Collecting Concurrent Validity, Reliability of Classification Decisions, Inter and Intra-Rater Reliability Evidence for the Furtado-Gallagher Computerized Observational Movement Pattern Assessment System

Mackinsey Woolever, M.S., & Ovande Furtado, Jr., Ph.D.

California State University - Northridge

ABSTRACT

Fundamental movement skills (FMS) are considered the building blocks for the development of specialized skills (Gabbard, 2007). In addition, FMS competency has been linked to decreased levels of obesity (Lopes et al., 2012) and increased levels of physical activity/sport participation (Castelli & Valley, 2007). Thus, teachers and practitioners working with younger children must conduct regular assessments to gather evidence about the student's level of achievement in FMS development. This study aimed to collect evidence for concurrent validity, reliability of classification decisions, and inter/intra rater reliability for the FG-COMPASS. The FG-COMPASS is a process-oriented and criterion-referenced instrument that assesses FMS development in children ages 5 to 10 years. The Test of Gross Motor Development 2 (TGMD-2) was used as the "gold standard" for concurrent validity. Participants were 34 children between the ages of 5 and 10 years. Pearson partial correlations (controlling for age) comparing the scores of both tests indicate a moderate to strong correlation for locomotion ($r_{xy.z} = .52, p < .01$), object manipulation ($r_{xy.z} = .59, p < .001$), and total scores ($r_{xy.z} = .63, p < .001$). The reliability of classification decisions was assessed by comparing the live score ratings of five raters with the primary investigator's video scores on 10 randomly selected children per skill. The inter-rater reliability was assessed by comparing the scores across the FG-COMPASS 5 raters. The intra-rater reliability was assessed by comparing each FG-COMPASS rater's live scores to their video scores. The weighted kappa values ranged from .51 to .83, .50 to .89, and .60 to .87 for reliability of classification decisions, inter and intra rater reliability, respectively. The results of this study provide further validity and reliability evidence for the FG-COMPASS. Further studies involving children with different ethnicity backgrounds and a larger sample size is recommended.

PURPOSES

The purpose of this study was to collect criterion-related (concurrent) validity for the FG-COMPASS by comparing its results (total and both subtest scores) to the results of the TGMD-2.

A second purpose of this study was to collect inter and intra-rater reliability for the FG-COMPASS.

HYPOTHESES

- 1) There will be a positive "strong" (.50 or higher) correlation between the total scores of both assessment tools.
- 2) There will be a positive strong correlation between the locomotor and object manipulation scores between both assessment tools.
- 3) There will be a "good" (weighted kappa = .61 or higher) inter-rater and intra-rater agreement for the total, locomotor, and object manipulation scores for the FG-COMPASS.

METHODS

Participants

- Twenty-two girls ($M = 8.14, SD = 1.78$) and twelve boys ($M = 8.44, SD = 1.49$)
- K-6 public school located in Southern California
- Procedures were approved by the LAUSD Committee of External Research Review and the CSUN Committee for the Protection of Human Subjects

METHODS cont.

Instruments

TGMD-2

- Children between the ages of 3 and 10 years
 - Assesses 12 gross motor skills (6 locomotor and 6 object manipulation)
- #### FG-COMPASS
- Children between the ages of 5 and 10 years
 - Assesses 8 gross motor skills (3 locomotor and 5 object manipulation)

Research Assistants

Sixteen kinesiology undergraduate students

- Ten raters (5 FG-COMPASS and 5 TGMD-2)
- Five Test Administrators (2 FG-COMPASS and 3 TGMD-2)
- Two Video Editors

Procedures

Data Collection

- Participants were assigned identification numbers (01 – 41).
- Followed a three-step process
 - (1) Each participant held up their identification number card and stated "I am (ID number)" to the camera before any instructions, demonstrations, or any skill performances began.
 - (2) Test administrators provided each participant with instructions and demonstrations for the skill and asked for any questions.
 - (3) The participant completed their performance trials, which were used for scoring.

- Data collection occurred over an eight-week period, with each week designed to match similar skills as close as possible between the two assessment tools.

TGMD-2

- Only the test administrators attended the testing site each week.
- Test administration procedures were outlined in the test's user manual and were followed closely including suggested materials and directions.
- All performances were video recorded for the rater's to refer to when scoring for the TGMD-2.
- All skill assessments sessions occurred within one week of the initial performance.

FG-COMPASS

- Raters and test administrators attended the testing site each week.
- Test administration procedures were outlined in the test's user manual and were followed closely including suggested materials and directions.
- All performances were recorded for the rater's to reassess twenty randomly selected participants.
- Raters recorded a live score for all participants and were asked to reassess twenty randomly selected participants per skill an additional two times. The first reassessment occurred one week from initial performance and the second reassessment occurred two weeks from initial performance.

Data Analysis

- Inter/intra rater analysis for the FG-COMPASS was assessed using weighted kappa statistics.
- Inter rater analysis for the TGMD-2 was assessed using the two-way random average measures intraclass correlation coefficient ICC(2,5).
- Raters vs. expert consensus for the FG-COMPASS was assessed using weighted kappa. Raters vs. expert consensus for the TGMD-2 was assessed using ICC(2,1).
- Concurrent validity was assessed using both partial (controlling for age) and zero-order Pearson product-moment correlation coefficient.

Results

Inter-rater reliability for the FG-COMPASS

- Weighted kappa values ranged from .50 to .74 ($M = .66$) for the locomotion subtest, .62 to .89 ($M = .77$) for the object manipulation subtest, and .73 for the total test.
- These results indicate a "good" agreement (Altman, 1991) for the locomotion, object manipulation subtests as well as for the total test.

Inter-rater reliability for the TGMD-2

- The ICC values ranged from .74 to .95 ($M = .87$) for the locomotion, .84 to .95 ($M = .91$) for the object manipulation subtests, and .74 to .95 ($M = .89$) for the total test.
- These results indicate an "excellent" agreement (Fleiss, 2011).

Intra-rater reliability for the FG-COMPASS

- The averaged weighted kappa values (Live vs. Video 1) ranged from .60 to .80 ($M = .69$) for the locomotion subtest, .64 to .87 ($M = .79$) for the object manipulation subtest, and .76 for the total test.
- These results indicate a "good" agreement for the locomotion, object manipulation subtests as well as for the total test.
- The consistency between Video 1 vs. Video 2 was relatively high with averaged scores ranging from .65 to .89 ($M = .78$) for the locomotion subtest, from .73 to .96 ($M = .84$) for the object manipulation subtest, and .81 for the total test.

Rater vs. expert consensus for the FG-COMPASS

- Weighted kappa scores ranged from .51 to .73 ($M = .63$) for the locomotion subtest, .52 to .83 ($M = .72$) for the object manipulation subtest, and .69 for the total test.
- These results indicate a "good" agreement for the locomotion, object manipulation subtests as well as for the total test.

Raters vs. expert consensus for the TGMD-2

- The ICC values ranged from .50 to .95 ($M = .68$) for the locomotion subtest, .67 to .92 ($M = .83$) for the object manipulation subtest, .75 for the total test.
- These results indicate an "excellent" agreement for the object manipulation subtest, while the value of the locomotion subtest and total test indicates "fair to good" agreement.

Correlations between the FG-COMPASS and the TGMD-2

- Pearson partial product-moment correlations comparing the scores of both tests indicate a moderate to strong correlation (see Table 1) for locomotion ($r_{xy.z} (31) = .52, p < .01$), object manipulation ($r_{xy.z} (31) = .59, p < .001$), and total scores ($r_{xy.z} (31) = .63, p < .001$)

Table 1. Correlations between the FG-COMPASS and the TGMD-2

FG-COMPASS	TGMD-2 (criterion)		
	Locomotion	Object manipulation	Total score
Locomotion	.52* (.60**)		
Object manipulation		.59** (.81**)	
Total score			.63** (.82**)

Notes: numbers inside of parentheses denote Pearson zero-order correlations, whereas numbers outside parentheses denote Pearson partial correlations with age (in years) as the covariate.

* $p < .01$; ** $p < .001$

Correlations between similar skills for the FG-COMPASS and TGMD-2

- There are seven common skills between the FG-COMPASS and the TGMD-2
- Pearson partial product-moment correlations
 - A "strong" correlation was found for the skill of hop ($r_{xy.z} (31) = .60, p < .001$)
 - A "moderate" correlation was found for overhand throw ($r_{xy.z} (31) = .43, p < .05$), hand dribble ($r_{xy.z} (31) = .37, p < .05$), catch ($r_{xy.z} (31) = .46, p < .05$), and batting ($r_{xy.z} (31) = .37, p < .05$)
 - A "small" correlation was found for jump ($r_{xy.z} (31) = .12, p > .05$) and kick ($r_{xy.z} (31) = .24, p > .05$).

Discussion & Conclusions

- The results supported the three hypotheses proposed for this study.
- This was the first study to test the concurrent validity of the FG-COMPASS; therefore, more studies are needed to support the results.
- Some studies (Sun et al., 2011) have used the Pearson zero-order product-movement correlation index when comparing the results of their tests with the TGMD-2.
- This study took a more conservative approach by relying on the Pearson partial product-moment correlation index to control for age in years. We believe this a more accurate estimation of the correlation between the two tests as age may influence the association.
- The results of this study show that the FG-COMPASS can be viewed as a reliable and valid assessment tool that can be used in a live setting to assess gross motor performances in children between the ages of 5 and 10 years.

Selected References

- Altman, D. (1991). *Practical statistics for medical research* (1st ed.). New York: Chapman and Hall.
- 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.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Psychology Press.
- Fleiss, J. L. (2011). *Design and analysis of clinical experiments*. John Wiley & Sons.
- Gabbard, C. (2007). *Lifelong motor development* (5th ed.). San Francisco, CA: Benjamin Cummings.
- Furtado, O., Jr. (2004). *Furtado-Gallagher Movement Skill Assessment: Collecting evidence for content-related validity* (Master's Thesis). University of Pittsburgh, Pittsburgh, PA. Retrieved from <http://etd.library.pitt.edu/ETD/available/etd-12042004-093510/>
- 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. <http://doi.org/10.1016/j.jsams.2011.07.005>
- Sun, S.-H., Sun, H.-L., Zhu, Y.-C., Huang, L., & Hsieh, Y.-L. (2011). Concurrent validity of preschooler gross motor quality scale with Test of Gross Motor Development-2. *Research in Developmental Disabilities*, 32(3), 1163–1168. <http://doi.org/10.1016/j.ridd.2011.01.007>
- Ulrich, D. A. (2000). *Test of gross motor development examiner's manual* (2nd Ed.). Austin, TX: Pro-ed.

Contact Information

Ovande Furtado, Jr., Ph.D.
Cal State University, Northridge
Kinesiology Department
18111 Nordhoff Street
Northridge, CA 91330

T: (818)-677-5968
E: ovande.furtado@csun.edu
<http://fgcompass.com>

goo.gl/GBy37h

Poster presented at the 2016 NASPSPA Conference (June 15-18) – Montreal, Quebec, CA.