Journal of Physical Education and Sport ® (JPES), Vol. 23 (issue 7), Art 209, pp. 1706 - 1713, July 2023
online ISSN: 2247 - 806X; p-ISSN: 2247 – 8051; ISSN - L = 2247 - 8051 © JPES

Original Article

The reasoning behind assessing push-up tests – an in depth analysis

PETR KELLNER[1], JIŘÍ NEUBAUER[2], ZDEŇKA KELLNEROVÁ[3], PETR ZAHRADNÍČEK[4], VIKTOR NOVOTNÝ[5], LIBOR WAWRZACZ[6], MICHAL PUNČOCHÁŘ[7]
[1,4,5,6,7]Physical Training and Sport Centre,
[2] Department of Quantitative Methods,
[3] School Regiment University of Defence, CZECH REPUBLIC

Published online: July 31, 2023
(Accepted for publication July 15, 2023)
DOI:10.7752/jpes.2023.07209

Abstract:
A good exercise does not necessarily make for a good measurement tool and using such a tool may lead to wrong conclusions if used for scientific measurement and personal evaluation, yet it happens when using push-up tests which are subjective and lack reliability. This study examined the reasons behind the questionable reliability of push-up testing. Material and Methods: Fifty videorecorded 30-second push-up test performances were evaluated by 10 highly experienced raters in two separate assessment trials. The assessment involved counting the number of acceptable repetitions and identifying any technical flaws in the execution of the exercise. The collected evaluations were analyzed using quantitative and qualitative methods. Results: Statistical analysis ($p \leq 0.05$) revealed significant inter-rater differences in counting in both trials. Comparable counting was only found among raters who marked the same technique as "perfect" and overall concordance on perfect execution was 79.4%. Intra-rater counting reliability ranged from $r = 0.57$ to $r = 0.92$. Three main areas of technique deterioration were identified: incomplete arm extension (10.2% of denied repetitions), inadequate arm flexion (7%), and failure to keep the body straight and rigid (6.3%), which was also the most disputed between the raters. Additionally, male raters were more lenient towards the technique imperfections of female subjects. Many miscalculations were also detected, often correlated with perfect technique execution (88% of cases). The second most common cause of miscalculating was raters' willingness to count a repetition that was interrupted mid-execution due to time constraints. Conclusions: The study findings indicate that push-up assessment is highly subjective and should be avoided in scientific or personal evaluations that require a higher level of precision. The reliability of the assessment heavily depends on the individual administering the test, and the average evaluator demonstrates only moderate reliability. To mitigate gender-based bias, considering a female evaluator for female examinees is recommended. Therefore, caution is advised when relying on push-up tests when more reliable alternatives are available.
Key Words: objectivity, subjectivity, reliability, technique, evaluation, gender

Introduction

During the past decades, multiple studies of the objectivity of push-up tests have been conducted, and the majority of these studies have reported low reliability. Intra-rater correlations have been documented to range from $r = 0.22$ to $r = 0.99$, while inter-rater correlations have ranged from $r = 0.1$ to $r = 0.99$ (Kellner et al., 2021; Fielitz et al., 2016; Plowman & Meredith, 2013; Hashim, 2013; Morrow, 2010; Baumgartner et al., 2009; Wood & Baumgartner, 2004; McManis et al., 2000; McManis & Wuest, 1994). Despite these findings, push-up tests are still frequently used for evaluating arm, shoulder, and upper body strength, power and strength endurance across a range of fields, including the military, sports, schools, scientific studies, and medical monitoring. Other exercises or more sophisticated measurement tools are also used (Bianchim et al., 2022; Davies et al., 2022; Soriano et al., 2022), but push-up tests remain popular for general observations and fitness evaluation programs (Crosby et al., 2023; Adams et al., 2022; Gorner & Reineke, 2020; Iermakov et al., 2021; Mischenko et al., 2020). In particular, some variation of the push-up test is commonly used in armies worldwide (Wiliams et al., 2020; Barringer et al., 2019; Tomczak & Haponik, 2016) and this trend is unlikely to change in the near future for various reasons, including the ease of administration, as no additional equipment is required, and the tradition that is a significant factor in the armed forces.

Disman (2022) emphasizes the importance of reliability in any scientific measurement. Inter-rater reliability (objectivity) refers to the degree of agreement in scores obtained from two or more raters, while intra-rater reliability (stability) refers to the agreement of evaluation within one rater (Lacy, 2018; Baumgartner, 2003). Using a nonreliable tool is likely to lead to incorrect conclusions, where results may be influenced by the measuring tool itself more than the actual measured quality. Motor skill assessments can be product- or process-oriented, where product assessments focus on the outcome of the skill execution, such as time, distance, or

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

successful attempts, while process assessments are concerned with how the skill is performed (Barnett et al., 2009). The push-up test is generally oriented towards the number of repetitions performed within a given time interval or cadence, with a single repetition defined as a movement cycle consisting of reaching the down and up positions while keeping the body straight (Kellner et al., 2021). The assessment is a combination of both product and process approaches, as it is first necessary to determine whether each repetition should be counted according to the testing standard. The decision-making process is influenced by the rater's conscious and unconscious biases (Osório, 2020). Fielitz et al. (2016) documented an increase in repetition counting due to rater drift in prolonged assessment sessions, caused by raters lowering their technique requirements. Kellner et al. (2021) documented a minor increase in repetition counting due to a higher exercise cadence. In none of the studies on the subject was a large portion of the counting discrepancy between raters explained. Previous studies focused solely on repetition counting, typically comparing total scores, while having an insufficient number of raters or repetitions assessed. This limitation may have led to the omission of important aspects of the matter. It is noteworthy that while anthropometric research in detailed knowledge is advancing, the foundations in some of more general aspects, like measurement, are being neglected. Therefore, a better understanding of the mechanisms behind the commonly used test administrator's decision-making process during the assessment is desirable and was researched in this study.

Material & Methods

Participants

The research sample for the 30-second push-up test consisted of 200 soldiers from the Army of the Czech Republic (76% male, 24% female) who were accustomed to both the push-up exercise and the test. The testing sessions were videotaped, and 50 randomly selected recordings were chosen for analysis after eliminating any erroneous recordings.

The raters consisted of 10 teachers of physical education at the University of Defence (9 male, 1 female), 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 or physical readiness training. All raters were accustomed to administering the push-up test on a regular basis multiple times every year.

The research was approved by the Ethical Board of the University of Defence, and each participant signed an informed consent form.

Procedure

The 30-second push-up test used in this study is a standardized test used in the Czech Army. From the front-leaning rest position with the hands fully extended (the up position) and shoulder-width apart, with feet together, the examinee lowers the body by flexing the arms to the down position, where the chest must touch the ground. It is required to maintain a "rigid" straight body shape from shoulders to ankles without any sway for the entire test. The number of repetitions performed during a 30-second interval is counted. Failure to reach either the down or the up position, or to keep the body rigid, leads to an uncounted repetition (Ministry of Defence, 2011). Video recordings were taken from a position mimicking the usual position of a scorer, which is 45° forward from the examinee's shoulder. During field testing, the scorer can place a hand on the ground below the examinee's chest to check if the down position is reached. For the purposes of the study, raters were allowed to accept the down position if the upper arm was at least parallel to the ground. Viewing of the recordings followed the typical conditions of an actual test by watching each recording only once, and no replays, slowing, pausing, or other manipulation of the recording was allowed. The assessment of the recordings was divided into multiple sessions to avoid rater drift (Wilson & Engelhard, 2000).

For stability purposes, two assessment trials, one year apart, were carried out. Raters recorded the number of accepted repetitions, whether the push-up technique was of high quality without any compromises, or if there were small but acceptable deteriorations from the technique standard, and alternatively, what technical flaws led to uncounted repetitions.

Data analysis

This study utilized both qualitative and quantitative analysis. Statistical software R was used for calculations, and tests were performed at a significance level of 0.05. Qualitative analysis of the reasoning behind the assessment was conducted by categorizing the registered technical flaws, comparing inter-rater assignments of features, retroactively assessing the recordings, and obtaining lessons-learned notes or brief unstructured interviews with the raters. The intraclass correlation coefficient (ICC) scores can range from 0 to 1, with interpretation of the score suggested as follows: "values less than 0.5 are indicative of poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values greater than 0.90 indicate excellent reliability" (Koo & Li, 2016). Inter-rater reliability was evaluated using the Friedman test, and the Nemeniy post-hoc test was used for a deeper understanding of rater agreement. Intra-rater reliability was evaluated using the Wilcoxon pair test. The cross-raters' number of agreements and number of agreements for each performance were used to evaluate agreement on technique quality assessment.

The evaluation of inter-rater and intra-rater reliability was also carried out separately for executions marked as high quality and for executions with technical flaws using the Kruskal-Wallis and Wilcoxon pair tests (Hendl, 2006).

Results

Data for the analysis was collected in two assessment trials in which 50 videotaped performances of the 30-second push-up test were assessed by 10 raters. The overall intra-rater reliability of all raters for repetition counting, as examined by the Pearson and Spearman correlation coefficients, was r = 0.74 and r = 0.75, respectively, indicating moderate reliability. Individual reliability values among the raters ranged from r = 0.57 to r = 0.92 and are shown in Table 1.

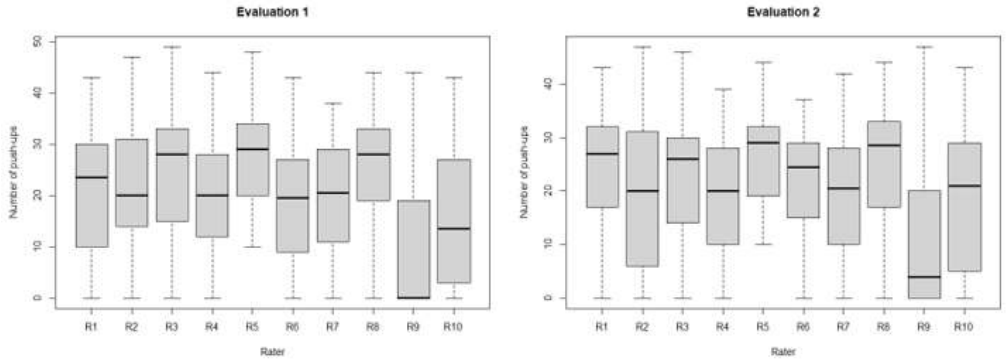
Table 1 – Individual intraclass correlation coefficients – repetition counting

	Pearson correlation	Spearman correlation
Rater 1	0,6332358	0,6154109
Rater 2	0,9266716	0,9282551
Rater 3	0,6820608	0,6917947
Rater 4	0,5736844	0,6056793
Rater 5	0,8881948	0,8791572
Rater 6	0,7215026	0,7310261
Rater 7	0,751104	0,7525979
Rater 8	0,8302266	0,8477048
Rater 9	0,6094051	0,5796897
Rater 10	0,6677212	0,6344716

Rater 9 showed an increase in repetition counting in the second assessment trial. In the first trial, many executions were denied and scored 0 because of extremely strict judging of hand placement width by the rater, whereas in the second trial, some of the previously denied executions were accepted and some (fewer) of the previously counted were denied. When the contradictory executions are eliminated from the calculation, rater 9's reliability increases to r = 0.94.

Analysis of intra-rater reliability using the Wilcoxon test (paired) documents significant differences in counting by raters 1, 5, and 6. The descriptive statistics for inter-rater assessments in both trials are shown in Figure 1. The clearly visible difference in scoring by rater 9 came from his extremely strict focus on hand placement.

Figure 1 – Boxplot of push-up repetition counting in both trials by individual raters



The Friedman test was used to evaluate whether the raters were counting similarly, and the results indicated statistically significant differences in both trials (p-values 5.913·10^(-16) and 4,16·10^(-11)). The Nemeniy post-hoc analyses presented in Table 2 show the rater differences. The number of differences varied from 1 to 7 raters in the first trial and from 1 to 5 in the second trial, with overall concordance being higher in the second trial.

Table 2– Comparison of evaluators – Trial 1 and 2, Nemenyi post-hoc test p-values (statistically significant differences at level 0.05 are in bold)

Evaluation 1	R1	R2	R3	R4	R5	R6	R7	R8	R9
R2	0,99571	-	-	-	-	-	-	-	-
R3	0,61201	0,98857	-	-	-	-	-	-	-
R4	0,91012	0,33898	**0,02100**	-	-	-	-	-	-
R5	**0,00113**	**0,03603**	0,44767	**0,00000**	-	-	-	-	-
R6	0,99999	0,93147	0,29948	0,99283	**0,00014**	-	-	-	-
R7	1,00000	0,98857	0,51752	0,94907	**0,00064**	1,00000	-	-	-
R8	0,70320	0,99571	1,00000	**0,03243**	0,35967	0,38094	0,61201	-	-
R9	0,18363	**0,01182**	**0,00014**	0,96905	**0,00000**	0,44767	0,24540	**0,00026**	-
R10	0,65836	0,12302	**0,00385**	0,99999	**0,00000**	0,91012	0,74588	**0,00643**	0,99907
Evaluation 2	R1	R2	R3	R4	R5	R6	R7	R8	R9
R2	0,99970	-	-	-	-	-	-	-	-
R3	0,99675	1,00000	-	-	-	-	-	-	-
R4	0,68101	0,96905	0,99283	-	-	-	-	-	-
R5	0,50574	0,13934	0,07555	**0,00242**	-	-	-	-	-
R6	1,00000	0,99995	0,99907	0,76622	0,41381	-	-	-	-
R7	0,88485	0,99757	0,99981	1,00000	**0,00988**	0,93147	-	-	-
R8	0,99718	0,87791	0,75614	0,15109	0,96318	0,99191	0,32886	-	-
R9	**0,00361**	**0,03795**	0,07555	0,57667	**0,00000**	**0,00604**	0,32886	**0,00007**	-
R10	0,30911	0,75614	0,87791	0,99994	**0,00022**	0,39177	0,99626	**0,03243**	0,89798

When the push-up technique assessment was taken into consideration and repetition counting was evaluated separately for raters who agreed on high technique quality and for those who did not, using the Kruskal-Wallis test showed comparable counting for "perfect technique" (p-values 0.6473 and 0.9354) and differences for "not perfect technique" (p-values 9.916·10^(-11) and 2.698·10^(-9)) in both trials.

In the technique quality assessment part, the raters were first required to mark executions with perfect technique by the testing standards. In the first trial, most raters assigned significantly fewer perfect technique marks (86 total) than in the second trial (124 total). Overall concordance between the two trials was 79.40%, and individual concordances are shown in Table 3. Based on the test of equal proportions, it can be stated that percentages are not the same (p-value 0.001919).

Table 3 – Individual technique quality assessment percentages

Rater 1	84,00%	Rater 6	78,00%
Rater 2	90,00%	Rater 7	90,00%
Rater 3	78,00%	Rater 8	56,00%
Rater 4	78,00%	Rater 9	82,00%
Rater 5	74,00%	Rater 10	84,00%

Noteworthy, rater 4 marked 0 techniques as perfect in the first trial and 11 techniques in the second. Based on the interview, we concluded that his mindset in the first trial derived from his refusal attitude towards participation in the study and overall burn-out syndrome (when he enjoys teaching sport games only) and therefore he stated no technique was good enough for him. Despite the issue mentioned above, rater 4 achieved similar percentages as other raters who assessed techniques in both trials to the best of their abilities, meaning that finding stability in assigning perfect technique marks is also very problematic.
The inter-rater agreement on technique quality ranged from 63.1% to 82.2% with an average of 77% in the first trial, and from 52.6% to 86.3% with an average of 71.7% in the second trial.

For the decision-making examination, raters had to justify their repetition counting by stating reasons for denying repetitions to be counted or allowed technical deterioration from the standard not resulting in the rejection of the repetition. The reasons were categorised, and 5 main groups and 3 sub-groups emerged. The "Rigidness failure (overall)" group presents an observation of the examinee's failure to hold a straight, rigid body

shape and has 3 sub-groups: the "Rigidness failure" group presenting more general, unspecified sway, the "Back extension" group presenting mostly lumbar hyperextension usually occurring during the return from the "down" position, and the "Forward bending" group presenting mostly thoracic flexion usually occurring during movement from the "up" position and in the majority of cases, it is linked to minimalistic or almost static vertical position of the hips. "The down position" group presents a failure to touch the ground, or, for the previously mentioned study purposes, reach at least a horizontal position of upper arms. "The up position" group presents a failure to reach the "up" position indicated by fully extended arms. The "Hand placement" and the "Feet placement" groups represent a failure to place hands below the shoulders or to place feet together.
Each group/sub-group was attributed by the number of rejected repetitions and by how many raters, and the number of raters mentioning the mild but acceptable technical deterioration from the standard in each of the recordings. Every video recording was also inspected for the exact number of movement cycles (repetition attempts) and if there were movement cycles exceeding the time limit. For each rater, it was also analysed if there was concordance between repetition counted, perfect technique mark, technical flaws, and movement cycles. The summary is shown in Table 4.

Table – Repetition counting reasoning summary (data form presented for technique deteriorations columns - A/B (C) where A is the number of rejected repetitions, B is the number of raters reporting the deterioration, C is the number of raters reporting the mild deterioration. Perfect technique – number of raters reporting the perfect technique. Arithmetic failure – form E/F where E is the number of miscalculated repetitions and F is the number of miscalculating raters)

	Movement cycles assessed	Rigidness failure (overall)	Rigidness failure	Back extension	Forward bending	The down position
absolute	13870	875/84 (82)	397/49 (48)	255/30 (32)	233/15 (8)	980/57 (16)
relative (%)	100	6,3	2,9	1,8	1,6	7
		The up position	Hand placement	Feet placement	Perfect technique	Arithmetic failure
absolute		1414/88 (28)	552/19 (32)	0/0 (23)	123	70/64
relative (%)		10,2	4	0	24,6	12,8

During the 50 recordings, a total of 1387 movement cycles were detected. This resulted in a total of 500 video observations and 13,870 movement cycle assessments, as each video was assessed by 10 raters. A failure to maintain a straight body position accounted for 875 (6.3%) declined repetitions, which was the most difficult technique for the raters to assess due to the variety in identifying this deterioration among the rater group. The most severe problem for the examinees was fully extending their arms, resulting in 1414 (10.2%) uncounted repetitions. Reaching the "down" position was the second most difficult for the examinees, with 980 (7%) failed attempts. Hand placement accounted for 552 (4%) of the failed attempts, with only 19 reported observations (each scored as 0 repetitions). It is clear that proper hand placement is crucial for the performance to be counted, although in real testing sessions, the examinee's hand misplacement is corrected before testing starts. The same is true for feet placement, which did not result in any uncounted repetitions, although it was marked as a mild mistake in 23 instances.

The last area of exploration was arithmetic failure. In 64 cases, discrepancies in entered numbers were found, resulting in 70 incorrectly counted repetitions. These mistakes seem to have two potential causes, although they are not necessarily exclusive. The first cause is identified as a mistake made by the rater in their arithmetic (ranging from 1 to 2 miscalculated repetitions). In 56 (88%) of these cases, the mistake was accompanied by the rater marking the execution as perfect. When not paired with the second cause mentioned below, the miscalculation was predominantly negative, meaning the rater scored a lower number than was actually executed and should be counted. From questioning the raters, it was determined that when assessing non-problematic, technically perfect or almost perfect execution, raters are more relaxed and their attention on the task can sometimes be diverted, resulting in missed counts. The second identified cause of miscalculation was found in the arguable ending of the test, meaning when the examinee exceeds the time limit to finish the last repetition. 12 of the recordings (and therefore 120 evaluations) were ended in such a manner, and in 34 (28%) of these cases, the raters counted the repetition as finished even though it occurred after the time limit. If a rater also marked the video as perfect technique, the counting past the limit rose to 91% of the cases.

Discussion

The theory of measurement emphasizes two fundamental attributes of measurement instruments: reliability and validity. Both are necessary for credible output from any measurement, and the use of a tool without these qualities is likely to lead to incorrect conclusions. The reliability of push-up tests has been examined multiple times in previous decades, and the reliability of the test has not often been found to be satisfactory. The present

study investigates the reasons for this through qualitative and quantitative analysis. Intra-rater reliability over two assessment trials was examined using Pearson (r=0.74) and Spearman (r=0.75) correlation coefficients, indicating moderate reliability. The reliability of individual raters ranged from r=0.57 to r=0.92, showing a range from moderate to excellent reliability. The Wilcoxon test (paired) showed that three raters counted with significant differences between the two trials. Rater 9 was extremely strict on the hand placement rule in the first trial, resulting in multiple recordings being scored as 0, while they were counted by others. It is noteworthy that this rater was extremely permissive when assessing female performances and counted executions that were rejected by many others. A certain degree of leniency in counting female performances was discovered among all male raters, except for one female rater. This leniency was mentioned by a majority of the raters in interviews. Inter-rater reliability, examined using the Friedman test, displayed significant differences in counting in both trials (p-values 5.913·10^(-16) and 4,16·10^(-11)). Comparable counting in both trials was only found among the raters who agreed on "perfect technique" using the Kruskal-Wallis test (p-values 0.6473 and 0.9354).

Overall, there was a concordance of 79.4% in the identification of the perfect technique between the two trials. In the second trial, there were significantly more performances (44% more) marked as having a perfect execution. Rater 4 refused to mark any performances in the first trial, but assigned 11 marks in the second trial. This change in attitude from refusal in the first trial to a more forthcoming attitude in the second trial might be due to many factors, including neophobia (Corey, 1978), as the rater had not participated in any research for many years and was only accustomed to teaching routine obligations. The inter-rater agreement on technique quality averaged 77% and 71.7% in the first and second trials, respectively, indicating that identifying a perfect execution is questionable and difficult. It is noteworthy that rater 2 was the only rater who, in the second trial, did not assign a technique mark to any performance that had not previously been marked by him, while other raters assigned techniques inconsistently. Rater 2 also had the highest intra-rater reliability in repetition counting. This exceptional performance is likely due to many factors outside the scope of this study, but the rater demonstrated great dedication to the study and had extensive experience in movement assessment and testing.

In the decision-making justification part, three categories for real testing repetition denial were identified: failure to maintain a straight body position (with subcategories of general rigidness failure, back extension, and forward bending) accounted for 875 (6.3%) rejected repetitions, incomplete arm extension accounted for 1414 (10.2%) rejections, and failure to reach the down position accounted for 980 (7%) failures. The most disputed category among raters was the rigidness of the body. The acceptable technical flaws were also the same as the rejection categories, with the addition of two categories: hand placement and feet placement.

The last area examined was miscalculations in repetition counting, where two causes were identified: a mistake made by the rater in their arithmetic (1 or 2 repetitions), and the rater's willingness to count a repetition that was interrupted mid-execution by the time limit. Interestingly, the occurrence of counting mistakes correlated with a perfect technique in 88% of the cases. When not combined with the counting past the time limit, the raters scored lower numbers than the number of movement cycles actually executed. From the interviews, it can be concluded that assessing non-problematic performance leads to a decrease in focus on the task and often results in forgetting to count. When the non-problematic performance ends mid-execution, the counting past the time limit occurred in 91% of cases, while if the performance requires decision-making for its execution quality flaws, the counting past the time limit was noticed in 28% of cases.

Conclusions

This study investigated push-up tests, which are used by many organizations worldwide and are likely to continue being a popular evaluation tool in the foreseeable future. Several important insights into the assessment of the test were revealed. First and foremost, the overall intra-rater reliability of repetition counting reached only moderate reliability. Even though two of the raters were able to exhibit excellent reliability, the majority presented moderate reliability, indicating the inherently subjective nature of the assessment. These findings suggest that achieving consistently high reliability in repetition counting depends heavily on the competence and motivation of the individual test administrator. Additionally, the study highlighted significant inter-rater differences in counting among specific raters, indicating varying levels of agreement in the assessment process. Moreover, the leniency displayed by male raters in assessing female performances raised concerns about potential gender-based biases, and it may be advised to use a female evaluator for female examinees. The evaluation of technique quality further emphasized the complexity of push-up test assessments. Identifying "perfect technique" proved difficult, with low inter-rater agreement observed. Similarly, the decision-making process during the assessment revealed various reasons for denying repetitions, including issues related to body posture, arm extension, and reaching the "down" position. The existence of these diverse criteria adds further layers of complexity to the assessment process and contributes to the overall subjectivity in the results, since concordance among raters on reaching these criteria varied strongly. One particularly concerning aspect uncovered in the study was the occurrence of counting mistakes made by raters, which were often accompanied by the assignment of a perfect technique mark. This phenomenon suggests that raters may inadvertently overlook counting repetitions when evaluating non-problematic, technically sound performances. Additionally, counting mistakes were observed in cases where the examinee exceeded the time limit for the last repetition, indicating

that raters may not be consistent in applying the time constraint criteria. The implications of these findings are significant for organizations and institutions that heavily rely on push-up tests for evaluations. While the push-up test can be a valuable tool for assessing upper body strength and endurance, it is crucial to recognize its inherent subjectivity and limitations. It is crucial to consider the qualifications and experience of the administrators when interpreting the results of push-up tests, and a deliberate approach is suggested when the use of push-up tests cannot be superseded by any more reliable tests. Future research in this area should focus on exploring ways to enhance the objectivity and reliability of push-up tests, potentially through inspecting the competence of raters, where evaluating the proficiency and accuracy of raters in their assessment process could be a crucial step in improving the overall consistency and validity of the test results. Furthermore, employing advanced technologies, such as computer vision and motion analysis, may lead to more precise and reliable results, ultimately enhancing the utility and credibility of push-up tests in various applications and settings.

References

Adams, M.M., Hatch, S.A., Winsor, E.G., Parmelee, C. (2022) Development of a Standard Push-up Scale for College-Aged Females. International Journal of Exercise Science, 15(4): 820-833.

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. (2009). Objectivity, reliability and validity for a revised push-up test protocol. Measurement in Physical Education and Exercise Science, 6, 4, 225-242.

Barringer, N.D., McKinnnon, C.J., O´Brien, N.C., Kardouni, J.R. (2019) Relationship of strength and conditioning metrics to success on the Army ranger physical assessment test. Journal of Strength and Conditioning Research, 33, 4, 958-964.

Barnett, L., Beurden, E., Morgan, P.J., Lincoln, D., Zask, A., Beard, J. (2009) Interrater objectivity for field-based fundamental motor skill assessment. Research Quaterly for Exercise and Sport, 80, 2, 363-368.

Bianchim, M.S., McNarry, M.A., Evans, R., Thia, L., Barker, A.R., Williams, C.A., Denford, S., Mackintosh, K.A. (2022). Calibration and Cross-validation of Accelerometry in Children and Adolescents with Cystic Fibrosis. Measurement in Physical Education and Exercise Science, 19, 9, 51-55.

Corey, D.T. (1978) The determinants of exploration and neophobia. Neuroscience & Biobehavioral Reviews, 2, 4, 235-253.

Crosby, B.J., Newton, R.U., Galvao, D.A., Taaffe, D.R., Lopez, P., Meniawy, T.M., Khattak, M.A., Lam, W., Gray, E.S., Singh, F. (2023) Feasibility of supervised telehealth exercise for patients with advanced melanoma receiving checkpoint inhibitor therapy. Cancer Medicine, 2023;00:1-13.

Davies, T., Halaki, M., Orr, R., Mitchell, L., Helms, E.R., Clarke, J., Hackett, D.A. (2022) Effect of Set-Structure on Upper-Body Muscular Hypertrophy and Performance in Recreationally-Trained Male and Female. Journal of Strength and Conditioning Research, 36, 8 ,2176-2185.

Disman, M. (2022) How Sociological Knowledge is Produced. Prague: Karolinum. ISBN 9788024650531 [in Czech]

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.

Gorner, K. & Reineke, A. (2020) The influence of endurance and strength training on body composition and physical fitness in female students. Journal of Physical Education and Sport, 20, 3, 2013-2020.

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) Overview of Statistical Methods for Data Processing: Data Analysis and Meta-analysis. 3rd edition. Prague: Portal. ISBN 80-7178-820-1. [in Czech]

Iermakov, S., Olha, I., Khudolii, O. (2021) Strength abilities: assessment of cumulative training effects of strength loads of a series of classes in 8 years old boys. Journal of Physical Education and Sport, 21, 2, 1242-1250.

Kellner, P., Neubauer, J., Polách, M. (2021) Objectivity of push-up test and technique assessment. Journal of Physical Education and Sport, 21, 4, 1629-1634.

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.

Lacy, A.C. & Williams S.M. (2018) Measurement and evaluation in physical education and exercise science. 8th edition. New York: Routledge. ISBN 9781138232341.

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). Normative Decree of the Minister of Defense No. 12/2011. Prague: Ministry of Defence. [in Czech]

Mishenko, N., Kolokoltsev, M., Romanova, E., Alontsev, V., Ustselemov, S., Strashenko, V., Andrianov, A. (2020) Program for improving strength abilities of 16-17-year-old students in the additional physical education system. Journal of Physical Education and Sport, 20, 5, 2796-2802.

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/

Soriano, M.A., Jimenez-Ormeno, E., Amaro-Gahete, F., Haff, G.G., Comfort, P. (2022). How does lower-body and upper-body strength relate to maximum split jerk performance? Journal of Strength and Conditioning Research, 36, 8, 2102-2107.

Tomczak, A & Haponik, M. (2016). Physical fitness and aerobic capacity of Polish military fighter aircraft pilots. Biomedical Human Kinetics, 8, 117-123, 2016.

Wiliams, A., Jacobson, Z., Montgomery, S., Gaddes, R. (2020). Foreign Military Physical Fitness Assessments. Insight Policy Research, Arlington, Virginia. Available at https://insightpolicyresearch.com/wp-content/uploads/2021/03/MV-DOD-Physical-Fitness-Assessments.pdf; accessed March 21, 2023.

Wilson, M. & Engelhard, G.Jr. (2000). Objective Measurement: Theory Into Pracise (Volume 5). New York, Ablex Publishing, ISBN-13 978-1567504330

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.