academicJournals

Vol. 5(2) pp. 18-28, May 2014
DOI: 10.5897/JPESM12.003
ISSN 1996-0794

http://www.academicjournals.org/JPESM

Journal of Physical Education and Sports Management

Full Length Research Paper

Reliability and validity of test of gross motor development-2 (Ulrich, 2000) among 3-10 aged children of Tehran City

Farrokhi A.[1], Zareh Zadeh M.[2], Karimi Alvar L.[3], Kazemnejad A.[4], and Ilbeigi S.[3]*

[1]Faculty of Physical Education, University of Tehran, Tehran, Iran.
[2]Faculty of Physical Education, University of Kerman, Kerman, Iran.
[3]Faculty of Physical Education, University of Birjand, Birjand, Iran.
[4]Faculty of Physical Education, University of Tarbiyat Modares, Tehran, Iran.

Received 24th January 2012; Accepted 24th January 2013; Published May 2014

The purpose of this study is to determine construct validity and three reliability aspects of Test of Gross Motor Development (TGMD-2; Ulrich, 2000) in Tehranian children aged 3:0-10:11. TGMD-2 which includes two subtests (locomotor and object control) is designed to assess movement pattern development of 12 fundamental movement skills. The TGMD-2 was administered to 1438 children. Internal consistency reliability for locomotor and object control score and also for total composite score averaged 0.78, 0.74 and 0.80, respectively. Internal consistency reliability was high for females and males, too. Test-retest reliability ranged from 0.65 to 0.81, and intra-rater reliability was above 0.95. To investigate construct validity, Ulrich's (2000) two-factor model was postulated and this hypothesis was tested through confirmatory factor analysis. According to the model, six variables or skills measuring child's ability for moving into space were loaded on one factor (locomotor), while the other six variables measuring ability for controlling and manipulating objects were loaded on the other factor (object control). According to the findings, two-factor structure of TGMD-2 and also proper assignment of skills to locomotor and object control factors were supported for our population, too. Additionally, the correlation coefficients between age and subtests' score provided support for another aspect of construct validity, that is, developmental nature of TGMD-2; the resulting coefficients indicated that TGMD-2 is capable of differentiating between ages. In conclusion, this study indicated that TGMD-2 can be used with confidence to assess gross motor development of the studied population.

Key words: TGMD-2, construct validity, internal consistency, test-retest, intra-rater reliability.

INTRODUCTION

The principal element in motor development is the fundamental movement skills (Reeves et al., 1999), which are included in the gross movements, that is, the movements which are related to the large force producing muscles

*Corresponding author. E-mail: saeed.ilb@gmail.com. Tel: +985612502124. Fax: +985612502124.

function (Thomas and Nelson, 1996). These skills which refer to the purposeful movement patterns include locomotor and object control movements. The locomotor movement patterns include the skills used for the purpose of transporting the body through space, like running and jumping. The object control movement patterns, however, are the skills that involve giving force to object or receiving force from objects, like throwing and catching the ball. These skills develop, in early childhood, as a function of physical maturation and practice and are the prerequisites for games and specialized movements (Gallahue and Ozmun, 2006).

The mastery of certain fundamental movement skills is a prerequisite for daily life functioning and participation in later physical or sport-specific activities. At an early age, gross movement skills are necessary to move, stabilize and control body and objects while exploring the environment. Later in life, well developed gross movement skills help individuals to function more smoothly (Cools et al., 2008). Studies have indicated that inadequate competence in these skill causes lack of success and the feeling of being incapable in games (Vira and Ruadsepp, 2003) which leads to the gradual elimination of physical activity through the life span (Deflandre et al., 2001) and the improper level of physical fitness (Reeves et al., 1999). In addition, children who are weak in these skills cannot easily establish social relations with others and have a higher level of anxiety and lower levels of self-esteem in comparison to more skillful peers (Piek et al., 2006). Also, delay in gross motor development usually is followed by visual perception disorders (Skordilis et al., 2004), dyslexia and linguistic disorder which, if not diagnosed on time, lead to learning and cognitive problems (Merriman et al., 1995). Thus, one aspect of the screening programs in early childhood should be measuring the development of fundamental movements which calls for the existence of standardized tests in this area. These tests can be divided into two categories, each with advantages and disadvantages, the measurement of motor performance quantitatively by speed, distance or number of successful attempts and the measurement of fundamental movement skills qualitatively by criteria for passing or technique components of the movement. In recent years, the most frequently used fundamental movement skills assessment tools with children employ qualitative measures, because the major advantages of qualitative assessment are the information can be used to inform the teacher which specific components of a skill an individual needs to practice, and the assessment can be undertaken in a more meaningful context than quantitative methods (Hands, 2002). In addition, technique components are not fully acquired by preschool, and their performance is easily influenced by testing conditions in quantitative methods. So it is not always valid to use quantitative scales with preschool children. In tests based on pass or fail criteria, however, motor ability is estimated by a specific performance criterion, movement patterns can be observed in daily life and special measurement tools and conditions are not needed (Goshi et al., 1999).

One of the best known tests in movement assessment area is Test of Gross Motor Development-2 (Ulrich, 2000) designed to measure the development of fundamental movements based on qualitative aspects of the skills in children of 3 to 11, and includes six locomotion skills and six object control skills. According to the author, the test can be used to identify children who are significantly behind their peers in gross motor performance, to plan programs to improve skills in those children showing delays and to assess changes as a function of increasing age, experience, instruction or intervention (Cools et al., 2008).

One of the qualities of a standardized test is validity. A valid test is one that theoretical and empirical evidences confirm the usefulness and appropriateness of interpretations and applications based on the test results (Van Waelvelde et al., 2004). One aspect of validity is construct validity. Construct is the characteristic reflected in the test scores (Burton and Miller, 1998), like intelligence or motor development. Construct validity refers to the extent to which the underlying construct (factors) or trait, claimed to be measured by the test, can be identified (Thomas and Nelson, 1996) and also refers to the extent to which test scores reflect the construct-related theoretical concept on which the test is based (Saif, 2004). To examine the construct validity, generally hypotheses are generated about the defined constructs assumed to be latent in test performance, and the hypotheses are verified by logical or empirical methods. To identify the number of constructs and the structural model on which the test is based, confirmatory and exploratory factor analyses are used (Ulrich, 2000). In the exploratory factor analyses, using the covariance analyses in the matrix of correlation between the test's measures, the measures that their variations are correlated are grouped as one factor. So, the underlying constructs are extracted and the factorial structure of the test is identified (Sarmad et al., 2004). But, the identified factorial structure is specific to the population the test is designed for. Therefore, the appropriation of the test's factorial structure in a new population is evaluated through confirmatory factor analysis and goodness of fit indexes (Thomas and Nelson, 1996).

The Test of Gross Motor Development was originally developed in the United States for typically developing (TD) children but it has been translated and validated in different countries for children with and without disability. Evidence of the construct validity of the TGMD-2 was reported in its test manual. TGMD-2 was validated on 1,208 American children via exploratory factor analysis using principal component analysis with promax rotation and confirmatory factor analysis with maximum likelihood mode of estimation. Although the exploratory factor

analysis results identified two factors of locomotor skills and object control skills, the loadings of two items (Strike and Jump) were unclear. The fit indexes provided by confirmatory factor analysis indicated that the two-factor model of the TGMD-2 produced a good approximation to data; however, the model did not represent a reasonable fit ($\chi 2 = 280.3$, df = 53, $\chi 2/df = 5.29$). In addition, the level of significance and some fit indexes such as root mean square error of approximation (RMSEA), standardized root mean square residual (SRMR), non-normed fit index (NNFI) and comparative-fit index (CFI), as well as, the path diagram with item loadings and correlation between two factors, provided a better understanding toward the underlying structure of a testing instrument, was not mentioned in the manual (Spessato et al., 2012). Wong (2006) tested 1251 TD Hong Kong Chinese children. A principal axis factor analysis with oblique rotation determined two factors, but five items were loaded on both factors. After eliminating these items progressively and computing additional principal axis factor analyses, the desirable exploratory factor analysis results were two factors with eight items loaded satisfactory on factors suggested by Ulrich (2000). Four two-factor models with 12-, 11-, 9-, and 8-item were tested using confirmatory factor analysis. Although the two-factor model of the TGMD-2 did not represent exact fit, goodness-of-fit indexes supported that the four models were tenable (Wong, 2006). For Brazilian Children Appropriate indices of the confirmatory factorial validity and adequate items correlations with the correspondent subtest were found, that provided reasonable support for the TGMD-2 two factor model (Valentini, 2012).

TGMD-2 has also been used to test children with sensory disability and those cognitively impaired. Satisfactory fit indexes were reported for Flemish children with cognitive disabilities (goodness-of-fit index [GFI] = .88; adjusted goodness-of-fit index [AGFI] = .82) (30) and visual impairment children from Netherlands (RMSEA = 0.07, GFI = 0.85) (Houwen et al., 2010).

Extensive research addressed the influence of age and gender on the proficiency of gross motor skills. It was reported that gross motor skills improved with age (Wong, 2006). So, In addition to factorial structure analysis, the other aspect of construct validity of the tests having construct with developmental nature, such as TGMD-2, is the investigation of age differentiation capability or the score improvement by age increasing (Saif, 2004).Thus, it could be hypothesized that there is a high correlation between the children's test performance and age. One way of establishing a test's construct validity is to study the performances of different groups of individuals on the test (Ulrich, 2000). Apart from age related differences, gender differences in gross motor skill performance have been established in both children and adolescents, with boys achieving higher scores than girls especially in object control skills (Okely and Booth, 2004; Spessato et al., 2012). Then with reference to the literature, it would be expected that boys would obtain higher object control skill scores. Ulrich (2000) indicated that both subtests were strongly related to the chronological age ($r=0.69$ to 0.72, $p<.05$ for locomotor; $r = 0.71$ to 0.75, $p<.05$ for object control). Ulrich did not report gender differences in the manual but the reported raw score means showed boys did better than the girls on the object control subtest (Ulrich, 2000). Niemeijer (2007) also, observed high correlations between age and subtests total score, ranged .66 to .81, among typically development Dutch children (Niemeijer, 2007). Simons et al. (2007) reported significant but low correlation between the age and object control subtest scores ($r = 0.25$) for children with mental retardation; ANOVA also revealed a significant age effect in the object control subtest but not in the locomotor subtest. Significant effect of gender was also observed in object control skills among mental retarded children (Simons et al., 2007). For children with visual impairment, there were significant partial correlations between age and the locomotor ($r = 0.36$, $p = 0.002$) and object control ($r = 0.53$, $p<.001$) subtests. A significant effect for gender in the object control subtest was found, but not in the locomotor subtest (Hands, 2002).

Reliability is the other feature of standardized tests and the prerequisite for validity, that is, an assessment instrument that is not reliable cannot be valid (Saif, 2004). Reliability pertains to repeatability and consistency of test scores. Theoretically, an observed score (X) is compromised of an examinee's true score (T) and error score (E), that is, $X = T + E$. Reliability, then, is the degree to which observed test scores match true score. With regard to psychometric definition, reliability is presented in terms of the variance components (S_T^2 / S_X^2). As the error variance decreases, greater proportion of the observed-score variance accounted for by the true-score variance, and reliability will increase (Burton and Miller, 1998). Thus, reliability reflects the degree to which test scores are free of measurement error or error variance (Thomas and Nelson, 1996).

Measurement error comes from three sources: content and questions, test administration conditions and scoring. To determine the content reliability of a test, internal consistency is examined. Internal consistency refers to the extent of test's item consistency and congruity (Okely and Booth, 2004). The error of administration is estimated through test-retest method. Particularly when the performance is expected not to change during a short time, the stability of performance is investigated during a given time period by test-retest (Saif, 2004). In this method, the test is administered at two different times but with the same conditions and the correlation between two set of scores is reported as "stability reliability" (Wong, 2006). The inter-rater and intra-rater reliability are two ways of

investigating the scoring reliability or objectivity. In this method, which requires video-recording, the performance is scored by two raters or by one rater at two different times, then the correlation between the scorings is calculated (Burton and Miller, 1998).

The degree of reliability is expressed by a correlation coefficient, ranging from 0.00 to 1.00, and generally coefficient of above .70 is good .The correlation coefficient is calculated through intra-class correlation (R) and Pearson correlation (r). Some rationale was presented for using ICC methods instead of Pearson r, such as ICC uses ANOVA to obtain reliability coefficient and changes in means and standard deviations from trial to trial or item to item can be assessed in the ICC. So it provides precise estimation of error variance. The best-known technique of estimating reliability based one ICC is Cronbach alpha coefficient (Morrow et al., 2005).

Satisfactory reliability results have been reported for locomotor, object control and composite score of TGMD-2 for different populations. For American children, internal consistency alpha coefficient was higher than .85 (locomotor α = .85, object control α = .88, and GMQ α = .91). Pearson correlation coefficient for the test retest was .88 for locomotor, .93 for object control and for total test .96. Inter rater reliability was .98 for all three aspects (Ulrich, 2000). Desirable test-retest values (r = .83 to .91) and inter- and intra-rater reliability (α = 0.86 to 0.99) were found for TD Brazilian children (Valentini, 2012). The internal consistency of the TGMD-2 was found to be high (α = 0.71 to 0.72) and the inter rater, intra rater, and test retest reliability were acceptable (ICCs ranging from 0.82 to 0.95) for Netherlands children with visual impairment of Netherlands (Houwen et al., 2010). Wong and Cheung (2006) also reported acceptable indices of internal consistency (α =.78 to .85) and Inter rater reliability (r = .82 to .94) for TD Chinese (Wong, 2006). Alpha values ranged from .82 to .90 for subtests and composite score internal consistency among Flemish children with cognitive disabilities. In addition, Spearman correlation coefficients for test retest reliability of locomotor and object control subtests were .90 for and .92, respectively (Simons et al., 2007).

Although there are evidences of reliability and validity of TGMD-2, the evidence of a measure's reliability and validity is sample-dependent and cannot be generalized to other cultural and geographical areas (Yun and Ulrich, 2002). This fact limits the applicability of TGMD-2 about Iranian children. Considering the sociocultural differences observed across different countries, application of a test in a population with different set of characteristics should be followed by the assessment of reliability and validity of the test to guarantee the correctness of the result (Chow et al., 2006). In Iran, there was no test available to measure the qualitative process of gross motor skills and the introduction of the TGMD-2 to kindergartens and primary schools seemed to be desirable. Due to this, and with regard to the unpleasant consequences of delayed motor development and the necessity of reliable and valid test existence for motor development assessment, the present study aims at determining the construct validity and three aspects of the reliability of TGMD-2: internal consistency, stability and objectivity among Tehranian children of 3 to 11.

METHODOLOGY

Sample selection

The population was 3 to 11 years old children of public kindergarten and elementary schools, supervised by Education Organization of Tehran City. The sample was selected through a stratified cluster sampling procedure. Based on information from the Education Organization, the sample was stratified by geographic region and education districts size as well as age and gender. Finally, the sampling procedure resulted in 1438 persons as target population for normalization and validation. All the children recruited for testing were healthy and free from diagnosed orthopedic, neurologic, physical impairments, developmental conditions and learning disability.

63 children were selected from the sample for the evaluation of test-retest reliability. Intra-rater reliability was evaluated by testing 32 children which were selected just for this purpose.

Instrument

Test of Gross Motor development-2 (Ulrich, 2000) was the instrument used for collecting the data. This test is a revised vision of TGMD (Ulrich, 1985). Ulrich standardized and validated this test using a sample of 1208 American children. TGMD-2 is a process oriented test which assesses development of movement pattern of 12 fundamental movement skills. The age range from 3 to 10 years old covers the period in which the most dramatic changes in a child's gross movement skill development occur. The test includes two subtests: object control and locomotor. Locomotor subtest includes running, galloping, hopping, jumping, leaping and sliding while object control subtest includes striking, dribbling, catching, kicking, throwing and rolling.

Each skill is evaluated based on some performance criteria. The content of each criterion is about one component of the advanced pattern of skill performance. Each subtest includes 24 performance criteria. For scoring, the child has to perform the task at two trials. Each criterion is given a 1, if the criterion is performed correctly; and a 0, if performed incorrectly. The test is administered in 15 to 20 min. Each subtest has a total raw score attained by summing the scores for the related skills. The maximum score is 48. The normative scores include the standardized score for the subtests, with mean and standard deviation of 10 and 3 respectively and also the Gross Motor Quotient which is a composite score based on the standardized scores (Ulrich, 2000). GMQ which has a mean and a standard deviation of 100 and 15 respectively is the best measure of a child's overall gross motor ability (Burton and Rodgerson, 2001).

Data collection procedure

The study was approved by research and ethical committee of Tehran Organization of Education, and the investigator got the

permission to perform the TGMD-2 in selected school. The school members were informed about the purpose and safety measures. After admittance of the school, students' date of birth, parents' education and health status were obtained from personal files and the information was considered in sampling procedure. Prior to data collection, the children's parents gave informed consent for their children's participation.

To assure measurement consistency, the same tester (investigator) observed and scored all children's performance. Prior to testing, training was done on administrating and scoring of the TGMD-2 based on the manual instructions. Before the actual data collection, a pilot-testing was conducted in order to evaluate the investigator's scoring competency and intra-rater reliability. For estimating intra-rater reliability, the investigator watched and scored the videotaped performance of selected participants twice, with a 12-day interval. The performance was video recorded by using Sony R, CCD-TRV418E, 1322101 Camcorder. The TGMD-2 was administered twice, with a 2-week interval between testing, for test-retest reliability. The testing was done during the physical education class. Arrangements were made to accommodate the test in a safe environment for the children in order to minimize administration time and distractions. The children were tested in small groups and each child finished the test in one session. The testing followed standardized test procedures as provided in the test manual (Ulrich, 2000).

Data analysis

The reliability of TGMD-2 was investigated through internal consistency, test-retest and intra-rater reliability methods. The scores of the entire normative sample served as subjects for internal consistency analysis and the data for evaluating stability and objectivity reliability were obtained from the sample selected for these two purposes. Alpha coefficient was computed for internal consistency on the subtest and GMQ using SPSS. But, internal consistency for GMQ was calculated using the following formula which is designed for composite scores (20).

$$\alpha = 1 - \frac{\sum \delta_{xi}^2 (1 - \rho_{xixi})}{\delta_z^2}$$

δ_{xi}^2 = subtests score's variance, ρ_{xixi} = subtest's reliability coefficient, δ_z^2 = composite score's variance.

Internal consistency coefficient was computed for age groups separately and in order to estimate total internal consistency reliability of subtests and GMQ regardless of age, the alphas were averaged using z-transformation method, the formula for calculating the mean of some correlation coefficients (Thomas and Nelson, 1996). Intrarater and test-retest reliability was evaluated using the intraclass correlation coefficient (ICC). To control any effects of age in the selected samples on the evaluation of test-retest and intra-rater reliability, the total score of each subtest was changed into standardized scores.

The construct validity of TGMD-2 was investigated through confirmatory factor analysis and examining age differentiation and gender differences. Maximum-likelihood confirmatory factor analyses were performed to test the goodness-of-fit of TGMD-2 skill assignment to the subtests, using Lisrerl (Lisrerl 8.8, Scientific Software International, Lincolnwood, IL, USA). It was assumed that the locomotor subtest includes the six items of running, galloping, hopping, jumping, leaping and sliding and object control subtest includes the six items of striking, dribbling, catching, kicking, throwing and rolling. Pearson correlation coefficients were calculated to examine the relationship between age and subtests total score. A 4×2 (Age × Gender) two-way analysis of variance was performed to examine further age related developmental changes and also gender differences on subtest performance. For the analysis, age bands were defined as age band 1, from 3-4 years; age band 2, from 5-6 years; age band 3, from 7-8 years; age band 4, from 9-10 years.

Table 1. Alpha coefficients for TGMD-2 scores at age intervals and gender.

Age (N)	Subtests		GMQ
	Loc	OC	
3 (159)	.88	.72	.86
4 (178)	.87	.69	.83
5 (190)	.83	.69	.80
6 (190)	.77	.75	.81
7 (184)	.75	.76	.77
8 (177)	.66	.77	.79
9 (187)	.74	.78	.84
10 (173)	.65	.71	.72
Avg. α	.78	.74	.80
Girl (719)	.92	.85	.91
Boy (719)	.91	.89	.92

RESULTS

TGMD-2 reliability

Table 1 shows the internal consistency reliability coefficients for the locomotor subtest (Loc), object control subtest (OC) and GMQ, and also shows the averaged coefficients. At the end of the table alpha for the two groups of boys and girls has been provided.

As shown in Table 1, the range of the internal consistency reliability coefficients for the eight age groups is from .65 to .88. The averaged alpha coefficients show that the total reliability is above .74. In addition, the alpha for each of the genders is above .85. Although the acceptable value of reliability coefficient is above .70 (4), the minimum acceptable level for alpha has been reported to be .65 (1). Thus, the resulted alpha values confirm internal consistency reliability of TGMD-2 for Tehranian children of 3 to 11.

Table 2 shows the ICCs and corresponding 95% Confidence intervals for, intrarater and test-retest reliability. In this table, the mean and standard deviation of the subtests' total score for each testing and scorings are provided. The content of Table 2 shows a little difference

Table 2. The test-retest and intra-rater reliability coefficient of TGMD-2.

TGMD-2 scores	Test-retest			Intra-rater		
	TM (SD)	TM (SD)	ICC(95٪CI)	TM (SD)	TM (SD)	ICC(95٪CI)
Loc	32 (3)	33 (3)	.65(.50-.79)	32 (3)	31(3)	.95(.91-.97)
OC	20 (6)	21 (5)	.85(.75-.91)	24 (5)	24 (4)	.99(.97-.99)
GMQ	100 (15)	100 (15)	.81(.70-.89)	100 (15)	100 (15)	.97(.94-.98)

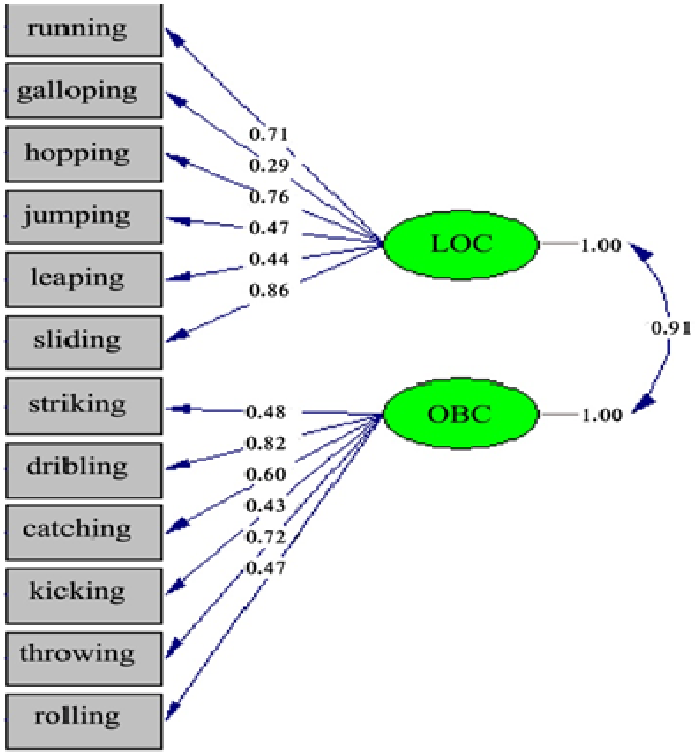

Figure 1. Pictorial profile of hypothesized two-factor model of TGMD-2.

between the means for each testing and scorings. for interpretation of the resulted coefficients we adopted the criteria recommended by Fleiss (1981) that consider ICCs of > 0.74 as excellent, < 0.40 as poor, 0.40-0.59 as fair and 0.60-0.74 as good. Based on the criteria, the coefficients for test retest and intrarater reliability were at good to excellent level (Fleiss, 1981).

Validity of TGMD-2

Validity was evaluated using confirmatory factor analysis. Figure 1 shows the schematic representation of the two-factor structure of the test and the internal relations of the components.

According to Figure 1, the loading of 12 skills on two factors (locomotion and object control) range from .29 to .86. The least value belongs to galloping and the highest value belongs to sliding. As the loading values below .30 are considered low (11), the resulted values shown in Figure 1 are desirable. All factors loadings were statistically significant ($p < .05$). The figure shows that the correlation between the two factors is high which indicates that a one factor solution would be appropriate for the data rather than a two factor solution. Thus, to investigate the correctness of the assumed factor structure, goodness of fit indexes are taken into account which is shown in Table 3.

The overall fit of the data to the evaluation of the proposed factor structure is initially based on the non-significance of chi-square ($\chi 2$) (Evaggelinou et al., 2002). A non-significant p value for $\chi 2$ means that the model is a good representation of the data and there is no reliable difference between the sample matrix (that is, the data)

Table 3. The values of goodness of fit indexes for the two-factor structure of TGMD-2.

Model	Fit Indexes								
	χ^2	df	χ^2/df	GFI	AGFI	RMSEA (90% CI**)	NNFI	CFI	SRMR
Two-factor model	303.9*	53	5.70	.97	.95	.05 (.05-.06)	.97	.98	.03

*p<.05 **Confidence Interval.

Table 4. Subtests' total score means (and standard deviations) and age differentiation correlation coefficients.

Age	Loc		OC	
	Girl M(SD)	Boy M(SD)	Girl M(SD)	Boy M(SD)
3	16(6.5)	17(6.5)	12(3)	13.5(4)
4	24.5(6)	25.5(6)	15(3.5)	17(4)
5	28.5(5)	29(5)	17(4)	19(4)
6	31.5(4)	31(3.5)	18.5(4)	22(5)
7	32(3.5)	33(3.5)	21(4.5)	26.5(5)
8	33.5(3)	33(3)	24(5)	28(5)
9	34(3)	34(4)	25(5)	30(5.5)
10	34(3.5)	34.5(2.5)	27(4)	32.5(4)
Correlation with age (r)	.68*	.66*	.76*	.79*

*p<.05.

and the fitted matrix (that is, the model) (Houwen et al., 2010). As shown in Table 3, this index is significant here. However, χ2 is strongly affected by sample size and in large samples there is likelihood that it would be significant (Evaggelinou et al., 2002). Thus, more valid indexes are used besides χ2 (Cools et al., 2008). The values of 2 to 5 for χ2/df show the reasonable fit of the assumed factorial model (Ulrich, 2000). The goodness-of-fit index (GFI) and adjusted goodness-of-fit index (AGFI) were examined to provide information about the proposed model estimate covariance and sample covariance (Valentini, 2012). The GFI was considered to be a measure of the proportion of variances and covariances that the proposed model was able to explain (Wong, 2006). Both GFI and AGFI indexes range from 0 to 1 and values greater than .90 are indicative of a good model (Evaggelinou et al., 2002). In terms of assessing the degree of lack of fit of the model, the root mean square error of measurement (RMSEA) was computed. For the examined model, the values between .05 and .08 indicate a reasonable approximation to the data (Wong, 2006). RMSEA values less than .05 would indicate a close fit of the model while values of about .08 would indicate a reasonable fit (Evaggelinou et al., 2002). The comparative-fit index (CFI) was used to estimate model discrepancies. The CFI compared the proposed model with the null model assuming no relationships between measures. The non-normed fit index (NNFI) was computed to examine the proportionate improvement in fit of the model compared to a baseline model in which all observed variables are uncorrelated. The values of NNFI and CFI in the .90 or above were considered as a reasonably good approximation of the data. The standardized root mean square residual (SRMR) was computed for the proposed model to provide a summary measure of standardized residuals. The small values of the SRMR for the model (<.05) demonstrated an acceptable fit (Valentini, 2012).

As shown in Table 3, although χ2 index is significant, the fit indexes are quite acceptable with respect to the criteria reported above. GFI and AGFI considered as absolute fit indexes (Evaggelinou et al., 2002) are close to 1, reflecting reasonably good approximations of the data. RMSEA value, another absolute fit index, demonstrates that the two factor structure of TGMD-2 was close fitted. The values of other indexes included as descriptive-fit and alternative-fit indexes provide additional support for the fit of the two-factor model.

Table 4 shows the correlations between age and the subtests' scores of TGMD-2. The means and the standard deviations for each subtest are presented. Because of the developmental nature of motor development, it was expected that there is a high and significant correlation between age and the total score of the two subtests.

Table 4 shows that the mean of the performance of the sample in the subtests improves as the age goes up. The obtained coefficients also show the TGMD-2 subtests are strongly related to age.

Factorial ANOVA revealed a significant main effect for gender in the object control subtest, $F(1, 1430) = 228.06$, $p < 0.001$, but not in the locomotor subtest, $F(1, 1430) = 2.79$, $p = 0.09$, with the boys doing significantly higher than the girls on the object control subtest. The results showed significant influence of age for locomotor subtest, $F(3, 1430) = 465.27$, $p < 0.001$, and object control subtest, $F(3, 1430) = 648.52$, $p < 0.001$. Post hoc analysis of influence of age using Tukey HSD test revealed that significant differences ($p < 0.001$) between the four age bands (3.0 to 4.0, 5.0 to 6.0, 7.0 to 8.0, and 9.0 to 10.0), with the elder bands outperformed younger bands on the performance of both subtest.

DISCUSSION

TGMD-2 reliability: In investigation of the internal consistency reliability of TGMD-2, the alpha coefficient values for the two subtests of locomotion, object control and the composite score of GMQ are .78, .74 and .80 respectively. As the values of higher than .70 are interpreted as good and those above .80 are interpreted as very good (Armitage and Theodore, 1998), the findings are indicative of the fact that the internal consistency reliability of TGMD-2 is acceptable for 3 to 11 years old children of Tehran City. It reflects the homogeneity and consistency of items' content among Tehranian children.

Comparing the results with reported coefficients for American children (Ulrich, 2000), we found that the internal consistency reliability of the two subtests and GMQ was higher (.85, .88 and .91 respectively) among American. One of the factors affecting reliability is the group variability. Considering the concept of reliability in the classical test theory which views reliability as the proportion of true score's variance to observed score's variance, the more heterogeneous the group the higher the reliability score would be, because of the broad range of scores (Kubiszyn and Borich, 1990). When the standard deviation of the subtests' score for American children were compared with those of the participants in the present study, it was found that, in most of the age groups, the standard deviation for the scores of the American children were higher (Ulrich, 2000). It would be attributed to cultural and physical education content differences of two societies.

The estimation of the test-retest or stability reliability showed that the ICC coefficients for the object control and locomotion subtests and also GMQ were .65, .85 and.81 respectively. With regard to the criteria of Fleiss (1981) the resulted coefficients show the acceptable stability of TGMD-2 scores over time. The results indicate that the scores of object control subtest are more stable than the scores of locomotion subtest. Although the comparison of the results to other studies is difficult due to the differences in the statistical analysis methods applied, Ulrich (2000) and Simons et al. (2007) also observed this difference among TD American children and mental retarded Flemish children, respectively (Simons et al., 2007;Ulrich, 2000). To justify this difference, we can refer to study of Lee et al. (2005). The researchers investigated the psychometric properties of TGMD-2 according to item response theory concepts. They found that the object control items were more difficult and more precise than those of locomotion, regardless of the group properties (Lee et al., 2005). The more precise the items on a test, it assesses a more extensive range of the ability it is assumed to assess, thus the test has a higher reliability (Baumgartner and Jackson, 1991). So, it can be said that the higher reliability of object control subtest relative to locomotion subtest is due to the more precise items.

The result showed that the high intra-rater reliability of TGMD-2. In all articles which were reviewed on TGMD reliability, the scoring or objectivity reliability of the test was reported as higher than .80. This is indicative of the clear and unambiguous administration and scoring of the test.

TGMD-2 validity: Although validity evidence for the TGMD-2 has been reported for typically developing children (Ulrich, 2000; Valentini, 2012; Van Waelvelde et al., 2004) and children with intellectual disabilities (Simons et al., 2007), research investigating the validity of TGMD-2 in different countries was insufficient. More importantly, the TGMD-2 had not been validated in Iran. Then we aimed to assess some aspect of the validity of the TGMD-2 for measuring fundamental movement skills in Tehranian children. The two-factor solution reported by Ulrich (2000) that consisted of 12 test items for representing locomotor skills and object control skills was tested with confirmatory factor analysis and the results compared to the those reported for American (Ulrich, 2000), Hong Kong Chinese (Valentini, 2012) and Brazilian (Van Waelvelde et al., 2004) children. The two-factor model of the TGMD-2 in this study did not result in exact fit as illustrated by significant chi-square value for the reason of sample size. The chi-square to degrees of freedom ratios, however, did not exceed 5 demonstrated that the model was acceptable. As displayed in Figure 1 each motor skill correlated satisfactory with the correspondent subtest and negative correlations were not observed between motor skills and subtests that supported the two factor model of TGMD-2 (Cronbach, 1989). In addition, the correlation between locomotion and object control was .91. This high correlation shows that these two factors both assess the gross motor skills. But, in some case where correlation coefficients show the

Table 5. Reported factor loading for two subtests of TGMD-2 in three validation studies.

	Ulrich (2000) (*N* = 1,208)		Wong and Cheung (2010) (*N* = 626)		Valentini (2012) (*N* = 2,674)	
	Subtests					
Skill	**Loc**	**OC**	**Loc**	**OC**	**Loc**	**OC**
Run	.52		.64		.46	
Gallop	.66		.68		.71	
Hop	.70		.71		.66	
Leap	.49		.56		.53	
Jump	.59		.57		.53	
Slide	.69		.78		.55	
Strike		.75		.63		.69
Dribble		.61		.66		.56
Catch		.57		.64		.59
Kick		.65		.50		.75
Throw		.75		.66		.69
Roll		.67		.72		.45

Table 6. Reported goodness of fit indexes of factorial structure of TGMD-2.

	Ulrich (2000)	Wong and Cheung (2010)	Valentini (2012)
χ2/df	5.29	3.40	*
GFI	.96	.95	.98
AGFI	.95	*	.95
RMSEA	*	.06	.06
CFI	*	.97	.88

*Not reported.

relation between the subtests of a battery, it is expected that the coefficients would be acceptably high; too high coefficient means all the subtests measure the same ability not different aspects of a construct (Ulrich, 2000). Thus, the observed correlation may be indicative of the fact that the two subtests are not independent and the two-factor structure of TGMD-2 does not fit to the data. But the two factor structure of TGMD-2 provided a sufficient fit to the data as presented by small RMSEA and SRMR, and high GFI, AGFI, NNFI and CFI, based on suggested values. Therefore the two-factor structure of the TGMD-2 seems to be tenable to Tehranian children. In this study the resulted loading of the factors and fit indices were similar to those reported for typically developing children from the United States (Ulrich, 2000), Hong Kong (Wong and Cheung, 2010) and Brazil (Valentini, 2012), as indicated in Tables 5 and 6. It was also observed relatively high correlation (.71) between the two latent constructs (locomotor and object control) for Hong Kong Chinese children (Wong, 2006). However, confirmatory factor analysis of one-factor solution with 12 items did not show an acceptable fit (χ2/df =7.86) compared to two-factor solution (χ2/df =3.40), and the two-factor model represented a significantly better fit than the one-factor. This further confirmed the two-factor structure of the TGMD-2.

By comparing reported χ2/df in Table 6 and the value obtained in this study it could be concluded that the factorial structure of TGMD-2 provided better fit to data for Chinese population than American and Iranian population.

The literature shows that as typically developing children grow older, their motor skill performance improves due to physical maturation and experience, and older children demonstrate higher mastery level of gross motor skill performance than younger children (13,14). So it could be expected that a valid motor development test should be capable to detect developmental changes in performance. The resulted correlations between the subtests total score and age were significant and ranged from .66 to .79, which are in the high to very high range, according to MacEachron's (1982) rule of thumb interpretations (Ulrich, 2000). The results of analysis of variance, also, provided further support for developmental nature of the test, with both gender's performance improving with age. Such results suggest that TGMD-2 has sufficient validity to assess age-related development of gross motor skills in Tehranian children. The observed correlations range indicates that correlation between age and locomotor subtest's scores was smaller than those for object control. Such a thing was observed among American (Ulrich, 2000) and Dutch children (Niemeijer, 2007). TGMD psychometric properties examining showed that the object control items have higher degrees of difficulty than the locomotor items (Cole et al., 1991; Lee et al., 2005). From a developmental view, one expects children of 3 to 10 years of age to perform better on locomotor skills than ball skills, as shown in Table 4. So the better performance of children in locomotor subtest is indicative of the fact that their scores are close to maximum. Thus, due to the ceiling effect, locomotor score progresses with age, especially in higher age groups, more slowly than object control score. It shows that age differentiation capability of object control items is better than locomotor items.

Our results indicated sex related differences in object control subtest, with boys outperforming girls. These results were in line with previous studies conducted in other countries (Okely and Booth, 2004; Spessato et al., 2012; Wong, 2006), and in special population such as children with autism (Berkeley et al., 2001), mental retardation (Simons et al., 2007) and visual impairment

(Houwen et al., 2007). According to the motor development literature the proficiency of gross motor skills differ between males and females because of biological differences and sociocultural factors. Then, based on the results, TGMD is a valid test for distinguishing between the two groups.

Conclusion

Based on the findings it was concluded that TGMD-2 is a valid and reliable instrument for measuring gross motor development of Tehranian children. Although the fit indexes of two-factor structure were seen reasonable as a result of confirmatory factor analysis, the correlation between two factors, that is, locomotor and object control was high; this refutes the assumed underlying traits of the test. So it would be necessary to examine fit indexes of single-factor model and compare them to two-factor model's fit indexes. According to the Yun and Ulrich (2002), the validation process should not be limited to one approach, so it is also recommended to examine content and criterion validity in future studies (Yun and Ulrich, 2002). It should be noted that some object control items like striking are not widely used in Iranian traditional games and sports, this issues should be considered in future examining internal consistency and construct validity especially through exploratory factor analysis. In addition, because the sample was limited to Tehranian children caution is needed in generalization of the results to other Iranian population and in the interpretation of data. Therefore, a comprehensive validation study is suggested through selection of a representative sample that is stratified by age, gender and geographical region of Iran.

Conflict of Interests

The author(s) have not declared any conflict of interests.

REFERENCES

Armitage P, Theodore C (1998). Encyclopedia of Biostatistics. John Wiley & Sons Ltd.

Baumgartner TA, Jackson AS (1991). Measurement for evaluation in physical education and exercise science. 4th ed. Wm. C. Brown publishers.

Berkeley SL, Zittel LL, Pitneey LV, Nichols SE (2001) Locomotor and object control skills of children diagnosed with autism. Adapted Physical Activity Q. 18(4):405-416.

Burton AW, Miller DE (1998). Movement skill assessment. Human Kinetics.

Burton AW, Rodgerson RW (2001). New perspectives on the assessment of movement and motor abilities. Adapted Physical Activity Q. 18:374-365.

Chow SMK, Hsu Y, Henderson SE, Barnett AL, Lo SK (2006). The movement ABC: A cross-cultural comparison of preschool children from Hong Kong, Taiwan, and the USA. Adapted Physical Activity Q. 23:31-48.

Cole E, Wood TM, Dunn JM (1991). Item response theory: A useful test theory for adapted physical education. Adapted Physical Activity Q. 8:317-332.

Cools W, De Martelaer K, Samaey C, Andries C (2008). Movement skill assessment of typically developing preschool children: A review of seven movement skill assessment tools. J. Sports Sci. Med. 8:154-168

Cronbach LJ (1989). Construct Validity after thirty years. In RL Linn (Ed.), Intelligence: Measurement, theory, and public policy. Bloomington, IL: University of Illinois pp.147-171

Deflandre A, Lorant J, Gavarry O, Falgairette G (2001). Determinants of physical activity and physical and sports activities in French school children. Perceptual Motor Skills 92:399-411.

Evaggelinou C, Tsigilis N, Papa A (2002). Construct validity of the test of gross motor development; A cross-validation approach. Adapted Physical Activity Q. 19:483-495.

Fleiss JL (1981). Statistical methods for rates and proportions, 2nd Edition. New York: Wiley

Gallahue DL, Ozmun JC (2006). Understanding motor development: Infant, children, adolescent, adult. 6th ed. McGraw-Hill International Edition.

Goshi F, Demura S, Kasuga K, Sato S, Minami M (1999). Selection of effective tests of motor ability in preschool children based on pass or fail criteria: Examination of reliability, objectivity, and rate of passing. Perceptual Motor Skills 88:169-181.

Hands BP (2002). How can we best measure fundamental movement skills? Health Sciences Conference Papers. http://researchonline.nd.edu.au/health_conference/5/

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 Q. 27:143-159.

Houwen S, Visscher C, Hartman E, Lemink KAPM (2007). Gross motor skills and sport participation of children with visual impairments. Res. Q. Exercise Sport. 78(1):16-23.

Kubiszyn T, Borich G (1990). Educational testing and measurement. 3th ed. Harper Collins Publisher.

Lee M, Zhu W, & Ulrich DA (2005). Many-Faceted Rasch calibration of TGMD-2. Internet Article.

Linn LL (editor) (1988). Educational measurement. 3th ed. American Council on Education & National Council on Measurement in Education. Part I: Reliability.

Merriman W, Barnet BE, Isenberg D (1995). A preliminary investigation of the relationship between language and gross motor skills in preschool children. Perceptual and Motor Skills 81:1211-1216.

Morrow JR, Jackson Aw, Disch JG, Mood DP (2005). Measurement and evaluation in human performance. 3th ed. Human Kinetics.

Niemeijer AS (2007). Neuromotor task training for children with developmental coordination disorder. Unpublished doctoral dissertation, University of Groningen, the Netherlands p.16.

Okely AD, Booth ML (2004). Mastery of fundamental movement skills among children in New South Wales: Prevalence and sociodemographic distribution. J. Sci. Med. Sport. 7:358-372.

Piek JP, Baynam GB, Barrett NC (2006). The relationship between fine and gross motor ability, self-perceptions and self-worth in children and adolescents. Hum. Movement Sci. 25:65-75.

Reeves L, Broeder CE, Kennedy-Honeycutt L, East C (1999). Relationship of fitness and gross motor skills for five to six yr. old children. Perceptual Motor Skills 89:739-477.

Saif AA (2004). Educational measurement, assessment and evaluation. 3th ed, Dowran publishing company (Persian publisher).

Sarmad Z, Bazargan A, Hejazi E (2004). Research method in behavioral sciences. Aagah publishing company (Persian publisher).

Simons J, Daly D, Theodorou F, Caron C, Simons J, Andoniadou E (2007). Validity and reliability of TGMD-2 in 7-10 yr. Flemish children with intellectual disability. Adapted Physical Activity Q. 25(1):71-82

Skordilis EK, Douka A, Spartali I, Koutsouki, D. (2004). Depth perception of elementary school students with qualitatively evidenced locomotor impairments. Perceptual and Motor Skills 99:501-518.

Spessato BC, Gabbard C, Valentini N, Rudisill M (2012). Gender

differences in Brazilian children's fundamental movement skill performance. Early Child Dev. Care pp.1-8.
Thomas JR, Nelson JK (1996). Research methods in physical activity. 3rd ed. Human Kinetics.
Ulrich DA (2000).Test of Gross Motor Development. 2nd ed. Austin, TX: Pro-Ed.
Valentini N (2012). Validity and Reliability of the TGMD-2 for Brazilian Children. J. Motor Behav. 44(4):275-280
Van Waelvelde H, De Weerdt W, De Cock P, Smits-Engelsman BCM. (2004). Aspects of the validity of movement assessment battery for children. Hum. Movement Sci. 23:49-60.
Vira R, Ruadsepp L (2003). Psychological correlates of physical activity among seven through eight grades. Hum. Movement Stud. 44:501-517.
Wong AKY (2006). Construct Validity of the Test of Gross Motor Development–2. Unpublished doctoral dissertation, Hong Kong Baptist University.
Wong AKY, Cheung SY (2010). Confirmatory factor analysis of the Test of Gross Motor Development-2. Meas. Phys. Educ. Exercise Sci. 14:202–209.
Yun J, Ulrich DA (2002). Estimating measurement validity: A tutorial. Adapted Physical Activity Q. 19:32-47.