European Journal of Adapted Physical Activity, 10(2), 20–30
European Federation of Adapted Physical Activity, 2017

Validity and Reliability Evidence of Smart Start in Preschoool-aged Children with/without a Developmental Delay and/or a Disability

Jaehun Jung[1], So-Yeun Kim[2], Lauriece L. Zittel[2], and Marilyn A. Looney[2]

[1]Oregon State University, USA; [2]Northern Illinois University, USA

The purpose of this study was to investigate validity and reliability evidence of Smart Start in male (n =35) and female (n = 25) children with/without a disability. Fundamental movement skills (FMS) of preschoolers (with a disability, n = 32; and without a disability, n = 28) aged between 37 to 64 months were examined using Smart Start and the Test of Gross Motor Development-Second Edition. The correlation between total scores of the two instruments was r = .89, p < .01. Across three raters, the average percentages of agreement, modified kappa coefficients, and intraclass correlation coefficients (ICCs) for intra-rater reliability using Smart Start on all participants were .92, .83, and .96, respectively. For inter-rater reliability, the average percentages of agreement, modified kappa coefficients, and ICCs based on all participants were .86, .71, and .93, respectively. The major findings provide preliminary evidence to support concurrent validity and intra- and inter-rater reliability of the Smart Start for assessing FMS of preschoolers with/without a disability.

Keywords: assessment, fundamental movement skills, preschooler

Introduction

Fundamental movement skills (FMS) are crucial for children in early childhood because children use these skills to interact with and respond to environmental challenges (Gallahue & Ozman, 2006). Evidence suggests that children with well-developed FMS tend to be more engaged in physical activity than children with low FMS (Cliff, Okely, Smith, & McKeen, 2009; Fisher et al., 2005; Graf et al., 2004; Okely, Booth, & Patterson, 2001; Williams et al., 2008; Wrotniak, Epstein, Dorn, Jones, & Kondilis, 2006). Inadequate proficiency of FMS in early childhood can have negative effects on physical activity competency later in life (Gallahue & Ozman, 2006). Also, FMS, especially manipulative skills, are predictors of adolescent physical activity (Barnett, van Beurden, Morgan, Brooks, & Beard, 2009).

FMS assessment is essential for preschoolers with a developmental delay and/or a disability. Preschoolers with a developmental delay and/or a disability may have difficulty acquiring the motor skill competency necessary to perform FMS in early childhood, thus putting them at risk for poor physical, social, and emotional development (Majnemer, 1998). Given the potential long-term, negative effect of delayed motor competency for children with developmental delays and/or disabilities, intervention and instruction to develop FMS competence is strongly recommended (Stodden et al., 2008).

Two standardized instruments are available to assess FMS of preschool-aged children with and without a developmental delay and/or a disability. Test of Gross Motor Development-Second Edition (TGMD-2; Ulrich, 2000) and Peabody Development Motor Scales-Second Edition (PDMS-2; Folio & Fewell, 2000) have been widely used for diagnosis and educational placement decisions for children in this preschool group (e.g., Hardy, King, Farrell, Macniven, & Howlett, 2010) with a developmental delay and/or a disability (e.g., Capio, Sit, & Abernethy, 2011; Houwen, Hartman, Jonker, & Visscher, 2010; Simons et al., 2008; Wuang, Wang, Huang, & Su, 2008).

Standardized assessment tools are particularly useful when making decisions for educational eligibility and placement. However, there are a few shortcomings of standardized tests, which have been recognized. The results of standardized tests can be inaccurate, particularly when administered to children with specific disabilities, a developmental delay, or who are at risk because the tests are primarily normed on typically developing children (McLean, 2005).

Furthermore, data from standardized tests may not be useful once eligibility has been determined. Especially, if the test items in the norm-referenced tests were

designed to discriminate between groups of children, rather than to reflect the knowledge or skills that children need to achieve (Neisworth & Bagnato, 2004). In other words, standardized tests may lack instructional relevance (Gickling & Thompson, 1985; Morison, White, & Feuer, 1996). The test items in these standardized tools may not reflect the instructional curriculum, thus may create a gap between assessment and instruction (Gickling & Thompson, 1985). Consequently, classroom teachers could have difficulty using data from these tests to identify specific goals and objectives for program and instructional planning.

Curriculum-based assessment (CBA) has advantages to link assessment to curriculum and instruction. According to Deno (1987), CBA can be defined as "direct observation and recording of a student's performance in the local curriculum as a basis for gathering information to make instructional decisions" (p. 41). The main purpose of using CBA is to determine the instructional needs of the learner (Burns, MacQuarrie, & Campbell, 1999). CBA test items are sampled from the curriculum (Potter & Wamre, 1990) and the test results are used to compare a child's performance to his or her previous performance criteria, rather than to compare a child's performance to peers or normative data (Burns et al., 1999).

Using this educational perspective, Smart Start Preschool Movement Curriculum (Smart-Start) (Wessel & Zittel, 1995) was developed. This curriculum-based and criterion-referenced instrument can be useful in designing movement programs for preschoolers with all abilities. This movement curriculum also provides help for teachers planning instruction responsive to the unique needs and interests of preschoolers with/without disabilities (Wessel & Zittel, 1995) and enable teachers to use developmentally appropriate practices to observe and assess preschoolers with/without disabilities.

Although Smart Start has been used to help preschool classroom teachers and physical educators plan and implement movement programs, there has been limited research on validity and reliability evidence for Smart Start. One study by Ong (2001) examined the reliability evidence of Smart Start for 28 preschoolers aged 36 to 72 months with/without a developmental delay and/or disability. The results of the study indicated that Smart Start had good inter-rater reliability (r = .77–1.00) and intra-rater reliability (r = .84–1.00) for seven locomotor skills (crawl/creep, walk, run, jump down, jump over, hop, and gallop).

There are a few limitations in the study by Ong (2001). First, the study examined rater reliability evidence for seven locomotor skills (crawl/creep, walk, run, jump down, jump over, hop, and gallop) although Smart Start also includes object control, orientation, and play skills. In order to determine FMS of preschoolers, both locomotor and object control skills should be examined. In this study, object control skills including strike, bounce, catch, kick, overhand throw, underhand throw, and roll a ball were included in order to establish intra- and inter-rater reliability evidence. The second limitation of Ong's study is that it focused on investigating rater reliability evidence, and did not examine validity evidence for Smart Start. Currently, Smart Start has content-related validity. According to Yun and Ulrich (2002), validity evidence about the data or inferences made based on the results of measurements cannot be provided by content validity alone. In this study, concurrent validity evidence involving preschoolers with and without a disabilities were investigated by comparing Smart Start and TGMD-2 (Ulrich, 2000).

The purpose of this study, therefore, was to investigate concurrent validity evidence and inter-rater and intra-rater reliability of Smart Start for preschoolers with and without a developmental delay and/or a disability. The Smart Start content used in this study included the updated locomotor and the updated object control checklists. Preschoolers with a disability in this study were defined as those with a developmental delay and/or a specific disability as well as those at risk for a developmental delay. High correlations (r > .7) between the total scores of Smart Start and TGMD-2 were hypothesized. For intra- and inter-rater reliability evidence, it was hypothesized that high percentage of agreement (above 80%), high modified kappa coefficients (k > .7), and high intraclass correlation coefficients (ICCs; R > .8) would be obtained after completing the rater training.

Method

Participants

A total of 60 participants (19 preschoolers with a developmental delay, 13 preschoolers at risk for a developmental delay who were randomly selected from 44 participants; and 28 preschoolers without a disability) were recruited from two early childhood centers in the northern Illinois area. Recruitment and data collection were conducted after receiving approval from a University Institutional Review Board (IRB). All the participants' parents or guardians signed an informed consent form in order for their child

to participate in this study, and participants gave their verbal assent prior to participating in this study.

Thirty-five boys and 25 girls participated in this study. The mean age of all participants in this study was 50.98 months (*SD* = 7.86, range 37–64). On average, the preschoolers in this sample with disability were almost five months older than those without disability. Most of the participants in this study were Caucasian (71.6%) and African American (16.7%). The directors of two early childhood centers confirmed the participants' disability diagnosis. The demographic characteristics of each group are presented in Table 1.

A developmental delay is defined as a clinical presentation with various disabilities related to age-specific deficits in learning skills and adaptation (Shevell et al., 2003). Preschoolers with a developmental delay may or may not have other disabilities, such as cerebral palsy and certain neuromuscular disorders. The program directors of the two early childhood centers identified participants' personal disability diagnosis information; however, the investigator did not have access to each participant's personal disability diagnosis information due to school district policy. In this study, children with a developmental delay, who have a physical disability, were not recruited.

Instruments

Smart Start: Preschool Movement Curriculum for Children of All Abilities (Smart-Start)

Smart Start is a movement curriculum designed for classroom teachers to observe and assess preschoolers as well as plan instruction responsive to the needs and interests of children of all abilities (Wessel & Zittel, 1995). Smart Start includes teaching materials for implementing a curriculum as well as curriculum-based and criterion-referenced assessment checklists for locomotor skills, object control, orientation, and play skills.

Smart Start has a few unique features as a CBA for preschool children. First, each child's assessment data can be used for planning instructional lessons. Each Smart Start skill is task-analyzed into specific key elements that preschoolers need to achieve, and the key elements become the instructional objectives for each program planned (Wessel & Zittel, 1995). Second, Smart Start offers practical procedures to develop and design effective instructional programs for all children. To assist teachers to develop and design an effective instructional program, the authors present the three Cs of curriculum design: Content, Construction, and Contact. The three C's provide teachers with direction for designing appropriate movement environments. The Locomotor, Orientation, Object control, and Play participation (LOOP) (Wessel & Zittel, 1995) model directly outlines the content related to the program goals and objectives.

Some Smart Start checklists have been recently revised, and the updated locomotor and object control checklists were used in this study. These checklists include 14 FMS divided into two subscales: locomotor (run, gallop, hop, leap, horizontal jump, jump down, and slide) and object control (strike, bounce, catch, kick, overhand throw, underhand throw, and roll a ball). Because the TGMD-2 does not include jump

Table 1
Demographic Characteristics of the Sample (Percentage)

		Groups		
		Total (N = 60)	PWD (n = 32)	PWOD (n = 28)
Ethnicity	Caucasian	71.6	71.9	71.4
	African American	16.7	12.5	21.4
	Hispanic	8.3	15.6	0.0
	Asian	3.4	0.0	7.2
Age (Month)	36–47	36.7	25.0	50.0
	48–59	46.7	50.0	42.9
	60–71	16.6	25.0	7.1
Sex	Male	58.3	59.4	57.1
	Female	41.7	40.6	42.9

Note. PWD = Preschoolers with a disability; PWOD = Preschoolers without a disability.

down and underhand throw skills, the investigator omitted the skills of jump down and underhand throw in this study.

Each movement skill has 3 to 5 key elements. Participant performance was rated as "1" if the key element within the skill was performed or "0" if the key element was not performed. Participants were asked to perform each skill twice. When the child displayed the key element during both trials, the examiner scored a "1". Raw scores for each subscale were calculated by summing participant's points. The highest possible total raw score for the locomotor subscale was 27 and for the object control skill subscale it was 30.

Test of Gross Motor Development-Second Edition (TGMD-2)

TGMD-2 is a norm- and criterion-referenced test that is designed to assess fundamental gross motor skills of children 3 to 10 years of age (Ulrich, 2000). This study used TGMD-2 as a criterion measurement because TGMD-2 has been examined for content, construct, and criterion validity evidence provided by the test author (Ulrich, 2000) and other researchers (e.g., Houwen et al., 2010; Simons et al., 2008). TGMD-2 has been widely used to assess FMS for typically developing children (e.g., Hardy et al., 2010; Robinson & Goodway, 2009), individuals with cerebral palsy (e.g., Capio et al., 2011), school-age children with a visual impairment (Houwen et al., 2010), and school-age children with intellectual disabilities (e.g., Simons et al., 2008).

Evidence of validity and reliability has been established for TGMD-2. Construct validity evidence of TGMD-2 has been reported in typically developing children from 3 to 10 years of age (e.g., Evaggelinou, Tsigilis, & Papa, 2002) as well as in children with a visual impairment from 6 to 12 years of age (e.g., Houwen et al., 2010). Additionally, an acceptable level of internal consistency, test-retest reliability and inter-rater reliability evidence were established when TGMD-2 was administered in children with intellectual disabilities from 7 to 10 years of age (Cronbach's $\alpha = .85$ and .88 for locomotor and object control skills, respectively) (Simons et al., 2008).

For the present study, each skill of TGMD-2 was measured twice. The skill's performance was evaluated by the sum of the points earned based on predetermined performance criteria for each skill (3 to 5 criteria, depending on skill). When a participant displayed the criteria correctly, one point was given for each predetermined performance. The highest possible total raw score for the locomotor was 48 and object control skills was 48. The raw scores can be converted into percentile ranks and standard scores.

Procedure

Data collection

All participants were asked to perform 12 FMS (run, gallop, hop, leap, horizontal jump, slide, strike, bounce, catch, kick, overhand throw, and roll a ball) in the indoor facilities of the two early childhood centers. Using a camcorder, their performances were recorded. Instructions for each skill were given and the investigator demonstrated the proper technique (described in both TGMD-2 and Smart Start) before each test.

Rater training

In addition to the primary investigator (rater 1), two graduate students (rater 2 and rater 3) with rater experience in administering TGMD-2 and teaching preschoolers with/without a disability were recruited for inter-rater reliability evidence. All three raters received training in scoring Smart Start and TGMD-2 prior to the start of data analysis to achieve both accuracy and consistency.

The training was based on the rater training procedure developed by Darst, Zakrajsek, and Mancini (1989). This training procedure was designed for systematic observation tools and consisted of five phases:

a) orientation to the system,
b) learning the categories,
c) using the coding form correctly,
d) initial coding practice, and
e) live observation practice.

The training procedure has been used in other studies by Brewer and Jones (2002), Partington and Cushion (2013), and Roberts and Fairclough (2012).

For the present study, the training procedure was modified. There were three sessions including

a) an introductory session to describe the instrument in detail (orientation to each instrument including the coding form),
b) a review of key elements (learning the skills), and
c) coding practice to establish intra- and inter-rater reliability using training video clips.

Live observation practice was not necessary for this study since actual data analysis was for coding video clips. The three raters received training for Smart Start first and then the same protocol was followed to train on TGMD-2.

In order to initiate actual data analysis, a percentage of agreement had to be at least 85%, and the modified kappa coefficient had to be estimated at

.70 or higher. During the first coding practice, disagreements among the three raters were found. To resolve these disagreements among the raters, discussion and formal seminars were conducted. After the discussion and seminars, the second coding practice was delivered. All raters met the criteria to initiate actual data coding in the second coding practice.

Data coding

The primary investigator, rater 1, scored all participants' (N = 60) videotaped performances twice using Smart Start and TGMD-2 with an interval of one week between the first and second scoring sessions. The two other raters each scored videotaped performances of 30 participants (14 from the group of preschoolers without a disability and 16 from the group of preschoolers with a disability) using Smart Start and TGMD-2 with an interval of one week between the first and the second scoring sessions.

Data coding was completed by the three raters in four weeks. During the four weeks, each rater independently scored the participants' videotaped performances using either Smart Start or TGMD-2. In the first week, raters 1 and 2 scored the participants' performances using Smart Start, while rater 3 scored 30 participants' performances using TGMD-2. In the second week, raters 1 and 2 scored the participants' performances using TGMD-2 and rater 3 scored participants' performances using Smart Start. In order to establish intra-rater reliability for Smart Start, the same protocol was followed during the third and fourth weeks.

Data Analysis

Means and SDs of raw scores for the participants' FMS using Smart Start and TGMD-2 were calculated by using SPSS (Version 20.0, IBM Corp. Released, 2011) statistical software. Normality tests were conducted by using Kolmogorov-Smirnov test and Shapiro-Wilk test. The raw scores for each locomotor and object control skill as well as sum of locomotor, object control skills, and the overall total scores using Smart Start and TGMD-2 were calculated. One participant refused to perform two FMS and two participants refused to perform one FMS. The missed skills were treated as missing values in the data set. Total score of each participant was calculated without the missing values, so the content validity of the subscales and total scales were not compromised.

Concurrent validity

To examine concurrent validity evidence for Smart Start, 9 Pearson correlation coefficients between Smart Start and TGMD-2 was calculated separately by groups and for all participants using the sum of locomotor skill scores, sum of object control skill scores, and the overall total scores. For establishing validity evidence for Smart Start, the first record of Smart Start and TGMD-2 by primary investigator (rater 1) were used for calculation.

Reliability

Using a percentage of agreement, modified kappa coefficient (*km*), and the intraclass correlation coefficient (ICC) with 95% confidence intervals, inter-rater and intra-rater reliability evidence for Smart Start were calculated separately for each group. For modified kappa coefficient (*km*), the following formula was used:

$$km = \frac{\text{observed agreement} - 1/2}{1 - 1/2}$$

In the formula, observed agreement stands for the proportion of agreement and chance is determined as $1/k$ where k equals the number of classification categories (Looney & Gilbert, 2012). For inter-rater reliability of the assessments, the first record of each rater were analyzed. The percentage of agreement and modified kappa coefficient were calculated using the raw scores of each key element of locomotor and object control skills. The ICC (1, 1) for a single measure with 95% confidence intervals were analyzed using the one-way analysis of variance (ANOVA) model. The proportion of agreement and modified kappa describe the agreement of decisions between raters or two coding sessions for one rater while the type of ICC calculated is a type of agreement index where the variance between the two raters or two coding sessions by one rater is considered error. ICC (2, 1) is an agreement coefficient whose value will be equal or greater than the ICC (1, 1) because the variance between the two raters or two coding sessions by one rater is not considered error. If the ICC (1, 1) is an acceptable value, then the ICC (2, 1), which was not computed, would certainly be acceptable (Shrout & Fleiss, 1979). The predetermined alpha level was .05. Separate analyses by group were performed using the sum of locomotor skills, sum of object control skills, and overall total score.

Results

Descriptive Statistics

Table 2 shows the mean for raw scores of Smart Start and TGMD-2 by groups and total participants. The first coding record of rater 1 (primary investigator) was used to calculate the mean of raw scores

Table 2
The Mean of Raw Scores of Smart Start and TGMD-2 for Participants

		Smart Start					
		Locomotor Skills		Object Control Skills		Total score	
Groups		*M*	*SD*	*M*	*SD*	*M*	*SD*
Total	(*N* = 57)	13.3	4.55	13.3	4.09	26.6	7.87
PWD	(*n* = 31)	11.8	4.51	11.5	3.42	22.3	6.92
PWOD	(*n* = 26)	15.0	4.00	15.4	3.83	30.4	7.20
		TGMD-2					
		Locomotor Skills		Object Control Skills		Total Score	
Groups		*M*	*SD*	*M*	*SD*	*M*	*SD*
Total	(*N* = 57)	21.8	6.01	23.7	6.13	45.5	10.6
PWD	(*n* = 31)	20.1	6.74	22.4	5.79	43.5	10.9
PWOD	(*n* = 26)	21.5	5.02	25.3	6.25	47.7	9.91

Note. M = mean; PWD = Preschoolers with a disability; PWOD = Preschoolers without a disability; SD = standard deviation. The max score of the Smart Start is 57 in this study.

Table 3
The Pearson correlation coefficients between Smart Start and TGMD-2

	Groups		
	Total	PWD	PWOD
	(*N* = 57)	(*n* = 31)	(*n* = 26)
Locomotor	.86	.89	.87
Object Control	.87	.89	.90
Total	.89	.92	.92

Note. PWD = Preschoolers with a disability; PWOD = Preschoolers without a disability.

of Smart Start by groups and total participants. In this sample, preschoolers without a disability had higher scores measured by Smart Start than preschoolers with a disability. The same trend was also seen with the TGMD-2 means. For normality tests, the results indicated that there is no evidence for non-parametric data (p = .2 for Kolmogorov-Smirnov test; p = .446 or above for Shapiro-Wilk test).

Concurrent Validity

Pearson correlation coefficients between total scores measured with Smart Start and TGMD-2 for all participants, preschooler with a disability, and preschoolers without a disability were r = .89, p < .01, r = .92, p < .01, and r = .92, p < .01, respectively (Table 3). Pearson correlation coefficients between the sums of locomotor skills measured with Smart Start and TGMD-2 for all participants, preschooler with a disability, and preschoolers without a disability were r = .86, p < .01, r = .89, p < .01, and r = .87, p < .01, respectively. Pearson correlation coefficients between the sums of object control skills measured with Smart Start and TGMD-2 for all participants, preschooler with a disability, and preschoolers without a disability were r = .87, p < .01, r = .89, p < .01, and r = .90, p < .01, respectively. Correlations above 0.75 are considered relatively strong; correlations between 0.45 and 0.75 are moderate, and those below 0.45 are considered weak (Shortell, 2001).

Rater Reliability

For intra-rater reliability evidence for Smart Start and TGMD-2, rater 2 and rater 3 scored 30 participants' videotaped performances (16 with a disability and 14 without a disability) twice and the scores were analyzed. For inter-rater reliability evidence, coefficients between rater 1 and rater 2 and rater 1 and rater 3 were calculated.

For descriptive data analysis, the means of the first coding scores using Smart Start for preschoolers with a disability by rater 1, rater 2, and rater 3 were 22.88 (SD = 7.09), 21.38 (SD = 6.39), and 20.94 (SD = 7.01), respectively. On average, Rater1's scores were higher than Rater 2's scores, F (1, 15) = 12.2, p = .003, Effect

size (ES) = .21. On average, Rater 1's scores were higher than Rater 3's scores, F (1, 15) = 7.2, p = .017, ES = .27. The result of a one-way repeated-measures ANOVA for total scores of Smart Start for preschoolers with a disability indicated on average the second coding scores were higher than first coding scores, F (1, 15) = 10.9, p = .005, ES = .31, within rater 3.

Table 4 shows the averaged results of percentage of agreement, kappa coefficients, and ICCs for intra- and inter-rater of Smart Start for preschoolers with and without a disability. The average percentage of agreement, across three raters, for intra-rater reliability using Smart Start on all participants was 92% (*SD* = 4%, range 76–100%). The mean of modified kappa coefficients, across the three raters, for intra-rater reliability on all participants using Smart Start was .83 (*SD* = .09, range .52–1.00). The mean of ICC's for a single measure of intra-rater reliability using Smart Start on all participants was .96. The average of percentage of agreement, across three raters, for inter-rater reliability using Smart Start on all participants was 86% (*SD* = 5%, range 74–95%). The mean of modified kappa coefficients for inter-rater reliability using Smart Start on all participants was .71 (*SD* = .11, range .42–.89). The mean of ICCs for a single measure of inter-rater reliability using Smart Start on all participants was .93.

The average percentage of agreement, across three raters, for intra-rater reliability using TGMD-2 on all participants was 91% (*SD* = 4%, range 79–100%). The mean of modified kappa coefficients, across the three raters, for intra-rater reliability on all participants using TGMD-2 was .84 (*SD* = .09, range .57–1.00). The mean of ICCs for a single measure of intra-rater reliability using TGMD-2 on all participants was .96. There were no statistically significant differences ($p > .05$) between means relative to respective ICCs, $p > .05$.

The average of percentage of agreement, across three raters, for inter-rater reliability using TGMD-2 on all participants was 87% (*SD* = 5%, range 75–99%). The mean of modified kappa coefficients for inter-rater reliability using TGMD-2 on all participants was .73 (*SD* = .09, range .50–.98). The mean of ICCs for a single measure of inter-rater reliability using TGMD-2 on all participants was .92. There were no statistically significant differences between means relative to respective ICCs, $p > .05$.

Table 4

The averaged results of percentage of agreement, Kappa coefficient, and ICC for the intra-rater and the inter-rater of Smart Start for preschoolers with and without a disability

	Intra-rater Reliability of SS for PWD						Inter-rater of SS for PWD			
	R1 (*n* = 31)		R2 (*n* = 15)		R3 (*n* = 16)		R1/R2 (*n* = 15)		R1/R3 (*n* = 16)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
% of Agreement	90	5	91	4	92	4	87	5	85	4
Kappa Coefficient	.80	.08	.82	.09	.85	.08	.73	.09	.70	.10
ICC	.95		.94		.94		.96		.93	
	Intra-rater Reliability of SS for PWOD						Inter-rater of SS for PWOD			
	R1 (*n* = 26)		R2 (*n* = 13)		R3 (*n* = 13)		R1/R2 (*n* = 13)		R1/R3 (*n* = 13)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
% of Agreement	90	5	90	3	93	3	83	5	86	6
Kappa Coefficient	.81	.10	.79	.09	.87	.07	.66	.10	.72	.12
ICC	.98		.95		.93		.87		.91	

Note. PWD = Preschoolers with a disability; PWOD = Preschoolers without a disability; SS = Smart Start; R1 = Rater 1; R2 = Rater 2; R3 = Rater 3.

Discussion

The purpose of this study was to investigate validity and rater reliability evidence of Smart Start for preschoolers with/without a disability. TGMD-2 is widely used to assess children's FMS, and the assessment tool has been investigated in various research settings (Valentini, 2012) and population (e.g., school-aged children with visual impairment and 7 to 10-year-old children with intellectual disability). Given these

facts, TGMD-2 was utilized to obtain concurrent validity evidence for Smart Start involving preschoolers with and without a disability in this study. Concurrent validity evidence for Smart Start was demonstrated by a strong positive relationship between Smart Start and TGMD-2 for all participants and for each of the preschool groups for total score, object control score and locomotor score.

The results are similar to two other studies that examined validity evidence of a FMS assessment tool using TGMD-2 as a criterion. Sun, Sun, Zhu, Huang, and Hsieh (2011) reported a large correlation between data of TGMD-2 and their assessment tool, Preschooler Gross Motor Quality (PGMQ; $r = .83$, $p < .001$ for total scores) in preschoolers without a disability aged from 3 to 6 years and attending kindergartens in central Taiwan. The PGMQ is a FMS assessment tool for preschoolers aged from 3 to 6 years, which includes 17 items in three subscales of locomotion, object manipulation, and balance. In the study by Williams et al. (2009), movement skills of preschoolers without a disability were examined using the Children's Activity and Movement in Preschool Study (CHAMPS) (Williams et al., 2009) and TGMD-2. CHAMPS consists of 12 items in two subscales of locomotor and object control skills. A large correlation between CHAMPS and TGMD-2 was reported, $r = .98$, $p < .001$ for total scores.

In the present study, the average percentage of agreement for both intra-rater and inter-rater reliability by group were higher than 85%. In general, rule of thumb suggested by other studies contends that values from 75% to 90% demonstrate an acceptable level of agreement when using percentage of absolute agreement (Graham, Milanowski, & Miller, 2012). These results indicated that there is a strong agreement within a rater and between the trained raters (rater 1/rater 2 and rater1/rater 3). Similar results of the percentages of agreement between raters have been reported with a previous version of Smart Start in a sample of 24 preschoolers aged from 36 to 72 months with and without a disability (Ong, 2001). In the study, six raters who were preschool classroom teachers had at least 90% agreement with the researcher for seven locomotor skills.

Modified kappa coefficient and ICC were also calculated in order to examine intra-rater and inter-rater reliability evidence of Smart Start. The modified kappa coefficients in this study were all above .65. There are no universally accepted criteria for reliability coefficients, but according to Altman (1990), a kappa of .61 to .80 represents good agreement. Therefore, there is a strong agreement within a rater and between the two trained raters. ICCs for both intra-rater and inter-rater reliability by group were all above .85. These results support agreement reliability evidence from well-trained raters using Smart Start in preschoolers with/without a disability given Sundvall, Ingersley, Knudsen, and Kirkegaard (2013) view ICC > .80 as almost perfect agreement.

Significant differences between the raters were found when computing one-way repeated-measures ANOVA. Rater 1 scored higher than rater 2 and 3 when the raters used Smart Start for preschoolers with a disability. This trend was also found in the mean scores using Smart Start for preschoolers without a disability. However, the difference in mean raw scores between raters may not be a meaningful difference because there were less than 2 points differences between raters. This could be explained by the fact that rater 1 had more experience and familiarity in both administration of assessment and children motor competence than the other two raters. Experience and familiarity that raters gain from training had an effect on overall total scores. According to Cusick, Vasquez, Knowles, and Wallen (2005), trained raters in the study had significantly higher raw scores compared to untrained raters. It is also important to highlight that inter-rater reliability between rater 1/rater 2 and rater1/rater 3 were acceptable and all raters achieved acceptable intra-rater reliability through the training. The previous studies involving preschoolers have reported higher coefficients than the present study. In the study by Valentini (2012), high Cronbach's alpha coefficient (α) results using TGMD-2 in a sample of 2,674 typically developing children (1,352 boys and 1,322 girls) from 3 to 10 years old ($M = 7.56$ years, $SD = 1.91$ years) have been also reported ($\alpha = .88$ for inter-rater reliability and $\alpha = .92–.99$ for intra-rater reliability). Research by Williams et al. (2009) reported high agreement of inter-rater reliability evidence for CHAMPS using ICC in a sample of preschoolers without a disability ($R = .94$ for total scores).

Although the coefficients calculated in this study were all acceptable, there may be a few reasons for the lower coefficients obtained in this study. First, it may be because the sample sizes of the previous studies were larger than the present study. According to Weiner and Craighead (2010), reliability coefficients are reliant on the number of participants. The number of participants in the research by Sun et al. (2011), research by Williams et al. (2009), and the research by Valentini (2012) were 135, 297, and 2,674, respectively. The second reason is that different types

of reliability coefficients were reported by Valenti (2012) and the present study. The type of ICC calculated in this study is a type of agreement index where the variance between the two raters or two coding occasions by one rater is considered to be error. Cronbach alpha coefficients reported by Valenti do not include the variance between two raters or two occasions as error thereby describing the consistency in the ranks of the participants' scores. As a result, Cronbach alpha coefficients will always be higher than the the type of ICC reported in this study when calculated on the same data. Last reason would be that the training hours in the present study were relatively shorter than the training done for the research by Williams et al. (2009). Williams et al. (2009) reported 51 hours of intensive training using videotapes and observations before coding FMS data using CHAMPS. In the present study, the three raters, who were graduate students with experience in administrating TGMD-2 and in teaching children with and without a disability, completed approximately 20 hours of intensive rater training including introductory, review, and coding practice sessions. The feasibility of fewer training hours while still achieving good coefficients may make training raters to use Smart Start more practical.

There were a few limitations of this study. The first limitation was the small number of participants. Small sample sizes prevented the analysis of the data by gender. However, the means of raw scores for this sample were lower for participants with disability than without disability. The trend seen for the TGMD-2 was also seen for the Smart Start, and it aligned with literature indicating that preschoolers with disability have lower motor competency than their peers without disabilities. This trend seen for the Smart Start can be additional evidence that Smart Start is measuring what it is supposed to be measuring.

Second, running distance was shorter (40 feet for participants with a disability, 30 feet for those without a disability) than TGMD-2 assessment protocol due to limited space of both early childhood centers. Lastly, there was an administrative error of sliding skill test for TGMD-2. The participants were only given one trial. The skill of slide should be completed for both right and left directions; however, the participants in this study performed the skill in right direction only. Despite of this administrative error of the skill, this error may not compromise this study results. All the participants were given only one trial for the skill, and the raters evaluated the skill in the same manners that they were trained.

Perspective Paragraph

The findings of this study support evidence of concurrent validity and intra-rater and inter-rater reliability of Smart Start for assessing FMS for preschoolers with/without a disability using trained raters. Smart Start is a CBA tool for assessing FMS for preschoolers with/without a disability. Although several assessment tools for assessing FMS exist, Smart Start is unique because it is designed for preschoolers with and without disabilities. In addition, this assessment tool can be particularly useful for a classroom teacher who may have limited knowledge of FMS for preschoolers with /without a disability because it provides the teacher with curriculum-based activities to use.

References

Altman, D. G. (1990). *Practical statistics for medical research*. Boca Raton, FL: CRC Press.

Barnett, L. M., van Beurden, E., Morgan, P. J., Brooks, L. O., & Beard, J. R. (2009). Childhood motor skill proficiency as predictor of adolescent physical activity. *Journal of Adolescent Health, 44*, 252–259.

Brewer, C. J., & Jones, R. L. (2002). A five-stage process for establishing contextually valid systematic observation instruments: The case of rugby union. *Sport Psychologist, 16*(2), 138–159.

Burns, M. K., MacQuarrie, L. L., & Campbell, D. T. (1999). The difference between curriculum-based assessment and curriculum-based measurement: A focus on purpose and result. *Communique, 27*(6), 18–19.

Capio, C. M., Sit, C. H. P., & Abernethy, B. (2011). Fundamental movement skills testing in children with cerebral palsy. *Disability & Rehabilitation, 33*, 2519–2528.

Cliff, D. P., Okely, A. O., Smith, L. M., & McKeen, K. (2009). Relationships between fundamental movement skills and objectively measured physical activity in preschool children. *Pediatric Exercise Science, 21*, 436–449.

Cusick, A., Vasquez, M., Knowles, L., & Wallen, M. (2005). Effect of rater training on reliability of Melbourne Assessment of Unilateral Upper Limb Function scores. *Developmental Medicine & Child Neurology, 47*(1), 39–45.

Darst, P. W., Zakrajsek, D., & Mancini, V. H. (1989). *Analyzing physical education and sport instruction*. Champaign IL: Human Kinetics.

Deno, S. L. (1987). Curriculum-based measurement. *Teaching Exceptional Children, 20*, 41.

Evaggelinou, C., Tsigilis, N., & Papa, A. (2002). Construct validity of the Test of Gross Motor Development: A cross-validation approach. *Adapted Physical Activity Quarterly, 19,* 483–495.

Fisher, A., Reilly, J. J., Kelly, L. A., Montgomery, C., Williamson, A., Payton, J. A., & Grant, S. (2005). Fundamental movement skills and habitual physical activity in young children. *Medicine & Science in Sports & Exercise, 37*, 684–688. doi: 10.1249/01.MSS.0000159138.48107.7D

Folio, M. K., & Fewell, R. (2000). *Peabody Developmental Motor Scales: Examiner's Manual* (2nd ed.) Austin, TX: PRO-ED, Inc.

Gallahue, D., & Ozman, J. (2006). *Understanding motor development: Infants, children, adolescents, adults* (6th ed.). New York, NY: McGraw-Hill Humanities.

Gickling, E. E., & Thompson, V. P. (1985). A personal view of curriculum-based assessment. *Exceptional Children, 52*, 205–218.

Graf, C., Koch, B., Kretschmann-Kandel, E., Falkowski, G., Christ, H., Coburger, S., . . . Dordel, S. (2004). Correlation between BMI, leisure habits and motor abilities in childhood (CHILT-Project). *International Journal of Obesity, 28*, 22–26. doi:10.1038/sj.ijo.0802428

Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings*. Online Submission. Retrieved from http://eric.ed.gov/?id=ED532068

Hardy, L. L., King, L., Farrell, L., Macniven, R., & Howlett, S. (2010). Fundamental movement skills among Australian preschool children. *Journal of Science and Medicine in Sport, 13*, 503–508.

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*, 143–159.

IBM Corp. Released (2011). *IBM SPSS Statistics for Windows,* Version 20.0. Armonk, NY: IBM Corp.

Looney, M. A., & Gilbert, J. (2012). Validity of alternative cut-off scores for the back-saver sit and reach test. *Measurement in Physical Education and Exercise Science, 16*, 268–283.

Majnemer A. (1998). Benefits of early intervention for children with developmental disabilities. *Seminars in Pediatric Neurology, 5*, 62–69.

McLean, M. (2005). Using curriculum-based assessment to determine eligibility: Time for a paradigm shift? *Journal of Early Intervention, 28*, 23–27.

Morison, P., White, S. H., & Feuer, M. J. (1996). *The use of IQ tests in special education decision making and planning: Summary of two workshops*. Washington, DC: National Academy Press.

Neisworth, J. T., & Bagnato, S. J. (2004). The mismeasure of young children: The authentic assessment alternative. *Infants & Young Children, 17*, 198–212.

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

Ong, C. D. (2001). *Intracoder and intercoder reliability of the key element scores from the Smart Start locomotor skill key element checklist*. (Unpublished master's thesis). Northern Illinois University, DeKalb, IL.

Partington, M., & Cushion, C. (2013). An investigation of the practice activities and coaching behaviors of professional top-level youth soccer coaches. *Scandinavian Journal of Medicine & Science in Sports, 23*, 374–382. doi: 10.1111/j.1600-0838.2011.01383.x

Potter, M. L., & Wamre, H. M. (1990). Curriculum-based measurement and developmental reading models: Opportunities for cross-validation. *Exceptional Children, 57*, 16–25.

Roberts, S., & Fairclough, S. (2012). A five-stage process for the development and validation of a systematic observation instrument: The system for observing the teaching of games in physical education (SOTG-PE*). European Physical Education Review, 18*, 97–113.

Robinson, L. E., & Goodway, J. D. (2009). Instructional climates in preschool children who are at-risk. Part I: Object-control skill development. *Research Quarterly for Exercise and Sport, 80*, 533–542. doi:10.1080/02701367.2009.10599591.

Cusick, A., Vasquez, M., Knowles, L., & Wallen, M. (2005). Effect of rater training on reliability of Melbourne Assessment of Unilateral Upper Limb Function scores. *Developmental Medicine & Child Neurology*, *47*(1), 39–45.

Graham, M., Milanowski, A., & Miller, J. (2012). Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings. *Online Submission*. Retrieved from http://eric.ed.gov/?id=ED532068

Shevell, M., Ashwal, S., Donley, D., Flint, J., Gingold, M., Hirtz, D., … Sheth, R. D. (2003). Practice parameter: evaluation of the child with global developmental delay: report of the Quality Standards Subcommittee of the American Academy of Neurology and The Practice Committee of the Child Neurology Society. *Neurology, 60*(3), 367–380.

Shortell, T. (2001). *An Introduction to Data Analysis & Presentation.* Retrieved from http://academic. Brooklyn. Cuny. edu/soc/courses/712/chap18. Html.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428.

Simons, J., Daly, D., Theodorou, F., Caron, C., Simons, J., & Andoniadou, E. (2008). Validity and reliability of the TGMD-2 in 7-10-year-old Flemish children with intellectual disability. *Adapted Physical Activity Quarterly, 25*, 71–82.

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*, 290–306. doi:10.1080/00336297.2008.10483582

Sun, S. H., Sun, H. L., Zhu, Y. C., Huang, L. C., & Hsieh, Y. L. (2011). Concurrent validity of preschooler gross motor quality scale with test of gross motor development-2. *Research in Developmental Disabilities, 32*, 1163–1168. doi:10.1016/j.ridd.2011.01.007

Sundvall, L., Ingerslev, H. J., Knudsen, U. B., & Kirkegaard, K. (2013). Inter-and intra-observer variability of time-lapse annotations. *Human Reproduction, 28*, 3215–3221.

Ulrich, D. (2000). *Test of gross motor development. Examiner's manual* (2nd ed.). Austin, Texas: PRO-ED.

Valentini, N. C. (2012). Validity and reliability of the TGMD-2 for Brazilian children. *Journal of Motor Behavior, 44*, 275–280. doi:10.1080/00222895.2012.700967.

Weiner, I. B., & Craighead, W. E. (2010). *The Corsini encyclopedia of psychology (Vol. 4).* Hoboken, NJ: John Wiley & Sons.

Wessel, J., & Zittel, L.L. (1995). *Smart Start: A Preschool Movement Curriculum for Children of All Abilities*. Austin, TX: PRO-ED.

Williams, H. G., Pfeiffer, K. A., Dowda, M., Jeter C., Jones S., & Pater R. R. (2009). A field-based testing protocol for assessing gross motor skills in preschool children: The

CHAMPS motor skills protocol. *Measurement in Physical Education and Exercise Science, 13,* 151–165. doi:10.1080/10913670903048036

Williams, H. G., Pfeiffer, K. A., O'Neill, J. R., Dowda, M., McIver, K. L., Brwon, W. H., & Pate, R. R. (2008). Motor skill performance and physical activity in preschool children. *Obesity, 16,* 1421–1426. doi: 10.1038/oby.2008.214

Wrotniak, B. H., Epstein, L. H., Dorn, J. M., Jones, K. E., & Kondilis, V. A. (2006). The relationship between motor proficiency and physical activity in children. *Pediatrics, 118*, 1758–1764.

Wuang, Y. P., Wang, C. C., Huang, M. H., & Su, C. Y. (2008). Profiles and cognitive predictors of motor functions among early school-age children with mild intellectual disabilities. *Journal of Intellectual Disability Research, 52*, 1048–1060. doi: 10.1111/j.1365-2788.2008.01096.x

Yun, J., & Ulrich, D. (2002). Estimating measurement validity: A Tutorial. *Adapted Physical Activity Quarterly, 19*, 32–47.

Corresponding author

Jaehun Jung
Email address | jungjaeh@oregonstate.edu