www.revistageintec.net
ISSN: 2237-0722
revista GEINTEC Gestão, Inovação e Tecnologias

Reliability, Validity, and Norm References of Standing Broad Jump

Zarizi Ab Rahman[1]; Azlan Ahmad Kamal[2]; Mohad Anizu Mohd Noor[3]; Soh Kim Geok[4]; Alnedral[5]

[1]Faculty of Education, Universiti Teknologi MARA, Selangor Branch, Selangor, Malaysia.

[2]Faculty of Education, Universiti Teknologi MARA, Selangor Branch, Selangor, Malaysia.

[3]Faculty of Sport Science and Recreation, Universiti Teknologi MARA, Selangor, Malaysia.

[3]mohadanizu@uitm.edu.my

[4]Faculty of Educational Studies, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia.

[5]Coaching Department, Faculty of Sport Science, Universitas Negeri Padang, Indonesia.

Abstract

Standing Broad Jump (SBJ) is a field test used to assess leg power. This study aims to determine the reliability, validity, and develop norm reference among adolescents. The evidence of the reliability, validity, norm need to establish in the particular population to support the interpretation of the scores. This study involved 60 subjects and six raters for reliability and validity. 417 subjects for norm development. The ICC, test-retest, and Pearson Correlation were used to determine the reliability. Independent sample t was used to determine the validity. The standard deviation method was used to construct the norm reference. Findings showed the ICC was high among male raters (.96) and female raters (.99). The consistency of the instrument excellent among males and female subjects (r = 96, r = .90). Independent sample t-test showed t value (58) = 3.395, p = 0.01 was significant. Finding showed a significant difference between the elite (M= 2.0871) and non-elite athlete (M= 1.897) for male. While, there are significant different for female subjects' t value (58) = 7.324, p = 0.00 was significant. The difference showed the SBJ has the construct validity evidence in this population. This study also indicated new norm for SBJ as (M = superior 2.54 above, F = 2.06 above, excellence, M= 2.26-2.53, F= 1.74-2.05, good, M= 2.97-2.25, F= 1.40-1.73, average, M= 1.69-1.96, F = 1.06-1.39, poor, M= 1.68 and below, F= 1.05 and below). The results suggest that SBJ are reliable and validity with the norm reference to assess leg power. This study also will enhance the quality of Physical Education teacher either local or abroad. Hence, quality teachers should produce pupils with the balance of intellectual, spiritual, emotional, and physical. This study also provides new direction for others researcher to conduct new study especially in term of methodology and population.

Key-words: Objectivity, Reliability, Construct Validity, Norm Reference, Elite, Non-elite, Known Group Method.

1. Introduction

Muscular power is the ability to generate maximum force in the fastest possible time (Miller, 2014). The importance of muscular power is well established in human sports performance (Taipale, Mikkola, Vesterinen, Nummela, & Hakkinen, 2013; Ronnestad, Kojedal, Losnegard, Kvamme, & Raastad, 2012). Muscular power also essential for health outcomes among youth and associated with bone health by increasing bone mass (Reid & Fielding, 2012; Ginty, Rennie, Mills, Stear, Jones, & Prentice, 2005) to protect from osteoporosis and other bone diseases. Daily life activities such as walking, climbing stairs, or standing from a seated position require muscular leg power. Hence, muscular power is very significant not only for the athletes but also on normal population. It also requires good muscular power, especially when involving in the occupation, which requires a lot of walking, climbing stairs, and more.

Standing Broad Jump (SBJ) or Standing Long Jump (SLJ) is a field test used to assess explosive leg power or the ability to apply force in a horizontal direction (Madruga-Parera, Bishop, Fort-Vanmeerhaeghe, Beltran-Valls, Skok, & Romero-Rodriguez, 2020; Stauffer, Nagle, Goss, & Robertson, 2010). Although lab tests such as the Wingate test cycle ergometer provide accurate measurement, it still lacks feasibility. Therefore, field tests can be used to estimate muscular power. The SBJ test used a simple protocol, time-efficient protocol, and does not require complicated equipment (Chung, Chow, & Chung, 2013; Burr, Jamnik, Baker, Macpherson, Gledhill, & McGuire, 2008). Furthermore, the test has also been proposed by AAHPER Youth Fitness to assess leg power (Morrow, Mood, Disch, & Kang 2015). Standing Broad Jump has been used in the Physical Fitness Test as one of the instruments to assess muscular power in the selection process among candidates of the Physical Education program in the Faculty of Education, Universiti Teknologi MARA (UiTM). During the four years of taking Physical Education program, students not only involved in the classroom learning process, but they need to complete practical activities outside of the classroom, such as outdoor education, sports, games, athletic, and other curriculum activities. Upon graduation, they will become a Physical Education teacher at the school. Therefore, it is essential for the candidates who take the Physical Education course to have a high level of physical fitness, including lower body muscular power, to effectively perform their daily tasks as students of Physical Education. Data collection of lower body muscular power is critical to ensure the selection process fulfils the Physical Education program's criteria.

Valid and reliable tests of motor competence such as SBJ are necessary to allow researchers or practitioners to identified motor competencies, skill deficiencies and determine the effectiveness of

motor skill performance (Hulteen, Barnett, True, Lander, del Pozo Cruz, & Lonsdale, 2020). Yelboga and Tavsancil (2010) explained one of the measurement errors, according to Classical Test Theory, is inconsistency in time (test-retest or stability). A Classical Test Theory coefficient known as the Pearson correlation coefficient (r) was calculated for each construct's score using test and retest. Pearson's *r* is a value between -1.0 and 1.0, indicating the linear relationship between two measurements (Streiner, Norman, & Cairney, 2015) to ensure the test instruments' stability. Generalizability (G) theory is a statistical theory to assess and refine measurement procedures' designs to ensure the reliability of measurements (Morrow & Safrit, 1989; Cronbach, Gleser, Nanda, & Rajaratnam, 1972). G Theory considers the number of observers or raters encoding behaviours in the measurements process. Medina and Noguera (1999) suggested it is essential to include more than one rater to ensure data consistency are not affected by the subjective judgment if one rater is involved. However, two or more raters may lead to differences of judgment or evaluation between raters. So that is very important to ensure the test should have interobserver or interrater reliability. Hence, interrater reliability will indicate the extent to which the data has remained stable throughout the process of measurement. Furthermore, Cronbach and Meehl (1955) explained various methods to investigate the validity, one of the methods was the known-groups method. The method referring to the test score should discriminate between groups by using information from means (Thomas & Nelson 2001). For instance, construct validity for the anaerobic power test can determine if the sprinters score significantly better than the distance runners. Hattie and Cooksey (1984) also stressed if a test is "valid," one criterion could be that test scores must discriminate across groups that are theoretically known to differ.

Baumgartner, Jackson, Mahar, and Rowe (2007) and Miller (2014) asserted a test that yields scores that allow valid interpretations for six-year-old children but might not yield scores for 15 years old. Therefore, we can assume that a test yields valid information only for individuals similar in age, gender, and experience to those on whom the original validation was conducted. Additionally, Baumgartner & Jackson (1998) also asserted that norms reference, which is more than five years, cannot describe the actual findings and can be regarded as out-dated and needs to be re develop. Although a few validation studies have been conducted for SBJ, the difference in population in term of age, gender, experience require a new study of the relevant population (Yin, Tang, & Tao, 2018; Reid, Dolan, & De Beliso, 2017; Fernandez-Santos, Ruiz, Cohen, Gonzalez-Montesinos, & Castro-Pinero, 2015). Besides, the process of validation and reliability is about obtaining and gathering the evidence to support the score interpretation. Hence, gathering validity and reliability is never fully established because of the differences in population. Therefore, the reliability, validity evidence, and

norm reference for lower body muscular power in this population need to establish to ensure that the data obtained are meaningful and the selection process is accurate as required. Even though Thomas, Petrigna, Tabacchi, Teixeira, Pajaujiene, Sturm, Sahin, Lopez, Pausic, Paoli, Alesi, & Bianco (2020) have come out with norm for SBJ, but the differences of age and population causes the validity and reliability can be questionable. Therefore, the purposes of this study are to establish reliability, validity evidence and develop a specific norm-referenced standard for lower body muscular power assessed by SBJ. It also can be used in the selection process of candidates for Physical Education programs in the Faculty of Education, UiTM.

2. Literature Review

Standing Broad Jump is still used in Malaysia's education system to assess students' leg power in school, college, and universities, especially in Physical Education and sport science classes. Ministry of Education in Malaysia also uses SBJ as one of the instruments to assess leg power in the Physical Education program's selection process in the Institute of Teacher Education in Malaysia (MOE,2019). Wakai & Linthorne (2005) divided SBJ performance into three parts: (a) the take-off distance, which is defined as the horizontal distance between the take-off line and the jumper's centre of mass at the instant of take-off, (b) the flight distance, which is the horizontal distance travelled by the centre of mass while airborne and (c) the landing distance, which is defined as the distance between the centre of mass and the heels of the feet at the instant of landing. Reliability is one of the critical characteristics of a good test. A reliable measure is consistently unchanged over a short period of time (Baumgartner, Jackson, Mahar & Rowe, 2007). For example, if an individual whose power ability has not changed is measured twice, the two scores will be identical or consistent within two days consecutively. Reliability is vital, and for a measurement to have validity, it must be reliable (Baumgartner, Jackson, Mahar & Rowe, 2007). Scholtes, Terwee, & Poolman (2011) and Khoo & Li (2016) stated that the type of reliability included using different sets of items from the same measurement instrument (internal-consistency), across time (test-retest), by different persons (variation between two or more raters) and on the same occasion (intra-rater). Baumgartner, Jackson, Mahar, and Rowe (2007) and Miller (2014) also stated lack of agreement among scorers, lack of consistent performance by the individual tested, failure of an instrument to measure consistently are the factors of measurement error. The higher the error in any assessment information, the less reliable it is, and the less likely it is to be useful. Hence, the lower the measurement error, the higher the reliability, and thus, the measurement instrument is said to be of good quality. A few methods

estimate reliability, such as test-retest, internal consistency, intraclass correlation, parallel forms method, and split-half method (Hashim, 2015; Miller, 2014; Baumgartner, Jackson, Mahar & Rowe, 2007).

The new definition of validity is related to validity evidence (Baumgartner, Jackson, Mahar & Rowe, 2007). Validity evidence refers to empirical evidence that supports the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment (Messick, 1989b). While Cronbach (1971) asserted that what needs to be valid is the meaning or interpretation of the scores. Baumgartner, Jackson, Mahar, and Rowe (2007) and Cronbach (1971) also stated that we do not validate the test but collecting evidence to validate the interpretations made from the test score. Validity evidence must be collected to support the interpretation of the scores, either logically or statistically. A few methods of the construct validity evidence of physical activity can be investigated. Mahar & Rowe (2002) stated some parts of validity theory in psychology and educational measurement does not seem to fit the different types of research in exercise science. Therefore, a strong method of construct validation that will fit a wide variety of constructs and contexts, especially those relating to the study of physical activity, needs to be established.

Baumgartner, Jackson, Mahar & Rowe (2007) suggested construct validity evidence for exercise science and Physical Education can be determined based on judgment by expertise in a related area of the variables, comparison of the performance of the group before and after instruction or training, and statistical procedure namely factor analysis to identify constructs and the test that yield score leading to valid interpretation. Miller (2013) also reported comparing the mean difference for elite and non-elite performers as one of the procedures to determine the construct validity evidence. That method is also known as known different group validity, referring to a test that discriminates between two groups known to differ on the variable interest (Davidson, 2014; Mahar & Rowe, 2002). This type of evidence is similar to the "known groups" method, originating by Cronbach and Meehl (1955). Mahar and Rowe (2002) asserted construct validity evidence exists if two or more populations differ on a construct. This should be reflected in significant mean differences on a measure of that construct. Group differences are determined using a parametric or nonparametric statistic that allows group comparisons such as a t-test or analysis of variance (ANOVA), with post-hoc analysis (McConnell, Kolopack, & Davis, 2001).

The previous researcher in exercise science and Physical Education has used several methods in determining the reliability and construct validity evidence of SBJ. Reid, Dolan, & De Beliso (2017) used interclass and intra class reliability coefficients (ICC) to determine the test battery's

reliability. According to the researchers, it was found that SBJ is a highly reliable (interclass r = .99, intraclass r = 0.99) field test to assess leg power for collegiate track and field athletes. While Almuzaini & Fleck (2008) also used the intra class correlation coefficient (ICC) to determined reliability of the SBJ test and found high reliability for SBJ (ICC r = .97). However, a study by Ayan-Perez, Cancela-Carral, Lago-Ballesteros, & Martinez-Lemos, (2017) was used test-retest and Pearson correlation coefficients to examine the reliability of vertical jump for children. Castro-Pinero, Ortega, Artero, Girela-Rejon, Mora, Sjostrom, M., & Ruiz (2010) also added nor fatigue effects were found. The SBJ test is reliable for assessing lower body muscular strength for both male and female adolescents.

Evaluation is the process of giving meaning to measurement by judging it against some standards. The two most widely used in Physical Education, namely Norm Reference Grade and Criteria Reference Grade (Hashim, 2015; Miller, 2014). A criterion reference standard is used to determine if someone has attained a specific level, while a norm reference standard judges an individual's performance about other members' performance (Baumgartner, Jackson, Mahar & Rowe, 2017). A criterion reference standard is useful for setting the performance standard for all. Whereas a norm-referenced standard is valuable for comparisons among individuals when the situation requires a degree of selectivity. The standard deviation and percentile method have been widely used to compare student achievement in Physical Education when selecting the subject are needed. Previous studies on the construction of physical fitness norms have been used by the percentile method (Hashim & Gunathevan, 2015; Saint-Maurice, Laurson, Kaj, & Csanyi, 2015; Sharma, 2014; Godara, 2014). Percentile methods are best used when not involving large sample sizes. However, some researchers have used standard deviation and mean methods to generate physical fitness norms, such as Sookhanaphibarn & Choensawat (2015), setting the norm for adolescent physical fitness components in Thailand and Kanniyan (2016), setting the norm for skills in football games. The mean and standard deviation methods are best used when there is a large sample size and required the normally distributed scores (Miller, 2014). According to Chan (2014), grading by standard deviation has more advantages and is more flexible. The standard deviation method seemed to produce better cut-offs in allocating an appropriate grade to students more according to their differential achievements. Norm-referenced tests and their resulting scores provide data that help educators determine the performance level for a specific domain and compare student's achievement to other similar students.

3. Methods

Reliability and validity evidence for the SBJ was determined by ICC method, test-retest, and comparative design suggested by Miller (2014) and Baumgartner, Jackson, Mahar, and Rowe (2007). A total of sixty subjects (30 males, 30 females) and six assistant raters were involved in this study consisting of three males and three females from the Physical Education Program at the Faculty of Education. Therefore, the ICC method was used to obtain inter-rater reliability. Each of the raters is given intensive training on the administration to enhance the objectivity of the raters. The raters will provide the same protocol before they administer the test. Each rater will measure SBJ to all the subjects separately and independently on the same day. The time interval will be given to the raters after each test is completed. The ICC method was used to obtain an agreement between raters. While test-retest with 24 hours' time interval and Pearson Correlation was used to determine the instruments' reliability. The shorters of time interval are considering because no fatigue effects were found for the SBJ (Artero, Girela-Rejon, Mora, Sjostrom, & Ruiz, 2010) and enhancing the reliability of the instrument (Bishop, 2008).

On the other hand, evidence of construct validity was obtained by comparing mean scores between 30 elite athletes and 30 non-elite male and female athletes. The elite players consider representing the campus for an inter-campus tournament, while non-elite players represent the faculty for an inter-faculty tournament in handball. The handball players are selected because the handball players required explosive leg power for throwing the ball with power and speed, which are met through jumping and physical contact with the opponent (Akilan, & Chittibabu, 2014). The elite athlete is known to be better of ability compared to the non-elite. Hence, the instruments have construct validity evidence whenever the elite athlete's mean score is superior to the non-elite athlete (Miller, 2014; Baumgartner, Jackson, Mahar, & Rowe, 2007; Thomas & Nelson 2001). An independent t-test was conducted to determine the significant difference between the two groups. Norm reference-grade with the standard deviation method was used to establish the norm in this study. The total selected subject is 417 (Male = 207, Female = 210). According to Morrow, Jackson, Disch, and Mood (2005), the sample size needed for the norm development should be at least 200 people for each variable. All selected subjects for norm development were adolescents age 19-22 male and females who were candidates that undergo fitness test for admission to the Physical Education program in Faculty of Education, UiTM for 2019 and 2020 intakes. Permission from the parents or guardians and the declaration of health status were received before the test administration. The SBJ test obtained data in this study. The test was performed on a hard surface, and participants

were required to jump as far as possible (horizontal direction) following standardized procedures. Participants started from a standing position, with both feet touching a starting line, and were allowed to swing their arms before the jump. Both the take-off and landing phases of the jump had to be done with both feet. The distance between take-off and the heel of the closest foot at landing was recorded in centimetres. Each participant completed two trials, and the best score was recorded and used for analysis.

4. Results and Discussion

Data were analysed using SPSS for Windows ver.26.0. Descriptive statistical methods were used to obtain the mean and standard deviation. While ICC for inter rater reliability of objectivity, test-retest, Pearson correlation for reliability evidence, and independent t-test used to establish construct validity evidence, and standard deviation method for norm development. Table 1 showed a high degree of reliability was found between male and female inter raters for SBJ measurements. The average measure ICC was .968 with a 95% confidence interval from .941 to .984 (F (29,58) = 30.639, p<.001). Findings also showed a high degree of reliability was found between female inter raters for SBJ measurements. The average measure ICC was .996 with a 95% confidence interval from .993 to .998 (F (29,58) = 269.457, p<.001).

Table 1 - ICC Between Raters for SBJ Measurement

			95% Confidence Interval		F Test with True Value 0			
		Intraclass Correlation[b]	Lower Bound	Upper Bound	Value	df1	df2	Sig
Male	Average Measures	0.968	0.941	0.984	30.639	29	58	.000
Female		0.996[c]	0.993	0.998	269.457	29	58	.000

Table 2 showed the reliability evidence for SBJ male and female subjects. Findings indicate the SBJ instrument has high reliability both in male and females subject (r = .96 & .90) after test-retest.

Table 2 - Reliability Evidence for SBJ Instruments

		Male		***Female***	
		Test	Retest	Test	Retest
Test	Pearson Correlation	1	0.965**	1	0.900**
	Sig. (2-tailed)		.000		.000
	N	30	30	30	30
Retest	Pearson Correlation	.965**	1	0.900**	1
	Sig. (2-tailed)	.000		.000	
	N	30	30	30	30
**. Correlation is significant at the 0.01 level (2-tailed).					

Table 3 showed the descriptive result SBJ test for elite athlete and non-elite athlete. Findings showed mean for male elite athlete (M= 2.0871, SD= 0.236) better than non-elite athlete (M= 1.897, SD = 0.192) for male. Findings also showed mean for female elite athlete (M= 1.597, SD = 0.147) also better than non-elite athlete (M= 1.337, SD = 0.125).

Table 3 - Descriptive Statistics for SBJ Elite and Non-Elite

Instrument	Gender		N	Mean	Std. Deviation	Std. Error Mean
SBJ	Male	Elite	30	2.087	0.236	0.042
		Non-elite	30	1.897	0.192	0.035
	Female	Elite	30	1.597	0.147	0.026
		Non-elite	30	1.337	0.125	0.023

An independent sample t-test was used to determine construct validity evidence based on the comparison between elite athletes and non-elite athletes. Tables 4 and 5 showed the findings of independent sample t-test showed t value (58) = 3.395, p = 0.01 was significant. Findings showed there were significant different between elite (M= 2.0871, SD .2365) and non-elite athlete (M= 1.897, SD= .1925). This study also showed significant differences among female subjects. Independent sample t-test showed t value (58) = 7.324, p = 0.00 was significant. Findings showed there were significant different between elite (M= 1.5971, SD .1473) and non-elite athlete (M= 1.33, SD= 1252). The difference showed the SBJ has the construct validity evidence in this population.

Table 4: Independent Sample T Test for Construct Validity Evidence for Male Subjects

Levene's Test for Equality of Variances						t-test for Equality of Means				
						Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
		F	Sig.	t	df				Lower	Upper
SBJ	Equal variances assumed	0.693	0.408	3.395	58	0.001	0.18986	0.05592	0.07792	0.30179

Table 5 - Independent Sample T Test for Construct Validity Evidence for Female Subjects

Levene's Test for Equality of Variances						t-test for Equality of Means				
						Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
		F	Sig.	t	df				Lower	Upper
SBJ	Equal variances assumed	0.030	0.862	7.324	58	0.000	0.25951	0.03543	0.16514	0.35388

Table 6 showed the descriptive statistical analysis of the SBJ test showed that the mean score and standard deviation for the entire male subjects were (M = 2.11, SD = 0.287) while for female subjects (M= 1.56, SD = 0.339). The standard deviation method assumes the data are normally distributed. Hence, skewness and kurtosis were used to determine the normality. Hair, Black, Babin, & Anderson (2010) and Bryne (2010) asserted that data is considered normal if skewness is between -2 to +2. The values for kurtosis between -2 and +2 are considered acceptable to prove normal distribution (George & Mallery, 2010). Table 10 showed that the distribution of data is considered normal. Therefore, the standard deviation method can be used for norm development in this study.

Table 6 - Descriptive Statistics SBJ Males and Females

Test	**N**	**Min**	**Max**	**Mean**	**SD**	**Skewness**	**Kurtosis**
SBJ (M)	207	1.44	2.81	2.1148	0.28744	0.207	0.133
SBJ (F)	210	1.09	2.54	1.5600	0.33996	1.08	0.477

Table 7 showed the arrangement for the standard deviation method suggested by Miller (2014) for five grades. Based on the findings, table 7 showed five categories: superior, excellence, good, average, and poor for adolescents aged 17 until 22, male and female.

Table 7 - SBJ Norms for male and female subjects age17-22

Performance	Male	Female
	Score (cm)	Score (cm)
Superior	2.54 and above	2.06 and above
Excellence	2.26 - 2.53	1.74 – 2.05
Good	1.97 – 2.25	1.40 – 1.73
Average	1.69 – 1.96	1.06 – 1.39
Poor	1.68 and below	1.05 and below

These findings illustrate the ICC for male raters (.96) and female raters (.99) was excellent. Portney & Watkins (2000) suggested that ICC values less than 0.5 indicate poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability. Values greater than 0.90 indicate excellent reliability. Baumgartner, Jackson, Mahar, & Rowe (2007) stated that the inter scorer objectivity coefficient should be at least .80. Hence, this study showed that different raters could easily administer the SBJ protocols without the raters having different scores. The instrument's consistency was determined by test-retest, and Pearson Correlation showed r = .96 for male subjects, and r = .90 for female subjects was very high. Miller (2014) suggested the correlation coefficient for reliability between ±.80 to 1.00 (very high), ±.60 to .79 (high), ±.40 to .59 (moderate), ±.20 to .39 (low), below .20 (extremely low). The instrument's high reliability due to the close time interval (24 hours) between the test and retest. Bishop (2008) suggested the shorter the time interval, the higher the reliability of the instrument. The ~~l~~ongest time intervals may cause some physical changes and affect the measurement process. The independent sample t-test showed there are significant differences for leg power between elite and non-elite players. This finding indicates that the SBJ test can discriminate the subjects' abilities in terms of leg power. One of the characteristics of a good test should discriminate students' abilities (Jacob & Rothstein, 2016). The validity of the instruments will determine whether it can be measured the construct and yield valid interpretation. The known difference group evidence can determine to construct validity evidence if two or more populations differ on a construct (Mahar & Rowe, 2002). The analysis of the norm development for SBJ illustrates Standing Broad Jump score by gender. The grade provided in this study allows for comparisons of leg power with other populations in the same categories. For example, average standing broad jump scores obtained in this sample shows not much of a difference to what was reported in Saint-Maurice, Laurson, Kaj, & Csanyi~~,~~ (2015) and Sharma (2014). The SBJ norms also allow the proper selection of the candidates for Physical Education programs because the objectivity, reliability, and construct validity process in this population was determined. So, the grade should provide meaningful interpretation in the population.

5. Conclusion

The main conclusion that can be drawn is that the precise SBJ test battery in terms of reliability, construct validity, and the latest norm will enhance the success of the evaluation and interpretation for all candidates to be selected as Physical Education teacher candidates in Malaysia as well as abroad countries. The proper selection process will ensure the candidates are healthy and fit to fulfil the requirements of the Physical Education program throughout the study period. Only quality teacher candidates can equip themselves with all the skills required throughout the study. Furthermore, these findings also will improve the Physical Education teacher to serve better in school and contribute significantly to all the pupils to achieve the aims of education worldwide. Moreover, quality teachers should prepare pupils with the balance of intellectual, spiritual, emotional, and physical, especially in volatility, uncertainty, complexity, and ambiguity (VUCA) due to the rapidly changing and hyper-connected world. Every teacher needs to equip with all the necessary skills to face the VUCA world to provide high quality of teaching and learning in school. This study also provides a new direction for other researchers to conduct future research, especially with different methodology and population.

References

Akilan, N., & Chittibabu, B. (2014). Comparison of leg explosive power between volleyball and handball players. *Paripex Indian Journal Research*, *3*, 55-56.

Almuzaini, K. S., & Fleck, S. J. (2008). Modification of the standing long jump test enhances ability to predict anaerobic performance. *The Journal of Strength & Conditioning Research*, *22*(4), 1265-1272.

Ayan-Perez, C., Cancela-Carral, J. M., Lago-Ballesteros, J., & Martínez-Lemos, I. (2017). Reliability of sargent jump test in 4-to 5-year-old children. *Perceptual and Motor Skills*, 124(1), 39-57.

Baumgartner, T. A., & Jackson, A. S. (1998). *Measurement for Evaluation in Physical Education and Exercise Science* (Ed. 6). WCB/McGraw-Hill.

Bishop, P.A. (2008). *Measurement and Evaluation in Physical Activity Applications*. Arizona: Holcomb Hathaway.

Bulten, R., King-Dowling, S., & Cairney, J. (2019). Assessing the validity of standing long jump to predict muscle power in children with and without motor delays. *Paediatric Exercise Science*, 31(4), 432-437.

Burr, J. F., Jamnik, R. K., Baker, J., Macpherson, A., Gledhill, N., & McGuire, E. J. (2008). Relationship of physical fitness test results and hockey playing potential in elite-level ice hockey players. *The Journal of Strength & Conditioning Research*, 22(5), 1535-1543.

Byrne, B. M. (2010*). Structural Equation Modelling with AMOS: Basic Concepts, Applications, And Programming (Multivariate Applications Series).* New York: Taylor & Francis Group.

McConnell, S., Kolopack, P., & Davis, A. M. (2001). The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC): a review of its utility and measurement properties. *Arthritis Care & Research: Official Journal of the American College of Rheumatology, 45*(5), 453-461.

Medina, J., & Noguera, M. Á. D. (1999). Observer training methodology for research on Physical Education and Sport in which observation is used as a method. *European Journal of Human Movement*, (5), 69-86.

Castro-Pinero, J., Ortega, F. B., Artero, E. G., Girela-Rejón, M. J., Mora, J., Sjostrom, M., & Ruiz, J. R. (2010). Assessing muscular strength in youth: usefulness of standing long jump as a general index of muscular fitness. *The Journal of Strength & Conditioning Research, 24*(7), 1810-1817.

Chan, W. S. (2014). A better norm-referenced grading using the standard deviation criterion. *Teaching and Learning in Medicine, 26*(4), 364-365.

Chung, L. M. Y., Chow, L. P. Y., & Chung, J. W. Y. (2013). Normative reference of standing long jump indicates gender difference in lower muscular strength of pubertal growth. *Health 5*(6A), 6-11.

Cramer, G. D., Cantu, J. A., Pocius, J. D., Cambron, J. A., & McKinnis, R. A. (2010). Reliability of zygapophysial joint space measurements made from magnetic resonance imaging scans of acute low back pain subjects: comparison of 2 statistical methods. *Journal of Manipulative and Physiological Therapeutics,* 33(3), 220-225.

Cronbach, L. J. (1971). *Test validation. Educational measurement. In R.L. Thorndike, Educational Measurement (2nd ed).* Washington, DC: American Council on Education.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements*. New York: Wiley.

Davidson, M. (2014). Known-groups validity. *Encyclopedia of Quality of Life and Well-Being Research*, 3481-3482.

Davies, B. N. (1990). The relationship of lean limb volume to performance in the handgrip and standing long jump tests in boys and girls, aged 11.6–13.2 years. *European Journal of Applied Physiology and Occupational Physiology, 60*(2), 139-143.

Fernandez-Santos, J. R., Ruiz, J. R., Cohen, D. D., Gonzalez-Montesinos, J. L., & Castro-Pinero, J. (2015). Reliability and validity of tests to assess lower-body muscular power in children. *The Journal of Strength & Conditioning Research, 29*(8), 2277-2285.

Frey, B. B. (Ed.). (2018). *The Sage Encyclopaedia of Educational Research, Measurement, and Evaluation*. Sage Publications.

George, D. (2011*). SPSS For Windows Step by Step: A Simple Study Guide and Reference, 17.0 Update,* 10/e. Pearson Education India.

Ginty, F., Rennie, K. L., Mills, L., Stear, S., Jones, S., & Prentice, A. (2005). Positive, site- specific associations between bone mineral status, fitness, and time spent at high impact activities in 16- to 18-year-old boys. *Bone, 36*(1), 101-110.

Godara, H.L. (2014). Construction of physical fitness norms for School Students. *International Journal of Research,3*,162-165.

Hair, J., Black, W. C., Babin, B. J. & Anderson, R. E. (2010). *Multivariate data analysis (7th ed.). UpperSaddle River*, New Jersey: Pearson Educational International.

Hashim, A. &Gunathevan. (2015). 900 Push-Up Test Norms Sport Sport Science Students Sultan Idris Education University. *International Journal of Development and Emerging Economics, 31*,1-9.

Hashim, A. *Physical Education Measurement and Assessment Test*. Selangor: Dubook Press Sdn Bhd.

Hattie, J., & Cooksey, R. W. (1984). Procedures for assessing the validities of tests using the"

known-groups" method. *Applied Psychological Measurement, 8*(3), 295-305.

Hulteen, R. M., Barnett, L. M., True, L., Lander, N. J., del Pozo Cruz, B., & Lonsdale, C. (2020). Validity and reliability evidence for motor competence assessments in children and adolescents: A systematic review. *Journal of Sports Sciences, 38*(15), 1717-1798.

Jacob, B., & Rothstein, J. (2016). The measurement of student ability in modern assessment systems. *Journal of Economic Perspectives, 30*(3), 85-108.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155-163.

Madruga-Parera, M., Bishop, C., Fort-Vanmeerhaeghe, A., Beltran-Valls, M. R., Skok, O. G., & Romero-Rodríguez, D. (2020). Interlimb asymmetries in youth tennis players: Relationships with performance. *The Journal of Strength & Conditioning Research*, 34(10), 2815-2823.

Mahar, M. T., & Rowe, D. A. (2002). Construct validity in physical activity research. *Welk*, GJ, editor.

McConnell, S., Kolopack, P., & Davis, A. M. (2001). The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC): a review of its utility and measurement properties. Arthritis Care & Research: *Official Journal of the American College of Rheumatology, 45*(5), 453-461.

Messick, S., & Linn, R. L. (1989). *Educational measurement. New York: American Council on Education (3rd ed).* London: Macmillan Pub. Co, 13-103.

Miller, D. (2014). *Measurement by the physical educator why and how*. McGraw-Hill Higher Education.

Ministry of Education. (2019). *Skill Test Manual for Physical Education*. Malaysia Institute of Teacher Education.

Morrow Jr, J. R., Mood, D., Disch, J., & Kang, M. (2015). *Measurement and Evaluation in Human Performance*, 5E. Human Kinetics.

Morrow,J,R, Jackson, A,W.,Disch, J.G, & Mood, D.P. (2005). *Measurement and Evaluation in Human Performance (3rd ed.).* Champaign IL: Human Kinetics.

Morrow, J.R., & Safrit, M.J. (1989). Measurement concepts in physical education and exercise science. Human Kinetics, Champaign, Il

Owens Jr, E.F., Hart, J.F., Donofrio, J.J., Haralambous, J., & Mierzejewski, E. (2004). Paraspinal skin temperature patterns: an interexaminer and intraexaminer reliability study. *Journal of Manipulative and Physiological Therapeutics*, 27(3), 155-159.

Pasand, F., Keshavarz, M., Mirzaei, S. & Zeinali. (2015). Evaluation of Standing Long Jumping Fundamental Skill Developmental Sequence in Children. *Int J Biol Pharm Allied Sci–IJBPAS, 4*(7), 4568-4578.

Tejero-Gonzalez, Carlos Mª, David Martinez-Gomez, Jorge Bayon-Serna, Rocio Izquierdo-Gomez, Jose Castro-Pinero, and Oscar L. Veiga. (2013). Reliability of the ALPHA health-related fitness test

battery in adolescents with Down syndrome. *The Journal of Strength & Conditioning Research* 27(11), 3221-3224.

Thomas, J. R., & Nelson, J. K. (2001). *Research Methods in Physical Activity (4th ed.)*. Champaign, IL: Human Kinetics.

Portney, Leslie Gross, and Mary P. Watkins. (2009). *Foundations of Clinical Research: Applications to Practice. Vol. 892*. Upper Saddle River, NJ: Pearson/Prentice Hall.

Reid, C., M. Dolan, and M. DeBeliso. (2017). The reliability of the standing long jump in NCAA track and field athletes. *International Journal of Sports Science 7*(6), 233-238.

Reid, Kieran F., and Roger A. Fielding. (2012). Skeletal muscle power: a critical determinant of physical functioning in older adults. *Exercise and Sport Sciences Reviews 40*(1), 4.

Ronnestad, Bent R., Oystein Kojedal, Thomas Losnegard, Bent Kvamme, and Truls Raastad. (2012). Effect of heavy strength training on muscle thickness, strength, jump performance, and endurance performance in well-trained Nordic Combined athletes. *European Journal of Applied Physiology* 112(6), 2341-2352.

Saint-Maurice, Pedro F., Kelly R. Laurson, Mónika Kaj, and Tamás Csanyi. (2015). Establishing normative reference values for standing broad jump among Hungarian youth. *Research Quarterly for Exercise and Sport 86*(1), S37-S44.

Sharma, R. J. (2017). Preparation of physical fitness norms for boys' aspirants for entrance test. *International Journal of Physical Education, Sports and Health*, 4(2), 135-136.

Stauffer, Kory, Elizabeth Nagle, Fredric Goss, and Robert Robertson. (2010). Assessment of anaerobic power in female division I collegiate basketball players. *Journal of Exercise Physiology Online, 13*(1), 1-10.

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health Measurement Scales: A Practical Guide to Their Development and Use*. Oxford University Press, USA.

Taipale, R. S., J. Mikkola, V. Vesterinen, A. Nummela, and K. Hakkinen. (2013). Neuromuscular adaptations during combined strength and endurance training in endurance runners: maximal versus explosive strength training or a mix of both. *European Journal of Applied Physiology, 113*(2), 325-335.

Thomas, E., Petrigna, L., Tabacchi, G., Teixeira, E., Pajaujiene, S., Sturm, D. J., Sahin, F.N, Lopez, M.G., Pausic, J., Paoli, A., Alesi, M., & Bianco, A. (2020). Percentile values of the standing broad jump in children and adolescents aged 6-18 years old. *European Journal of Translational Myology, 30*(2).

Wakai, Masaki, and Nicholas P. Linthorne. (2005). Optimum take-off angle in the standing long jump. *Human Movement Science 24*(1), 81-96.

Yelboga, A., & Tavsancil, E. (2010). The Examination of Reliability According to Classical Test and Generalizability on a Job Performance Scale. Educational Sciences: *Theory and Practice, 10*(3), 1847-1854.

Yin, L., Tang, C., & Tao, X. (2018). Criterion-Related Validity of a Simple Muscle Strength Test to Assess Whole Body Muscle Strength in Chinese Children Aged 10 to 12 Years. *Biomed Research International, 2018*, 2802803. https://doi.org/10.1155/2018/2802803