Journal of Physical Education and Sport ® (JPES), Vol. 21 (4), Art 206, pp. 1629 - 1634, June 2021
online ISSN: 2247 - 806X; p-ISSN: 2247 – 8051; ISSN - L = 2247 - 8051 © JPES

Original Article

Objectivity of push-up tests and technique assessment

PETR KELLNER[1], JIŘÍ NEUBAUER[2], MICHAL POLÁCH[3]
[1,3] Physical Training and Sport Centre
[2]Department of Quantitative Methods, University of Defence, CZECH REPUBLIC

Published online: June 30, 2021
(Accepted for publication June 15, 2021)
DOI:10.7752/jpes.2021.04206

Abstract:
Push-up tests are frequently used to examine arm and upper body strength and endurance in both civilian and military sectors worldwide. The purpose of this study was to explore the objectivity of push-up performance assessments. Ten experienced raters individually assessed 50 videotaped push-up test performances using three different approaches. The three methods were 1) repetition counting with real-time view of videos, 2) repetition counting after in-depth analysis of each performance, and 3) marking performances that met the high-quality technique execution standard. Statistical analysis showed significant inter-rater differences in counting for both methods 1 and 2. The intra-rater comparison of methods 1 and 2 examined by the Wilcoxon pair test showed significant differences in the counting by 7 raters. The overall comparison of the methods showed significantly higher repetition counting for method 1. In the technique assessment part of the study, the overall agreement between raters was 77.7%. None of the performances received approval from all 10 raters, but 18 of the performances did not meet the requirements of any rater; thus, poor technique was more likely to be identified than high quality technique. When the technique assessment was taken into consideration, the raters who agreed on quality execution of the performances counted similarly using both methods 1 and 2, whereas without this agreement, the counting in both methods 1 and 2 was statistically different. The findings of this study suggest that push-up tests might be a reliable tool for physical examination only if the high-quality technique standard is met by the examinee.
Key Words*:* **reliability, testing, examination, strength, performance, quality**

Introduction

The push-up exercise is an excellent exercise to develop arms, shoulders, and upper body strength/endurance; therefore, it is widely used in physical preparation and training (Cogley et al., 2005; Mayhew et al., 1991). For the same reason and for its ease of realization, it is also very commonly used for physical evaluation in both military and civilian sectors worldwide. Even though there are multiple push-up tests mostly oriented toward the number of repetitions during a given time interval or cadence, the general concept usually requires the examinee to hold the body straight at all times while alternately reaching the down and up positions (Plowman & Meredith, 2013; North, 2013; Department of the Army, 2012; LaChance & Hortobagyi, 1994). A scientific measurement instrument should display a sufficient level of objectivity, reliability, and validity. Objectivity is also referred to as rater reliability, and it is based on the extent to which the instrument is free from the rater´s personal opinion, and it is a condition of reliability (Gwet, 2014; Kimberlin & Winterstein, 2008; Hendl, 2006; Baumgartner et. al., 2003).

A push-up test not only requires the rater to count the number of repetitions executed but also to decide if the repetition technique is acceptable by the test standards and can be counted. This decision-making is influenced by the rater´s conscious and unconscious biases (Osório, 2020; Janack, 2002). There have been a number of studies on push-up test objectivity and reliability conducted in recent years, and although the particular methodology of each study differed to some extent, the overall findings and conclusions have been often contradictory. Most studies used a 90° push-up test variant, and some examined a bent-knee push-up test. For intra-rater reliability, the documented correlations range from r = 0.22 to r = 0.99 (Fielitz et al., 2016; Plowman & Meredith, 2013; Hashim, 2013; Morrow, 2010; Baumgartner et al., 2009; Wood & Baumgartner, 2004; McManis & Wuest, 1994). For inter-rater reliability, the documented correlations range from r = 0.10 to r = 0.99 (Fielitz et al., 2016; Hashim, 2013; Baumgartner et al., 2009; McManis et al., 2000; McManis & Wuest, 1994). The rater drift documented by Fielitz et al. (2016) is noteworthy because it resulted in a very substantial increase in repetition counting during prolonged assessment sessions due to the increased tolerance for exercise technique imperfection. According to their results, some study conclusions support the use of push-up tests as a reliable and acceptable instrument for physical fitness assessment, while others suggest that a more reliable test should be used. The purpose of this study was to investigate the quality assessment of push-up technique execution and repetition counting.

Corresponding Author: PETR KELLNER, E-mail: petr.kellner@unob.cz

Material & Methods

Participants The research sample performing the 30-second push-up test included 200 soldiers of the Army of the Czech Republic studying at the University of Defence, who were accustomed to executing the test. They were videotaped during the test, and 50 randomly selected records (76% male and 24% female) were chosen for study analysis. The rater group included 10 teachers of the Physical Training and Sports Centre of the University of Defence. The raters comprised six military and four civilian personnel with an average age of 46 ± 10.44 years and an average of 18.2 ± 11.51 years of experience in teaching physical education. All raters were experienced in assessing push-up tests and have administrated them on a regular basis multiple times every year.

Procedure The 30-second push-up test is a standardized test used in the Czech Army for annual physical testing. The examinee assumes the front-leaning rest position with the hands fully extended and shoulder-width apart with feet together, and the body should form a generally straight line from shoulders to ankles; it is required to maintain the "rigid" straight body shape without any sway for the entire test. By flexing the arms, the examinee lowers the body to the down position where the chest must touch the ground. Failure to keep the body rigid or not to reach either up or down position leads to an uncounted repetition. The number of correct repetitions performed during a 30-second interval are counted (Ministry of Defence, 2011). The scorer usually takes the position 45° angle forward from the examinee´s shoulder, which is close enough to be able to put a hand on the ground below the examinee´s chest to check if the down position is reached. In cases where the chest-ground contact was not distinguishable enough, the raters were instructed to accept the down position if the upper arm position was at least parallel to the ground. The viewing angle of the recordings mimicked the usual rater viewing angle.Three methods of evaluation were used. In the first method, the raters replicated the typical conditions of an actual test, which means repetition counting during a single real-time view of the recordings. In the second method, the raters were allowed to view videos with multiple playbacks with slow motion and pausing without any limitations. No additional specialized graphical analytic or editorial software was allowed. The last assessment method was carried out after the first view of the video and required raters to decide if the push-up technique execution was high quality without any (even small and still acceptable) compromises. To avoid rater drift deterioration (Wilson & Engelhard, 2000), the raters were instructed to assess only a few videos in a single session and to take breaks for as long as they needed to feel fully rested.

Data analysis In this study, multiple analyses were performed. Inter-rater reliability was evaluated for both assessment methods 1 and 2 using the Friedman test, and the Nemeniy post-hoc test was used for deeper insight into rater agreement. The intra-rater agreement and overall comparison of the two assessment methods were achieved using the Wilcoxon pair test. The second area examined was the inter-rater agreement on quality assessment of push-up technique using cross-raters number of agreements and number of agreements for each performance. The third area of analysis was evaluation of inter-rater reliability and intra-rater agreement separately for executions marked as high quality and for executions with technical flaws using the Kruskal–Wallis and Wilcoxon pair tests. Statistical software R was used for calculations, and tests were performed at a significance level of 0.05.

Results

Data for the analysis were collected from 10 raters assessing 50-videotaped performances of the 30-second push-up test. Inter-rater reliability of both counting method 1 (single real-time view) and method 2 (detailed view) showed that the raters did not count equally (the Friedman test p-values were $5.913 \cdot 10^{-16}$ for method 1 and $1.446 \cdot 10^{-14}$ for method 2). The descriptive statistics are shown in Figures 1 and 2. The clearly visible difference in scoring from rater 9 came from an extremely strict focus on the shoulder-width hand placement requirement, resulting in multiple performances scored with 0 repetitions counted, which were considered acceptable for counting by other raters.

Fig. 1 – Boxplot of push-up repetition counting via Method 1 by individual raters

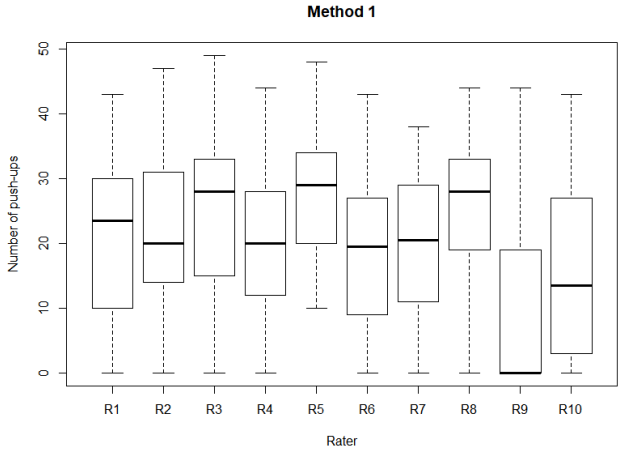
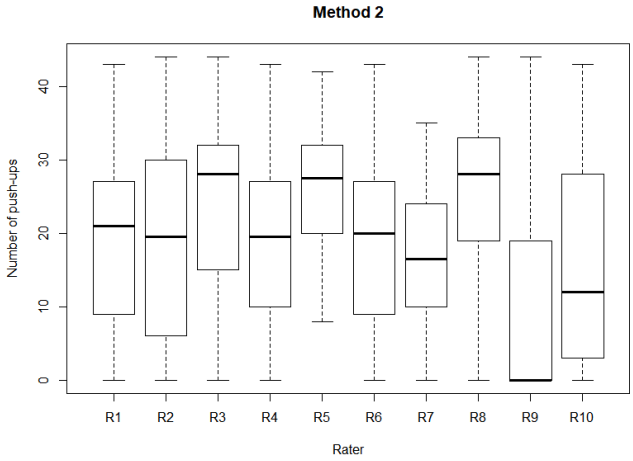


Fig. 2 – Boxplot of push-up repetition counting via Method 2 by individual raters



The Nemeniy post-hoc analyses presented in Table 1 and Table 2 show the rater differences. The number of differently counting raters varied from 1 to 7 raters for method 1 and from 1 to 5 raters for method 2. Raters 1, 6, and 7 scored mostly corresponding with other raters that used method 1, whereas for method 2, the most corresponding scorings were detected for raters 1 and 2. Rater 5 scored the least similar for both methods.

Table 1 – Comparison of evaluators – Method 1, Nemenyi post-hoc test p-values (statistically significant differences at level 0.05 are in bold)

Method 1	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Rater 9
Rater 2	0,99571	-	-	-	-	-	-	-	-
Rater 3	0,61201	0,98857	-	-	-	-	-	-	-
Rater 4	0,91012	0,33898	**0,02100**	-	-	-	-	-	-
Rater 5	**0,00113**	**0,03603**	0,44767	**0,00000**	-	-	-	-	-
Rater 6	0,99999	0,93147	0,29948	0,99283	**0,00014**	-	-	-	-
Rater 7	1,00000	0,98857	0,51752	0,94907	**0,00064**	1,00000	-	-	-
Rater 8	0,70320	0,99571	1,00000	**0,03243**	0,35967	0,38094	0,61201	-	-
Rater 9	0,18363	**0,01182**	**0,00014**	0,96905	**0,00000**	0,44767	0,24540	**0,00026**	-
Rater 10	0,65836	0,12302	**0,00385**	0,99999	**0,00000**	0,91012	0,74588	**0,00643**	0,99907

Table 2 – Comparison of evaluators – Method 2, Nemenyi post-hoc test p-values (statistically significant differences at level 0.05 are in bold)

Method 2	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Rater 9
Rater 2	1,00000	-	-	-	-	-	-	-	-
Rater 3	0,64689	0,81367	-	-	-	-	-	-	-
Rater 4	0,78582	0,61201	**0,00988**	-	-	-	-	-	-
Rater 5	0,05676	0,11792	0,97419	**0,00006**	-	-	-	-	-
Rater 6	1,00000	0,99996	0,47070	0,90418	**0,02617**	-	-	-	-
Rater 7	0,78582	0,61201	**0,00988**	1,00000	**0,00006**	0,90418	-	-	-
Rater 8	0,39177	0,57667	1,00000	**0,00242**	0,99821	0,24540	**0,00242**	-	-
Rater 9	0,26270	0,14512	**0,00038**	0,99870	**0,00000**	0,41381	0,99870	**0,00007**	-
Rater 10	0,42499	0,26270	**0,00122**	0,99995	**0,00000**	0,60027	0,99995	**0,00024**	1,00000

The intra-rater agreement of each rater examined by comparing their scorings via method 1 and 2 using the Wilcoxon pair test showed significant differences in scoring by 7 raters. Raters 6, 9, and 10 scored equally using both methods. The overall comparison of method 1 and 2 using the Wilcoxon pair test showed a significantly higher repetition counting during method 1, as shown in Table 3.

Table 3 – Descriptive statistics (n–the number of observations; Mean–arithmetic mean; Median–median; St. dev. –standard deviation; Min–minimum value; Max–maximum value; Q0.25–lower quartile; Q0.75–upper quartile; Skewness–skewness, Kurtosis–kurtosis)

	n	Mean	St. dev.	Median	Min	Max	Q0.25	Q0.75	Skew	Kurtosis
Method 1	500	20,392	12,719	21	0	49	11	30	−0,196	−1,018
Method 2	500	19,178	12,426	20	0	44	9	29	−0,140	−1,113

The inter-rater agreement for push-up technique assessment presented in Table 4 ranged from 63.1% to 82.2% with an average of 77.5 ± .72%. The least agreement in technique assessment was displayed by rater 5 (63.1%) and rater 8 (72.9%), while the remaining 8 raters ranged from 76.9% to 82.2%. The highest agreement reached was shown by raters 10 (82.2%) and 6 (81.8%). The average performances marked for high-quality by each rater was 8.6 ± 5.87 with a modus of 7 and median of 7.

The number of agreements in the assessment for each video are shown in Table 5. None of the performances received the approval of all 10 raters, and 18 of the performances did not meet the requirements of any rater. The average high-quality approvals was 1.72 ± 2.07 with a modus of 0 and median of 1. Of note, rater 4 did not mark any of the performances as a high-quality execution, and rater 5 marked 22 as high-quality performances.

Table 4 – Push-up technique quality assessment (Good – video marked as a high-quality performance; Bad – video marked as involving technique flaws)

Video	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
Good	5	1	8	0	0	3	2	3	1	0
Bad	5	9	2	10	10	7	8	7	9	10
Video	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20
Good	0	0	7	2	0	3	0	1	0	2
Bad	10	10	3	8	10	7	10	9	10	8
Video	V21	V22	V23	V24	V25	V26	V27	V28	V29	V30
Good	1	4	4	2	3	3	6	1	1	0
Bad	9	6	6	8	7	7	4	9	9	10
Video	V31	V32	V33	V34	V35	V36	V37	V38	V39	V40
Good	0	3	1	0	0	0	0	1	7	1
Bad	10	7	9	10	10	10	10	9	3	9
Video	V41	V42	V43	V44	V45	V46	V47	V48	V49	V50
Good	0	1	0	3	2	0	1	0	0	3
Bad	10	9	10	7	8	10	9	10	10	7

When the push-up technique assessment was taken into consideration and repetition counting was evaluated separately for raters who agreed on the technique execution quality, two new categories were established for both assessment methods 1 and 2: “Good“ for performances marked as high-quality execution and “Bad“ for performances marked as involving technique flaws.

For inter-reliability of repetition counting, the Kruskal–Wallis test was used. For the Good category, there were 86 cases of agreements, and the repetition counting was discovered to be comparable for both methods 1 and 2 (the p-value was 0.6473 for method 1 and 0.6399 for method 2). For the Bad category, there were 414 cases of concordance, and repetition counting was found to be unequal for both methods 1 and 2 ($p = 3.916 \cdot 10^{-11}$ for method 1 and $p = 3.176 \cdot 10^{-10}$ for method 2).

The overall comparison of methods 1 and 2 in repetition counting using the Wilcoxon test showed agreement for the Good category (p = 1; the average differences were 0.012 ± 0.242) and differences for the Bad category (p-value was almost 0; average differences were 1.468 ± 4.855). In the Bad category, method 1 displayed significantly higher repetition counting then method 2.

The intra-rater comparison for both technique categories using the Kruskal–Wallis test showed comparable repetition counting for the Good category (p = 0.2137) and differences for the Bad category ($p = 3.533 \cdot 10^{-13}$), as shown in Figures 3 and 4.

Fig. 3 – Boxplot of differences in counting between methods 1 and 2 for the Good category

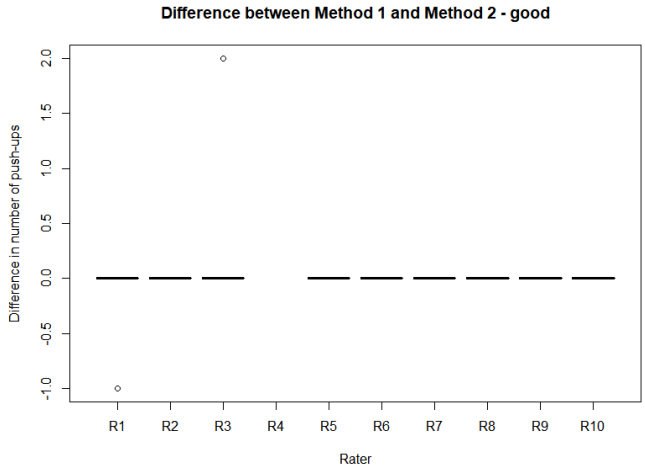
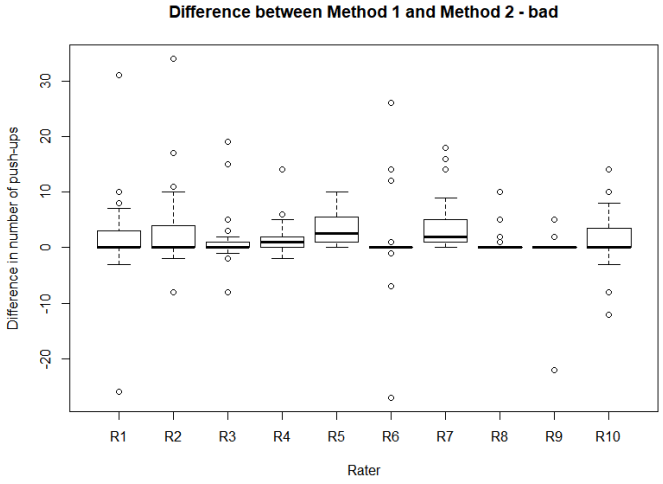


Fig. 4 – Boxplot of differences in counting between methods 1 and 2 for the Bad category



Discussion

Push-up tests are used to determine the physical capabilities of an examinee. Even more importantly, the results are often taken into consideration for personal evaluation in military organizations and physical education classes; therefore, accurate and proper grading is desirable to ensure equal opportunity for any person tested. The studies conducted to determine objectivity and reliability of push-up tests provides a variety of often contradicting results and conclusions. The presented study explored the objectivity of push-up repetition counting, which proceeds mainly from assessment of the technique of exercise execution, and whether each repetition performed should be counted. The 30-second push-up test used in this study is distinguished by rapid cadency often exceeding 1 push-up per second, which makes proper assessment difficult for raters. Inter-rater reliability examined by repetition counting during a single real-time view of videos (method 1) using the Freidman test showed significant differences in counting ($p = 5.913 \cdot 10^{-16}$). Similar significant counting differences ($p = 1.446 \cdot 10^{-14}$) were discovered for detailed video analysis when multiple playbacks, slow motion, and pausing were allowed (method 2). The similarities in counting varied for most raters for methods 1 and 2. Only one rater scored mostly the same using both methods, while others varied from 1 to 7 agreements with other raters. One rater scored the least similar using both methods. Based on these data, the low reliability reported in some previous studies seems to be more accurate than the opposite findings of some other investigations. The discrepancy might be caused by using only 2 raters for inter-reliability evaluation (Baumgartner et al., 2009; Hashim, 2013), while employing 8 raters (Fielitz, 2016) or 10 raters (in this study) produced a generally wider variety in assessment, even though few raters scored similarly.

Our intra-rater comparison of methods 1 and 2 showed significant differences in the counting for 7 raters. Method 1 showed a significantly higher repetition counting then method 2. These findings suggest that a higher exercise cadency leads to lowering standards by the raters. This statement should not be misinterpreted for slower execution that results in better reliability because scoring via method 2 showed similarly significant

differences in counting despite the fact that the raters were given all the time and performance execution speed needed to judge every repetition to the best of their abilities.

For the inter-rater quality assessment of technique execution, the average agreement was 77.5 ± .72%. Rater 5 with 22 high-quality assessments achieved 63.1% agreement with other raters, and the same rater also scored the least similarly in both counting methods 1 and 2. Three raters reached almost 80% agreement, and 4 raters were above 80%. Although eighteen performances were marked as high quality by 0 raters and 11 performances received only one approval, only 3 performances received 7 or 8 high-quality approvals. These findings surprisingly suggest that the opinion of high-quality technique performance varies between raters, unless the execution is truly exceptional, whereas poor execution is more likely to be identified. Consideration of technique assessment for counting evaluation led to vast repetition counting reliability changes. The high-quality performances received similar scoring via both methods 1 and 2, whereas the performances without the high-quality mark remained unequally counted in both methods. A similar trend was discovered during comparison of methods 1 and 2 in which high-quality performances received almost identical scores via both methods, whereas performances without the quality approval were still counted significantly higher in method 1.

Conclusions

Our study findings support the statement of low assessment objectivity of push-up tests discovered in some previous studies. Assessing proper execution seems to be very difficult by the naked eye only, and even though a clicker board or lasers might be able to facilitate the measurement of the "up" and "down" positions, rigid straight body position assessment remains extremely problematic. A high rater reliability was discovered only for performances with high-quality push-up technique, which allows raters to count every repetition without hesitation. Push-up technique evaluation methods are recommended for further research. Until a new approach for performance assessment is discovered, technical devices for automated movement analysis might be the only option to assure fair evaluation of push-up tests in the general population.

References

Baumgartner, T.A., Strong, C.H, Mahar, M.T., Rowe, D.A. (2003) *Measurement for evaluation in physical education and exercise science (7th. Ed.)*. New-York: McGraw-Hill.

Baumgartner, T.A., Oh, S., Chung, H., Hales, D. (2002). Objectivity, reliability and validity for a revised push-up test protocol. *Measurement in Physical Education and Exercise Science*, 6, 4, 225-242.

Department of the Army. (2012) *Army field manual (FM 7-22)*. Washington, DC. Available at https://www.atu.edu/rotc/docs/aprt_7-22.pdf; accessed December 21, 2020.

Fielitz, L., Coelho, J., Horne, T., Brechue, W. (2016). Inter-Rater Reliability and Intra-Rater Reliability of Assessing the 2-minute Push-Up Test. *Military Medicine*, 181, 2:167-172.

Hashim, A. (2013). Objectivity, reliability and validity of the 90°push-ups test protocol among male and female students. *Global journal of medical research interdisciplinary*, 13, 5.

Hendl, J. (2006) *Přehled statistických metod zpracování dat: Analýza a metaanalýza dat*. 3. vydání. Praha: Portál. ISBN 80-7178-820-1.

LaChance, P., Hortobagyi, T. (1994). Influence of Cadence on Muscular Performance During Push-up and Pull-up exercise. *Journal of Strength and Conditioning*, 8,2, 76-79.

McManis, B.G., Baumgartner, T.A., Wuest, D.A. (2000). Objectivity and reliabilty of the 90° push-up test. *Measurement in Physical Education and Exercise Science*, 4, 1, 57-67.

McManis, B.G., Wuest, D.A. (1994). Stability reliability of the modified push-up in children [Abstract]. *Res Q Exerc Sport*, 65, 54-59.

Ministry of Defence (2011). Normativní výnos ministra obrany č.12/2011. Praha: Ministerstvo obrany.

Morrow, J.R., Martin S.B., Jakcson A.W. (2010). Reliability and validity of the Fitnessgram: quality of teacher-collected health-related fitness surveillance data. *Res Q Exerc Sport*, 81, 3, 24-30.

Osório, A. (2020). Performance Evaluation: Subjectivity, Bias and Judgment Style in Sport. *Group Decision and Negotiation*, 29, 655-678.

Plowman, S.A. & Meredith, M.D. (2013). *Fitnessgram/Activitygram Reference Guide (4th Edition).* Dallas, TX: The cooper institute.

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Wood, H.M., Baumgartner, T.A. (2004). Objectivity, reliability and validity of the bent-knee push-up for college-age women. *Measurement in Physical Education and Exercise Science*, 8, 4, 203-212