DOI 10.26773/smj.200614

SportMont

ORIGINAL SCIENTIFIC PAPER

Indonesian Air Force Physical Tester Reliability in Assessing One-Minute Push-Up, Pull-Up, and Sit-Up Tests

Samsul Arifin[1], Heri Retnawati[2] and Himawan Putranta[1]

[1]Yogyakarta State University, Graduate School, Yogyakarta, Indonesia, [2]Yogyakarta State University, Lecturer of Graduate School, Yogyakarta, Indonesia

Abstract

The physical fitness test is a form of assessment to determine the level of physical fitness of a person, both general and specific (muscle). The purpose of this study was to assess the correlation among testers on pull-ups, sit-ups, and push-ups for one minute, and to determine the lowest reliability of the three tests. This study uses a sample of five people who are physical fitness testers of the Indonesian Air Force (TNI AU) who are experienced and active in conducting tests. The subjects were 25 males 18–22 years old. All testers assessed each subject by recording the results of the repetition of movements on all three tests. The data obtained were then converted based on the Indonesian Air Force physical fitness technical guidelines book. After analysis with Anova and ICC, it was found that the data produced by the five different testers had the ICC coefficient values that varied the least on the push-up test. Increased reliability of the testers can be accomplished through practice, tester selection, and paying attention to the ability of the tester. Also, the development of assessment tools and the development of alternative forms of testing are needed.

Keywords: *reliability, tester, push up test, pull up, sit up*

Introduction

Physical fitness has an essential role in supporting one's physical activities so that they can carry out their duties optimally. The degree of physical fitness has a linear relationship with the level of achievement, work success, and other physical activities (Widiyanto & Hartono, 2018). Many institutions require a certain level of fitness, so systems and tools that can measure and assess someone's fitness level ae needed. Harsono (2015) argues that physical fitness components that can be measured and assessed include strength, endurance, muscular power, speed, flexibility, agility, coordination, balance, accuracy, and reaction. In measuring physical fitness, the aspects that must be measured are the basic motor skills, which include strength, endurance, speed, flexibility, and coordination (Bompa & Haff, 2009).

Muscle strength and endurance are essential components of physical fitness (McManis, Baumgartner, & Wuest, 2000). The level of strength and endurance of muscles affects the ability of individuals to perform daily functions and various physical activities. A physical fitness test is needed to produce data about physical abilities, both in monitoring the physical development of coaching and in the context of selection. In the Indonesian Air Force, one-minute model pull-ups, push-ups, and sit-ups are part of a form of physical fitness test conducted to determine the strength and endurance of muscles without using assistive devices (Hartono, Widodo, Wismanadi, & Hikmatyar, 2019). Pull-up and push-up tests are used to assess and develop the strength of the shoulders, arms, and upper body, while sit-up tests are used to measure the strength and endurance of the abdominal muscles (Fox, 1988; Piscopo &

Correspondence:

S. Arifin
Graduate School, Yogyakarta 55281, Yogyakarta State University, Indonesia
E-mail: samsularifin.2018@student.uny.ac.id

Baley, 1981; TNI AU, 2011; TNI AU, 2013).

The results of these three tests are based on the results of the tester, who interprets the pull-ups, sit-ups, and push-ups. Based on the technical guidance of the Indonesian Air Force soldiers' physical safety test, the pull-up movement is done by lifting the body with the strength of the arm so that the chin passes above the bar and then drops off to the starting posture followed by lifting the body; this is repeated as much as possible without resting for a maximum of one minute. In the sit-up test, in the initial stance, participants lie on their backs with their legs bent 90 degrees, their feet flat against the floor and knees approximately 20 cm apart, hands placed behind the head, fingers placed with legs held in place to keep them from moving.

The movement starts with rising and sitting and bending down until the nose touches the right or left knee, and one of the elbows is between the knees; the subjects then quickly goes down, lying on his back as in the starting posture, and then repeats the motion for a maximum of one minute.

In the push-up test, the starting position is with both hands under the shoulders, arms bent at the side of the body, legs straight with toes resting on the floor and the distance between the hands as wide as the body. The subject straightens his arms to lift the body so that it is raised with the legs and body straight; He then bends his arms so that the body lowers; his chest touches the floor while his stomach should not; the head turned to the right or left, and the movement is repeated for one minute.

In connection with a test, according to Miller (2002), some physical ability test requirements are valid, reliable, objective, economical, attractive, and should be implemented. The study of McManis et al. (2000) reveals some problems that are often encountered in the pull-up test when test participants take a down position, and it is difficult to make an assessment. Likewise, in the push-up test, many assessors have difficulty in determining the movements, so the measurement results differ between assessors. According to Barnett et al. (2009), their research showed some problematic motor skills to be assessed, which result in a low-reliability score; they highlighted some obstacles in determining the reliability values in field-based research with direct observation rather than research that uses assistive devices and is more controlled.

In contrast, Baumgartner and Gaunt (2005), in their research on push-up movements, stated that the problem in push-up tests was to determine the position so that the tester could assess accurately. The tester must decide on an assessment of whether the movement of the tester is the correct movement and results in getting a score. The position of the part of the body determines the movements performed, including the number of times the movement can be repeated. This is consistent with what was stated by Cogley et al. (2005) in their research on free movement that different hand positions in push-ups affect the results -ups.

It is crucial that the tester can carry out measurements to produce accurate test data. In tests, errors in measurement are difficult to avoid, so what the tester can do is to anticipate the smallest possible error. The implementation of a mass test with a large number of participants requires many testers to be involved so that there may be no similarities between the measurement and assessment data. Tests involving a large number of testers must pay attention to the agreement between the testers (Putranta & Supahar, 2019). Putranta and Supahar's research (2019), shows that when the total score resulting from inter-assessor measurements and the results of the appraiser's agreement is examined, the scores are almost always not identical.

Kozlowski and Hattrup (1992) define agreement as interrater consensus and reliability as interrater consistency. One way to determine the ability of a tester to take measurements and assessments compared to other testers is called reliability inter-rater. There are many ways to obtain the value of the reliability coefficient inter-rater (ICC), but the basic technique is based on analysis of variance and estimation of various components of variance (Bartko, 1966). The ICC approach is used to assess the consistency of measurements made by several testers on test-takers. Various indices to measure the agreement between several assessors regarding the presence or absence of different measurement results can be interpreted as an intra-class correlation coefficient (Rae, 1984).

Fielitz, Coelho, Horne, and Brechue (2016) found that the coefficient among raters on a two-minute push-up test was small, which is also in line with Mathews' (2013) research on the reliability of rater in pull-up and push-up tests, which states that the results of this experiment illustrate the fact that the ability of the rater to measure physical fitness index is not better when carried out alternately or simultaneously. Also, the learning factor certainly helps to calculate a more valid score, so measurement needs to be preceded by training.

This study aims to obtain the level of reliability of the Indonesian Air Force physical testers among testers one-minute push-up, pull-up, and sit-up tests, and to determine the lowest reliability of three tests. Futhermore, the aims of this study also was used material for correction, training, and guidance in testing in the future.

Methods

Respondents in this study consisted of 25 young male civilians and 18–25-year-old male students who were part of the physical fitness development group at Adi Sucipto Air Force Base, Yogyakarta, Indonesia. As many as five randomly selected testers came from the Air Force Physical Development unit and were experienced and often involved in physical fitness testing. An assessor is an active military member who is male and aged 25-50 years and is still actively involved in physical fitness testing in the Indonesian Air Force.

All procedures for carrying out a pull-up, sit-up, and push-up are guided by technical guidelines for physical fitness tests issued by the Indonesian Air Force Headquarters. Participants carry out pull-up for one minute alternately in the order given by the assessor. During the test, each subject is rated by five testers. The tester only assesses the correct movements performed by the subject for one minute. If the participant stops even though one minute has not expired, the test is considered complete, and the tester records the results obtained. The same procedure is also done on sit-ups and push-ups, with the same subject, but before carrying out the next test, the subject is given sufficient rest time.

The data generated in the form of the results achieved by the subject for one minute of each type of test based on the results of the number of times able to make the correct movements in each test that has been recorded by the tester is then converted to the value of the ability to perform the exercises according to the assessment table contained in the manual for physical fitness test of the Indonesian Air Force on a scale of 0-100. Then the converted value is processed by Anova and

ICC analysis with the SPSS software.

Results

Descriptive data analysis results obtained that the five testers have different ratings on the results of the pull-up, sit-up, and push-up assessment. The results of the assessment by the five testers are in the form of the average value and the complete standard deviation in Table 1.

Table 1. Results of Pull Up, Sit Up and Push Up Tests

	Pull up	Sit up	Push up
	Mean±SD	Mean±SD	Mean±SD
Tester 1	39.40±21.75	79.04±16.23	47.36±15.44
Tester 2	35.04±22.79	69.80±18.07	39.16±27.55
Tester 3	47.76±26.12	81.68±14.97	35.76±19.59
Tester 4	49.04±24.94	79.84±15.86	32.64±25.96
Tester 5	59.00±23.25	86.20±11.99	53.56±22.74

In the pull-up test, the fifth tester has the most substantial average rating with an average value 59.00±23.255. The smallest assessment results, with an average 35.04±22.79, were obtained from the second tester. The results of the sit-up assessment also show almost the same results, namely the five testers have a diversity of test results for which the largest average is obtained from the fifth tester rating with an average of 86.20±11.99 while the smallest assessment with an average of 69.80±18.07 obtained from the second tester. In the push-up test, the most significant average rating is obtained from the fifth tester with an average 53.56±22.74, while the smallest assessment with an average 32.64±25.96 obtained from the fourth tester. From these data, it appears that the fifth tester tends to give a high rating compared to other testers, and the second tester tends to give a low rating. Differences in the results from the four testers above can also be proven through Anova analysis, as presented in the following Table 2.

Table 2. ANOVA Analysis Results for Differences in Assessment

	F critical	F hit	Sig
Pull up		17.407	.000
Sit up	2.87	28.174	.000
Push up		12.239	.000

Table 2 shows that all the results of the assessment of the five testers through three types of tests differ significantly with the calculated F value greater than the F critical and the significance value p=0.0000. In the pull-up test, F count=17.407, the sit-up test F value=28.174 and push-up test F value=12.239 all of which showed a value greater than F critical=2.87. Relating to the level of reliability of the tester in noncritical assessments on a pull-up, sit-up, and push-up tests, the magnitude of the correlation values among testers through the Inter Correlation Class analysis can be seen in Table 3.

Table 3. Correlation Values Inter-Raters (ICC)

	Intra Class Correlation	95% Confidence Interval	
		Lower limit	Upper limit
Pull up	0.782	0.657	0.882
Sit up	0.868	0.782	0.931
Push up	0.706	0.556	0.835

The results of the calculation of correlations among testers in Table 3 use the ICC type of consistency approach, which emphasizes the similarity of ratings between testers. This type of approach is suitable if used to measure abilities that emphasize the differences in each subject and the achievement of predetermined criteria. Table 3 data shows that in the three types of tests, between testers have varying correlation coefficient values: the pull-up test with the ICC coefficient = 0.782 with the correlation range 0.657-0.882, the sit-up test ICC coefficient = 0.868 with the correlation range 0.782-0.931 and the push-up test with the ICC coefficient = 0.706 with a correlation range of 0.556-0.835.

Discussion

Pull up, sit-up, and push-up tests are essential components of physical fitness, especially muscle strength and endurance. The equipment that is used is as simple as a crossbar for pull-up tests while none is needed for the sit-up and push-up tests. Another consideration is that it can be used to test many participants within a limited period, such as tests at military institutions with many test subjects. This component is important for someone who engages in many physical activities, especially muscle strength and endurance, such as athletes and soldiers. According to D'Isanto et al. (2019), the assessments produced through tests serve to define the anthropometric and psychomotor profiles of a person who is used to help determine the goals needed to set a training programme.

Accurately assessing the three tests is difficult because the focus is to obtain as many results as possible with a one-minute repetition of movements. Circumstances with rapid repetition of such movements would certainly make it difficult for the tester to be able to judge carefully and produce accurate data.

The assessment results from several testers appear to vary, including within the same test. Analysis based on the variance values of the above results leads to the conclusion that there are differences in the assessment made by the five testers who have a high significance value with p=0.0000.

While the variation in the value of the inter-rater correlation coefficient shows that the inter-rater correlation value on the push-up test has the smallest value with the value of ICC = 0.687, but the ICC value of pull-up and sit-up tests has values > 0.8. This study also obtained that the range of correlation coefficient values of the five testers in each test has a fairly long range, so the reliability of the tester can be concluded not yet fully adequate. Koo and Li (2015) stated that the ICC coefficient value below 0.50 is bad, between 0.50 and 0.75 in the medium category, between 0.75 and 0.90 the good category and above 0.90 is excellent.

Meanwhile, Artero, España-Romero, and Castro-Piñero (2011) suggested that the ICC between 0.70-0.80 is still questionable or doubtful, and 0.90 is considered high. Thus the reliability between testers on the pull-up, sit up and push up tests needs to be improved. Bajpai, Bajpai, and Chaturvedi (2015) state that it is essential to realize that it is not possible to reach a perfect agreement between testers and that a professional tester and experience are needed to obtain high coefficient values between them. There are many concrete steps to improve the consistency of the assessment by the tester and increase the value of the ICC coefficients of multiple testers, namely through the training of assessors, the selection of appraisers, and the ability to judge. Several studies have been carried out to improve the reliability of testers in carrying out physical tests.

In the study of McCunn et al. (2017), it is stated that filming aids in realizing a better agreement among the judges. A similar study by Mischiati et. Al. (2015) shows that the reliability value between assessors is acceptable. Rogers et al. (2017) demonstrated that ICC scores are better produced by assessors who use video assistance compared to direct measurement.

Further research needs to be done to develop tester tools in assessing pull-up, sit-up and push-up tests, such as the use of infrared motion sensors and cameras. Alternative forms of testing to measure muscle strength and power while still paying attention to practicality in the implementation of tests with a large number of participants are needed. Tester training, the selection or selection of testers involved in the test and paying attention to the ability of the tester are ways of improving the reliability of the tester. It is necessary to develop assistive devices used to pull up, sit up, and push tests to assist the tester in making decisions in assessing the correct movements, especially for tests with a large number of subjects.

Assessments made by multiple testers on a pull-up, sit-up, and push-up tests that rely on humans in their measurements are prone to differences in measurement. This can be seen from the test results data that are varied or significantly different with a significance of P = 0.000. This study also shows the differences in the level of reliability of the tester value of the correlation coefficient among tester (ICC), for which the reliability of the tester on the exercises are all in the medium range; therefore, for use in military institutions, it is necessary to make efforts to improve reliability. In addition to these, other alternatives are needed, such as the use of other forms of testing, as well as the use of assistive devices to facilitate measurement and training for the tester.

Acknowledgements

We are grateful to the Postgraduate Programme Lecturer, Yogyakarta State University and the Head of the Physical Development Sub-Department of the Indonesian Air Force's Health Service for guiding this research.

Conflict of Interest

The authors declare that there is no conflict of interest.

Received: 21 January 2020 | **Accepted:** 27 March 2020 | **Published:** 01 June 2020

References

Artero, E.G., España-Romero, V., & Castro-Piñero, J. (2011). Reliability of field based fitness tests in youth. *International Journal of Sports Medicine, 32*, 159–69.

Bajpai, S., Bajpai, R.C., & Chaturvedi, H.K. (2015). Evaluation of inter-rater agreement and inter-rater reliability for observational data: An overview of concepts and methods. *Journal of the Indian Academy of Applied Psychology 41*(3), 20-27.

Barnett, L., Beurden, E., Morgan, P.J., Lincoln, D., Zask, A., & Beard, J. (2009). Interrater objectivity for field-based fundamental motor skill assessment. *Research Quarterly for Exercise and Sport, 80*(2), 363-368. doi: 10.1080/02701367.2009.10599571

Bartko, J.J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports, 19*, 3-11.

Baumgartner, T.A., & Gaunt, S.J. (2005). Construct related validity for the Baumgartner modified pull-up test. *Measurement in Physical Education and Exercise Science, 9*(1), 51-60. doi: 10.1207/s15327841mpee0901_4

Bompa, T.O., & Haff, G.G. (2009). *Periodization: Theory and methodology of training. [5-th Edition]*. Champaign, IL, USA: Human Kinetics.

Cogley, R.M., Archambault, T.A., Fibeger, J.F., Koverman, M.M., Youdas, J.W., & Hollman, J.H. (2005). Comparison of muscle activation using various hand positions during the push-up exercise. *Journal of Strength and Conditioning Research, 19*(3), 628–633.

D'Isanto, T., D'Elia, F., Raiola, G., & Altavilla, G. (2019). Assessment of sports performance: Theoretical aspects and practical indications. *Sport Mont, 17*(1), 79-82. DOI: 10.26773/smj.190214

Fielitz, L., Coelho, J., Horne, T., & Brechue, W. (2016). Inter-rater reliability and intra-rater reliability of assessing the 2-minute push-up test. *Journal of Military Medicine, 181*(2), 167-175.

Fox, E.L. (1988). *Physiological basis of physical education on athletics*. Philadelpia: Saunders College Pub.

Harsono. (2015). *Sports coaching, theory, and methodology*. Bandung: Remaja Rosdakarya.

Hartono, S., Widodo, A., Wismanadi, H., & Hikmatyar, G. (2019). The effects of roller massage, massage, and ice bath on lactate removal and delayed onset muscle soreness. *Sport Mont, 17*(2), 111–114. DOI: 10.26773/smj.190620

Koo, T.K., & Li, M.Y. (2015). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 5*(3), 190-197.

Kozlowski, S.W., & Hattrup, K. (1992). A disagreement about the within-group agreement: Disentangling issues of consistency versus consensus. *Journal of Applied Psychology, 77*(2), 161–167. DOI:10.1037/0021-9010.77.2.161

Mathews, D.K. (2013). Comparison of testers and subjects in administering physical fitness index tests. *Physical Education and Recreation, 24*(4), 442-445. DOI: 10.1080/10671188.1953.10761987

McCunn, R., der Fünten, K.A, Govus, A., Julian, R., Schimpchen, J., & Meyer, T. (2017). The intra-and inter-rater reliability of the soccer injury movement screen (SIMS). *The International Journal of Sports Physical Therapy, 12*(1), 176-180.

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

Miller, D.K. (2002). *Measurement by the physical educator 4th edition*. San Fransisco: McGraww Hill.

Mischiati, C.R., Comerford, M., Gosford, E., Swart, J., Ewings, S., Botha, N., Stokes, M., & Mottram, S.L. (2015). Intra and inter-rater reliability of screening for movement impairments: Movement control tests from the foundation matrix. *Journal of Sports Science and Medicine, 14*, 427-440.

Piscopo, J., & Baley, J.A. (1981). *Kinesiology the science of movement*. New York: John Wiley & Sons, Inc.

Putranta, H., & Supahar, S. (2019). Development of physics-tier tests (PysTT) to measure students' conceptual understanding and creative thinking skills: A qualitative synthesis. *Journal for the Education of Gifted Young Scientists, 7*(3), 747-775. DOI: 10.17478/jegys.587203

Rae, G. (1984). On measuring agreement among several judges on the presence or absence of a trait. *Educational and Psychological Measurement, 4*, 247-253.

Rogers, D.K., McKeown, I., Parfitt, G., Burgess, D., & Eston, R.G. (2017). Reliability of the athletic ability assessment in Australian rules football. *Journal of Strength and Conditioning Research,* 3(1), 180-187. doi: 10.1519/JSC.0000000000002175.

Indonesian Air Force. (2011). *Regulation of the Air Force Chief of Staff Mighty*. The Air Force's Technical Guidance on Physical Development. Jakarta: Mabesau.

Indonesian Air Force. (2013). Decree of the Chief of Staff of the Air Force. The Air Force's Technical Guidebook on Physical Fitness Tests. Jakarta: Mabesau.

Widiyanto, & Hartono, S. (2018). The effects of hyperbaric oxygen and active recovery on lactate removal and fatigue index. *Sport Mont, 16*(3), 15–18. doi:10.26773/smj.181003