rater 1 | Rater 2 | Rater 3 | Rater 4 | Rater 5 | Mean | SD |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| LFMS |  |  |  |  |  |  | **0.64** | 0.12 |
|  | Hop (17) | 0.71 | 0.77 | 0.50 | 0.65 | 0.70 | 0.67 | 0.09 |
|  | Jump (18) | 0.74 | 0.83 | 0.78 | 0.75 | 0.57 | 0.73 | 0.09 |
|  | Skip (18) | 0.67 | 0.76 | 0.59 | 0.29 | 0.35 | 0.53 | 0.18 |
| MFMS |  |  |  |  |  |  | 0.74 | 0.08 |
|  | Throw (19) | 0.92 | 0.79 | 0.65 | 0.71 | 0.69 | **0.75** | 0.10 |
|  | Kick (19) | 0.82 | 0.70 | 0.51 | 0.78 | 0.82 | 0.73 | 0.12 |
|  | Dribble (17) | 0.91 | 0.77 | 0.94 | 0.77 | 0.89 | 0.86 | 0.07 |
|  | Catch (20) | 0.89 | 0.84 | 0.83 | 0.82 | 0.76 | 0.83 | 0.04 |
|  | Strike (19) | 0.61 | 0.57 | 0.50 | 0.49 | 0.40 | 0.51 | 0.07 |
| TFMS |  |  |  |  |  |  | **0.70** | 0.10 |

*Note*. Kappa scores are displayed without 95% confidence intervals. The mean and SD columns represent each skill’s average kappa score and dispersion, respectively. For LFMS, MFMS, and TFMS, these represent averages of the skills under each category. Sample sizes for each skill are indicated in parentheses after the skill name.

# Discussion

This study aimed to establish the FG-COMPASS’s concurrent validity by comparing its results to those of the TGMD-2. The study also sought further evidence of the FG-COMPASS’s inter-, intra-, and expert-rater reliability.

## Concurrent Validity Between the FG-COMPASS and the TGMD-2

The concurrent validity assessment between the FG-COMPASS and the TGMD-2 provides insights into the agreement and potential discrepancies between the two tests. Establishing concurrent validity is important when testing motor skills because it shows how much a new instrument, like the FG-COMPASS, can be used in place of or with an established test, like the TGMD-2.

For the LFMS, the observed ICC of 0.68 falls within the ‘good’ range of agreement. This suggests that while both tests are aligned in their assessment of locomotor skills to a considerable extent, there are nuances or specific elements captured differently by each test. The Bland-Altman analysis further elucidates this observation. With a mean bias close to zero, there is virtually no systematic difference between the two tests regarding the locomotor subtest. The 95% limits of agreement, however, ranging from −1.95 to 1.95, show that individual scores can differ by almost two units (z-scores) in either direction. This range of disagreement suggests that, while the tests are generally aligned, they might occasionally produce different scores for the same individual.

The MFMS demonstrated a stronger alignment between the FG-COMPASS and TGMD-2, with an ICC of 0.89. This level of agreement underscores that both tests have a similar evaluation framework for manipulative skills. The Bland-Altman analysis further supports this, with a mean bias close to zero, indicating negligible systematic differences. The 95% limits of agreement, spanning from −1.25 to 1.25, are tighter than those of the locomotor subtest, indicating more consistent agreement between the tests for individual scores in this category. It should be noted that the FG-COMPASS has three locomotor and five manipulative skills compared to the TGMD-2, which has six skills in each subtest. This imbalance in the number of skills in the locomotor subtest between the two tests likely affected the observed difference. This can be verified in future studies since two new locomotor skills have recently been added (Perez, 2018) to the locomotor subtest of the FG-COMPASS.

For the TFMS, the ICC was also 0.89, mirroring the ‘excellent’ agreement observed in the manipulative subtest. The Bland-Altman analysis revealed a bias close to zero, again suggesting no meaningful systematic difference. The 95% limits of agreement ranged from −1.24 to 1.24, comparable to the manipulative subtest, showing a consistent level of individual score agreement across the entirety of the tests.

In light of these findings, the FG-COMPASS has acceptable concurrent validity with the TGMD-2, especially for assessing manipulative skills. Although there is ‘good’ agreement in the locomotor subtest, the results indicate that the FG-COMPASS can be reliably used in contexts where the TGMD-2 is the standard. However, practitioners should exercise caution and consider potential discrepancies, especially when assessing locomotor skills. The observed differences could be attributed to the inherent variations in test structures, scoring criteria, or specific motor skills targeted. Further investigation could explore these nuances to improve the FG-COMPASS or offer more specific direction on its use in conjunction with or instead of the TGMD-2.

## Inter-Rater Analysis

The FG-COMPASS’s inter-rater reliability assessment provided insights into the agreement of different raters’ ratings. The weighted kappa values showed ‘good’ agreement for the LFMS, ‘excellent’ for the MFMS subtests, and ‘good’ for the overall test (TFMS). The agreement for the manipulative skills (0.77) was slightly higher when compared to the locomotor skills (0.66) - see Table 2. This trend has been seen in previous research when measuring inter-rater reliability (Houwen et al., 2010; Valentini, 2012). In addition, a previous study (Furtado & Gallagher, 2018) found similar but higher kappa scores for LFMS (0.89), MFMS (0.88), and TFMS (0.89) when investigating the inter-rater reliability of the FG-COMPASS. These scores suggest an ‘excellent’ agreement. In the Furtado and Gallagher (2018) study, raters assessed performers based on videos, whereas in the present study, we assessed live performances.

For the LFMS subtest, ‘Skip’ achieved the highest mean reliability (0.74), followed by ‘Jump’ (0.73), suggesting a ‘good’ agreement among raters for both skills. This is consistent with Furtado and Gallagher (2018) study, in which the inter-rater reliability for ‘Jump’ was found to be 0.92. On the other hand, ‘Hop’ had a lower mean reliability of 0.50, implying more variability in raters’ scores. More training on the performance criteria for ‘Hop’ was necessary to avoid guessing when classifying the performers. A potential misunderstanding of the criteria may also have led to guessing and poor agreement among the raters. Thirdly, borderline performances may have led to disagreement among raters. Performances from individuals who are transitioning between levels are difficult to classify. After closely examining the recorded performances, it was noted that approximately two-thirds of the performers could be considered “borderline.”

Within the MFMS subtest, ‘Dribble’ and ‘Catch’ had particularly high agreement scores of 0.89 and 0.86, respectively. These values are similar to Furtado and Gallagher (2018) study and reinforce the robustness of the scoring system for these skills. In contrast, the ‘Throw’ skill had an interesting result. While the kappa value (0.62) was considered ‘good’, it had a significantly higher standard deviation (0.22) than the other skills. The asterisks in the Table 2 further highlight substantial disagreements between rater #5 and the other raters for this skill, suggesting potential inconsistencies or misinterpretations in the rating criteria or procedure for ‘Throw’. In the current study, the weighted kappa of 0.62 (good) would increase to 0.79 (excellent) if rater #5 was removed from the analysis. Rater #5 showed the largest disagreement compared to the other four raters (see Table 2). The estimated agreement values were 0.45, 0.35, 0.31, and 0.40 between rater #5 and raters 1, 2, 3, and 4, respectively. These values are considerably lower when compared to the pair agreement values for the other raters, which ranged from 0.71 to 0.85. Such a discrepancy suggests that rater #5 misunderstood the criteria associated with ‘Throw’. The kappa values for ‘Kick’ (0.68) and ‘Strike’ (0.78) are slightly lower when compared to the Furtado and Gallagher (2018) study (e.g., Kick = 0.90, Strike = 0.86). The reason for the difference can be traced back to the fact that the raters in the current study evaluated performances in a live setting. In contrast, the raters in the 2018 study evaluated performances recorded on video.

The TFMS combined score reliability of 0.72 reflects a balanced representation of the LFMS and MFMS subtests, adding to the tool’s reliability. However, it is important to address the discrepancies and variations among the raters, especially the pronounced differences with rater #5 for the skill of ‘Throw’. Such discrepancies could arise from several factors, including differing interpretations of scoring criteria, rater training, or even the subjective nature of certain skills. Future attempts to collect inter-rater reliability for the FG-COMPASS should give special attention to standardizing interpretations and emphasizing areas where disagreements are most apparent.

## Intra-Rater Reliability

Evaluating intra-rater reliability is crucial for assessing the consistency of measurements taken by raters at two different times, typically one week apart. Because the FG-COMPASS is being developed to assess motor skill performance in a live setting, this study was particularly focused on how live assessments compare with video recordings of the same performance. Following the live assessment of performances, which were recorded on video, participants returned to the lab twice to re-assess the recorded performances: one week after the live performances and again two weeks after the live performances. The mean values presented in Table 4 provide a comprehensive overview of each rater’s consistency across the three combinations: live versus recording - time 1 (IR1), live versus recording - time 2 (IR2), and recording time 1 versus recording time 2 (IR3). Based on Table 4, the average reliability score for the IR1 comparison is a noteworthy finding. This evaluation is especially interesting because it assesses the agreement between real-time ratings and those gathered from video recordings one week after the initial assessments.

The results indicate a ‘good’ agreement for LFMS (0.69), MFMS (0.70), and TFMS (0.70). Regarding individual skills, ‘Skip’ demonstrated a high mean kappa intra-rater reliability of 0.80 for IR1. This suggests an ‘excellent’ agreement between the live and the first video assessments for this skill. On the other hand, the skill ‘Jump’ had a lower mean reliability score of 0.60, suggesting a possible inconsistency in intra-rater reliability. The TFMS mean kappa value (0.70) observed between live ratings and video recordings is considered ‘good’. This value is lower than that obtained (0.91) in a previous study (Furtado & Gallagher, 2018), which assessed intra-rater reliability from video recordings. This difference was expected since assessing performance from videos is less challenging compared to live assessment.

## Raters vs. Expert Agreement for the FG-COMPASS

The expert-rater agreement is the degree to which raters in this study agreed with an expert familiar with the FG-COMPASS testing protocol. Thus, one of the aims of this study was to measure the level of agreement between the ratings of five raters and the expert’s video scores.

An average kappa value of 0.64 indicates that the agreement between the raters and the expert in the LFMS category was generally ‘good’. For ‘Hop’, the raters’ evaluations closely aligned with the expert, yielding a mean kappa value of 0.67. The ‘Jump’ skill exhibited even greater alignment between the raters and the expert, with a kappa value 0.73. On the other hand, the ‘Skip’ skill demonstrated a fair agreement level, with an average kappa value of 0.53. The ‘Hop’ and ‘Jump’ results indicate a shared interpretation of the evaluation criteria between the raters and the expert. The variability in kappa values for ‘Skip’ indicates potential differences in interpretation or understanding between the raters and the expert, which may need further clarification or training for the raters.

When investigating agreement between trained raters and an expert from video performances, Furtado and Gallagher (2012) found ‘excellent’ and ‘good’ agreement between raters and an expert for ‘Skip’ (0.77) and ‘Jump’ (0.70), respectively, and ‘excellent’ agreement for ‘Hop’ (0.85). It is important to note that the decision tree for “Jump’ was modified following the Furtado and Gallagher (2012) study and re-assessed recently (Furtado & Gallagher, 2018). The agreement improved from 0.70 (good) to 0.88 (excellent). The lower kappa score (0.73) observed in the current study for ‘Jump’ is likely due to rater variability and the source of observation, which in the current study was from live performance compared to video performance in the 2018 study.

A mean kappa value of 0.74 indicated an ‘excellent’ alignment between the raters and the expert in the MFMS category. The evaluations for the ‘Throw’ and ‘Kick’ skills aligned well with the expert’s ratings, resulting in ‘excellent’ (0.75) and ‘good’ (0.73) agreement levels, respectively. The scores for ‘Dribble’ and ‘Catch’ had higher scores, with mean kappa values of 0.86 and 0.83, respectively. Like ‘Jump’, the decision trees for ‘Dribble’ and ‘Catch’ were modified following the Furtado and Gallagher (2012) study and re-assessed in the Furtado and Gallagher (2018) study. The kappa values improved from 0.72 (good) to 0.81 (excellent) and 0.72 (good) to 0.94 (excellent) for ‘Dribble’ and ‘Catch’, respectively. For the current study, ‘Strike’ displayed fair agreement (0.51). In its initial assessment (Furtado & Gallagher, 2012), the kappa value for ‘Strike’ was 0.79. This value, however, significantly improved to 0.93 in a subsequent study (Furtado & Gallagher, 2018). Similar to ‘Skip’, the differences in the kappa values for ‘Strike’ suggest that there may be discrepancies in how the raters and the expert interpret or understand the test’s protocol or criteria, which may require additional clarification or training for the raters in future studies.

When looking at all eight FG-COMPASS skills, an average kappa value of 0.70 for the total test (TFMS) suggests a good agreement between the expert and the raters. The empirical findings highlight the significance of familiarizing oneself with the test protocol before utilizing the FG-COMPASS, particularly in skills where agreement with the expert was moderate (e.g., skip and strike).

# Conclusion

This study looked closely at the FG-COMPASS. It focused on how well it worked with the TGMD-2, how reliable it was between and within raters, and how much an expert and raters agreed. The results demonstrate promising evidence of the FG-COMPASS’s applicability and robustness as a motor skill assessment tool. The assessment of concurrent validity with the TGMD-2 shows an excellent match between the two instruments. This shows that the FG-COMPASS could be a good alternative or supplementary tool. The ‘good’ agreement in the locomotor subtest calls for attention to specific nuances, which may be further explored in future studies. The evaluation of inter-rater reliability reveals ‘good’ agreement for LFMS, ‘excellent’ for MFMS, and ‘good’ for the overall test (TFMS). The discrepancies observed, especially with rater #5 for ‘Throw’, underline the importance of standardized interpretations and training among raters. The intra-rater reliability assessment, which explored consistency in live versus video assessments, indicates ‘good’ agreement. The variance in some skills, such as ‘Jump’, highlights the complexities of live assessment and warrants further examination. Finally, the analysis of agreement between raters and an expert emphasizes the significance of understanding and aligning with the test protocol. The agreement for LFMS was deemed ‘good’, while MFMS received an ‘excellent’ rating, resulting in an overall ‘good’ agreement. This reinforces the FG-COMPASS’s ability to provide a strong agreement between FMS classifications from raters and a ‘gold standard’.

## Implications and Future Directions

The findings collectively support the FG-COMPASS as a valuable instrument for motor skill assessment. The observed discrepancies in some areas provide constructive insights for refining the test, ensuring better alignment with established tests like the TGMD-2, and enhancing rater training and standardization. Future research should delve into the subtle differences in the locomotor subtest and seek to understand the variations in rater interpretations. Enhancing the FG-COMPASS through ongoing refinement and validation is an important step toward more accurate and consistent motor skill evaluations in research and practical settings. This study contributes to the growing body of knowledge supporting FG-COMPASS’s efficacy. It sets the stage for its broader implementation and potential impact on motor skill assessment.

# Limitations

Although precautions were taken, it’s important to note that this study has some limitations. Compared to the TGMD-2, which consists of six locomotor and six manipulative skills, the FG-COMPASS has three locomotor and five manipulative skills. This discrepancy may have affected the alignment between the two tests, particularly in the locomotor subtest. Second, the discrepancies observed with certain raters, such as rater #5, may point to inconsistencies in rater training or interpretation of the criteria. This could affect the overall reliability of the ratings. Third, the variability in kappa values between raters and the expert for certain skills indicates potential differences in interpretation or understanding of the evaluation criteria. This may warrant further clarification or training. Fourth, due to the small sample size, there may be limitations in generalizing the findings to broader populations.

## 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/>

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. *The Journal of Adolescent Health: Official Publication of the Society for Adolescent Medicine*, *44*(3), 252–259. <https://doi.org/10.1016/j.jadohealth.2008.07.004>

Bayer, O., Bolte, G., Morlock, G., Rückinger, S., von Kries, R., & GME-Study Group. (2009). A simple assessment of physical activity is associated with obesity and motor fitness in pre-school children. *Public Health Nutrition*, *12*(8), 1242–1247. <https://doi.org/10.1017/S1368980008003753>

Bellows, L. L., Davies, P. L., Anderson, J., & Kennedy, C. (2013). Effectiveness of a physical activity intervention for head start preschoolers: A randomized intervention study. *AJOT: American Journal of Occupational Therapy*, *67*(1), 28+. <http://go.galegroup.com/ps/i.do?id=GALE%7CA314443342&v=2.1&u=csunorthridge&it=r&p=HRCA&sw=w&asid=9496686804b2cdbe2f7f6b6205138068>

Bland, J. M., & Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, *327*(8476), 307–310. <https://doi.org/10.1016/S0140-6736(86)90837-8>

Castelli, D. M., & Valley, J. A. (2007). Chapter 3: The Relationship of Physical Fitness and Motor Competence to Physical Activity. *Journal of Teaching in Physical Education*, *26*(4), 358–374. <http://www.eric.ed.gov/ERICWebPortal/detail?accno=EJ776156>

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. <https://doi.org/10.1037/1040-3590.6.4.284>

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>

Cools, W., Martelaer, K. D., Samaey, C., & Andries, C. (2009). Movement Skill Assessment of Typically Developing Preschool Children: A Review of Seven Movement Skill Assessment Tools. *Journal of Sports Science & Medicine*, *8*(2), 154–168. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3761481/>

Fisher, A., Reilly, J. J., Kelly, L. A., Montgomery, C., Williamson, A., Paton, J. Y., & Grant, S. (2005). Fundamental Movement Skills and Habitual Physical Activity in Young Children. *Medicine & Science in Sports & Exercise*, *37*(4), 684–688.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, *33*(3), 613–619. <https://doi.org/10.1177/001316447303300309>

Furtado, O., Jr. (2004). *Furtado-Gallagher movement skill assessment: Collecting evidence for content-related validity* [Master’s thesis, University of Pittsburgh]. <http://etd.library.pitt.edu/ETD/available/etd-12042004-093510/>

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/10.1080/02701367.2012.10599872>

Furtado, O., Jr, & Gallagher, J. D. (2018). Expert-rater agreement and inter-/intrarater reliability for the Furtado–Gallagher Computerized Observational Movement Pattern Assessment System. *Perceptual and Motor Skills*, *125*(3), 423–437. <https://doi.org/10.1177/0031512518769205>

Gabbard, C. (2021). *Lifelong Motor Development*. Lippincott Williams & Wilkins.

Giavarina, D. (2015). Understanding Bland Altman analysis. *Biochemia Medica*, *25*(2), 141–151. <https://doi.org/10.11613/BM.2015.015>

Graf, C., Koch, B., Kretschmann-Kandel, E., Falkowski, G., Christ, H., Coburger, S., Lehmacher, W., Bjarnason-Wehrens, B., Platen, P., Tokarski, W., Predel, H. G., & Dordel, S. (2004). Correlation Between Bmi, Leisure Habits and Motor Abilities in Childhood (chilt-Project). *International Journal of Obesity*, *28*(1), 22–26. <https://doi.org/10.1038/sj.ijo.0802428>

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. <http://journals.humankinetics.com/AcuCustom/Sitename/Documents/DocumentItem/17945.pdf>

Lemos, A. G. (2012). Physical Education in Kindergarten Promotes Fundamental Motor Skill Development. *Advances in Physical Education*, *02*(01), 17–21. <https://doi.org/10.4236/ape.2012.21003>

Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, *76*(5), 365–377. <https://doi.org/10.1037/h0031643>

Lopes, V. P., Stodden, D. F., Bianchi, M. M., Maia, J. A. R., & Rodrigues, L. P. (2012). Correlation between BMI and motor coordination in children. *Journal of Science and Medicine in Sport*, *15*(1), 38–43. <https://doi.org/10.1016/j.jsams.2011.07.005>

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. <https://doi.org/10.1037/1082-989X.1.1.30>

McKenzie, T. L., Sallis, J. F., Broyles, S. L., Zive, M. M., Nader, P. R., Berry, C. C., & Brennan, J. J. (2002). Childhood Movement Skills: Predictors of Physical Activity in Anglo American and Mexican American Adolescents. *Research Quarterly for Exercise and Sport*, *73*(3), 238–244. <https://doi.org/10.1080/02701367.2002.10609017>

Perez, L. (2018). *Investigating expert-rater agreement and inter/intra-rater reliability of two fundamental movement skills for the locomotor subscale of the fg-compass* [Thesis, California State University, Northridge]. <http://dspace.calstate.edu/handle/10211.3/199459>

Provost, B., Crowe, T. K., & McClain, C. (2000). Concurrent validity of the Bayley Scales of Infant Development II Motor Scale and the Peabody Developmental Motor Scales in two-year-old children. *Physical & Occupational Therapy in Pediatrics*, *20*(1), 5–18. <http://www.ncbi.nlm.nih.gov/pubmed/11293915>

The jamovi project. (2022). *Jamovi* (Version 2.3.3). <https://www.jamovi.org>

Ulrich, D. (2000). *Test of gross motor development: Examiner’s manual*. Pro-Ed.

Valentini, N. C. (2012). Validity and Reliability of the TGMD-2 for Brazilian Children. *Journal of Motor Behavior*, *44*(4), 275–280. <https://doi.org/10.1080/00222895.2012.700967>

Valentini, N. C., & Rudisill, M. E. (2004). An Inclusive Mastery Climate Intervention and the Motor Skill Development of Children with and Without Disabilities. *Adapted Physical Activity Quarterly*, *21*(4), 330–347. <https://doi.org/10.1123/apaq.21.4.330>

Wiart, L., & Darrah, J. (2001). Review of four tests of gross motor development. *Developmental Medicine and Child Neurology*, *43*(4), 279–285. <https://doi.org/10.1017/s0012162201000536>

Woods, A. M., Bolton, K. N., Graber, K. C., & Crull, G. S. (2007). Chapter 5: Influences of Perceived Motor Competence and Motives on Children’s Physical Activity. *Journal of Teaching in Physical Education*, *26*(4), 390–403. <https://doi.org/10.1123/jtpe.26.4.390>