Scientific-Expert Journal of Anthropological Aspects of Sports, Physical Education and Recreation

Submitted: 17/05/2025
Accepted: 20/06/2025
Online first: 02/09/2025
UDK: to be registered

Original article

Test-Retest Reliability of the 20-Metre Shuttle Run Test in a Cohort of Police Trainees

[1,2]*Elisa F.D. Canetti, [1, 2] Ben Schram, [2,3] Rodney Pope, [4] Robert G. Lockie, [5,6]J.Jay Dawes, [2,7,8] Patrick Campbell, & [1,2]Robin Orr

[1]Faculty of Health Sciences and Medicine, Bond University, QLD Australia;
[2]Tactical Research Unit, Bond University, QLD, Australia;
[3]School of Allied Health, Exercise & Sports Sciences, Charles Sturt University, NSW, Australia;
[4]Center for Sport Performance, Department of Kinesiology, California State University - Fullerton, CA, USA;
[5]School of Kinesiology, Applied Health and Recreation, Oklahoma State University, OK, USA;
[6]Tactical Fitness and Nutrition Lab, Oklahoma State University, OK, USA;
[7]School of Behavioural and Health Sciences, Australian Catholic University, Brisbane, QLD 4014, Australia
[8]Sports Performance, Recovery, Injury and New Technologies (SPRINT) Research Centre, Australian Catholic University, Brisbane, QLD, Australia.

*Correspondence: ecanetti@bond.edu.au

Citation: F.D. Canetti, E et al. (2025). Test-Retest Reliability of the 20-Metre Shuttle Run Test in a Cohort of Police Trainees. *Sportlogia. 21*(2), 1-7. https://doi.org/10.63356/spl.2025.005

Abstract

The 20-metre Multistage Fitness Test (20-m MSFT) is commonly used to measure aerobic capacity in police trainees as an entry gateway or exit requirement. However, its test-retest reliability, or consistency of scores for individual candidates across successive days, has not been determined. The aim of this study was to evaluate the test-retest reliability of the 20-m MSFT in police trainees. Retrospective data for 13 police trainees who completed the 20-m MSFT on two occasions 48 hours apart (Trial 1 and Trial 2) were provided. Paired sample t-tests were used to detect differences between individual performances with intraclass correlation coefficients (ICC) investigating the test-retest reliability. A Bland Altman plot was created to inspect the limits of agreement between the two measures. Alpha levels were set at 0.05 whereby a p value of >0.05 indicated no significant difference in mean scores between the two trials. No significant differences (p=0.821) between the mean total numbers of shuttles completed in Trials 1 (mean = 70.4±19.7 shuttles; Level 8-9) and 2 (mean = 69.8 ±21.3 shuttles; Level 8-9) were found. Six trainees achieved higher total shuttle scores for Trial 1 (+8.0 ±3.2 shuttles) while seven trainees achieved higher total shuttle scores for Trial 2 (+5.9± 5.1 shuttles). Test-retest reliability across trials was 'excellent' (ICC(3,1)=0.922 [95%CI 0.766-0.976], mean difference between scores = 0.55±8.37 shuttles). While the 20-m MSFT has excellent test-retest reliability the small amount of variability in results suggests that retesting of candidates who fail to meet any discriminatory standard by a small margin should be considered.

Keywords: Law Enforcement, test consistency, police academy, shuttle run, PSRT, beep test

Introduction

Aerobic fitness is an important physical attribute for police trainees when completing their training academy (Lockie et al., 2019; Orr et al., 2020; Shusko et al., 2017; Tomes et al., 2020). Consequently, aerobic fitness assessments are often conducted during initial applicant screening and selection or during ab initio training (Lockie et al., 2019; Orr et al., 2020; Orr et al., 2022; Shusko et al., 2017; Tomes et al., 2020; Zulfiqar et al., 2021). During these assessments minimum requirements or 'cut scores' may be used to determine an applicant's suitability to train and potential to successfully perform the job tasks required of police officers. If aerobic fitness measures are to be used for hiring and retention purposes, it is important to understand whether scores on those measures are consistent between trials. In other words, how likely is it that the candidate would obtain a similar score if testing was conducted on another day relatively close to the time at which the initial test was performed (i.e., test-retest reliability). Test-retest reliability is of great importance given that potential variation could lead to candidate inclusion or exclusion or trainee pass or fail when on any other occasion the outcome may have been reversed.

Numerous factors may impact the results of physical fitness testing. Hopkins (2000) identified biological variability, and mental (e.g., effort, motivation) and physical (e.g., fatigue) state, as the main sources of within-subject deviations influencing test reliability. Furthermore, weather conditions (e.g., temperature, wind, humidity) (Sproule et al., 1993), environment (e.g., testing surface, noise, other people, etc.,) (Cooper et al., 2005), and hydration status (Lamb & Rogers, 2007) are all factors generally acknowledged to impact aerobic fitness testing results. Most notably for trainees undergoing ab initio training, insufficient recovery from daily stressors and training activities must be considered (Orr et al., 2016), as this has been shown to have a deleterious impact on performance (French & Ronda, 2021).

The 20-metre Multistage Fitness Test (20-m MSFT) is commonly used to assess aerobic fitness within law enforcement populations (Dawes et al., 2019; Lockie et al., 2021; Lockie et al., 2020; Orr et al., 2022; Zulfiqar et al., 2021). Considering this, broader research suggests that the test-retest reliability of the 20-m MSFT is generally 'high' to 'excellent', with reliability coefficients of between 0.87 and 0.98 (Aandstad et al., 2011; Knapik et al., 2004). However, these studies were conducted in military personnel, with no known similar study conducted among law enforcement personnel. This contextualization to the specific population is of importance given the myriad of factors discussed above that could influence assessment results. As such, the aim of this study was to investigate test-retest reliability of the 20-m MSFT in a population of police trainees.

Methods

Experimental approach to the problem

Retrospective data for 13 police trainees were provided by a state police agency. This data included trainee age and 20-m MSFT scores completed on two occasions, 48 hours apart (Trial 1 and Trial 2). All data were collected between 15:00-16:30 on the 28th and 30th November 2022, with each assessment preceded by a self-paced warm up. By ensuring the assessments were conducted at the same time of day, diurnal variations were mitigated. The processes put into place addressed concerns raised by French and Ronda (2021), with limited recovery between tests, lack of warm up protocols and diurnal variations being factors known to impact performance.

Participants

Data for 17 trainees who completed Trial 1 were provided. However, four (n=4) trainees were unable to attend Trial 2, due to either being on duty (n=3) or ill (n=1), leaving data for 13 trainees (males n=7; mean age = 29.6 ± 5.1 years; females n = 6, mean age = 27.7 ± 5.0 years) available for evaluation. The Bond University Human Research Ethics Committee granted ethics approval for this study (BUHREC, Research Protocol BS02086),

with approval for public release of this report provided by the state police agency within which this research took place.

Measurements and Procedures

Prior to attempting the 20-m MSFT, all trainees were briefed by a police Physical Education Officer (PEO) to identify any trainees who might be at risk of suffering from adverse events during exercise. Trainees were given 48 hours between assessments. While police academy training could not be ceased in order to mitigate fatigue, the trainees did not conduct another physical assessment in the intervening period. All activities were conducted outdoors (mean temperature = 19.2 ± 1.8 (range = 18.0-21.2) °C, mean humidity = 67.2 ± 8.1 (range = 57.8-78.4)%, mean wind speed= 7.5 ± 1.1 (range = 6.0-8.6) km/h. Following the safety briefing, trainees were given approximately 10 minutes to warm up and were instructed to include at least five 20-m shuttle runs.

The 20-m MSFT protocols are described in the literature (Dawes et al., 2019; Lockie et al., 2021; Lockie et al., 2020), but for ease will be briefly described here. The 20-m MSFT, also termed the 'progressive shuttle run test', 'beep test', or 'bleep test', has participants run back and forth between two lines spaced 20 meters apart. The speed of running starts at 8.5km/h and increases by 0.5km/h every level, with each level lasting approximately one minute (Dawes et al., 2017). The running speed was standardized by pre-recorded auditory cues (i.e., beeps), played on a handheld iPhone device (Apple Inc., Cupertino, California) connected via Bluetooth to a portable speaker (Ultimate Ears, UE Boom 3, California, US), with participants required to reach the opposing line by the next beep (Dawes et al., 2017). The test was terminated when: a) the PEO informed the trainee that they had failed to reach the lines three times in a row in accordance with the auditory cues, or b) the trainee voluntarily withdrew. The assessment was scored by the police PEO as is common in this police agency. Final scores were presented as level and shuttle (e.g., Level 7 – Shuttle 5, or 'Level 7-5') before being converted to total number of shuttles for the analysis.

Statistical analyses

Data were provided digitally in a Microsoft Excel spreadsheet and prior to analysis were examined for accuracy and cleaned for any errors, with improbable (e.g., Level 1-13 as opposed to Level 11-3) or hard to read results clarified and corrected. Data were then imported into JASP (JASP Team 2023; Version 0.16.4) for analysis. Given that Hopkins (2000) suggests a change in mean scores between trials can be used as a measure of reliability, paired sample t-tests were performed, comparing scores from the two trials. Alpha level was set at 0.05, whereby a p value of >0.05 indicated no significant difference in mean scores between the two trials. Cohen's d was used to assess the effect size indicated by the difference between the two related cohort means (Cohen, 2013), i.e. the standardized mean difference. Interpretation of the effect size (ES) followed the guidelines proposed by Hopkins (2009), whereby ES was considered very small (0.00-0.19), small (0.20-0.59), moderate (0.60-1.19), large (1.20-1.99), very large (2.00-3.99), or extremely large (≥4.00). Raincloud plots were created for visualization of trainee and mean data as well as data distribution using JASP (JASP Team 2023; Version 0.16.4) statistical software. An intraclass correlation coefficient (ICC) was calculated to investigate the test-retest reliability, as previously used in the literature (Aandstad et al., 2011; Cuenca-Garcia et al., 2022). Using JASP (JASP Team 2023; Version 0.16.4), a two-way mixed model was selected. The type of ICC was set to absolute agreement between results from the repeated testing, using single test measurements in each testing episode. The 95% confidence interval (95% CI) was also reported for the ICC. Interpretation of the level of agreement indicated by the ICC was as follows: <0.50, poor; between 0.50 and 0.75, fair, between 0.75 and 0.90 good; above 0.90, excellent (Koo & Li, 2016). A Bland Altman plot was created to inspect the limits of agreement between the two measures (Bland & Altman, 1999). Limits of agreement were calculated as mean difference ± 1.96 standard deviation.

Results

There was no significant difference (t(1,12)=0.232, p=0.821, ES=0.064) between total numbers of shuttles completed in Trial 1 (mean = 70.4 ± 19.7 shuttles; Level 8-9) and Trial 2 (mean = 69.8 ± 21.3 shuttles; Level 8-9). Six of the 13 trainees achieved a higher total shuttle score for Trial 1, completing on average 8.0 (+3.2) shuttles less in Trial 2. The other seven trainees achieved a higher total shuttle score for Trial 2, completing on average 5.9 (+ 5.1) shuttles more than in Trial 1. Trainee and mean data are shown in Figure 1.

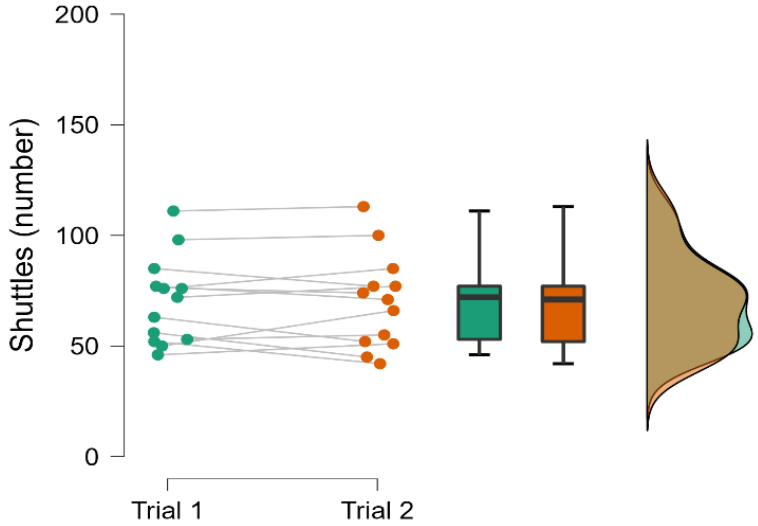

Figure 1. Depictions of individual trainee performance levels (scatter plot) and mean performance levels (box plots) as well as data distribution via a raincloud plot. Green markers denote Trial 1 results and orange markers, Trial 2. These plots show that the total shuttles performed in Trial 1 were similar to Trial 2.

Test-retest reliability of the 20-m MSFT for trial 1 and 2 was 'excellent', with an ICC(3,1)=0.922 [95%CI 0.766-0.976]. The mean difference in scores between trials was 0.55 [95%CI -4.52 to 5.60; SD 8.37] shuttles (Figure 2). The upper limit of agreement (LoA) was 16.95 [95%CI 8.19 to 25.71] shuttles and the lower LoA was -15.87 [95%CI -24.64 to -7.11] shuttles, with these LoA indicative of the 'typical error' in scores.

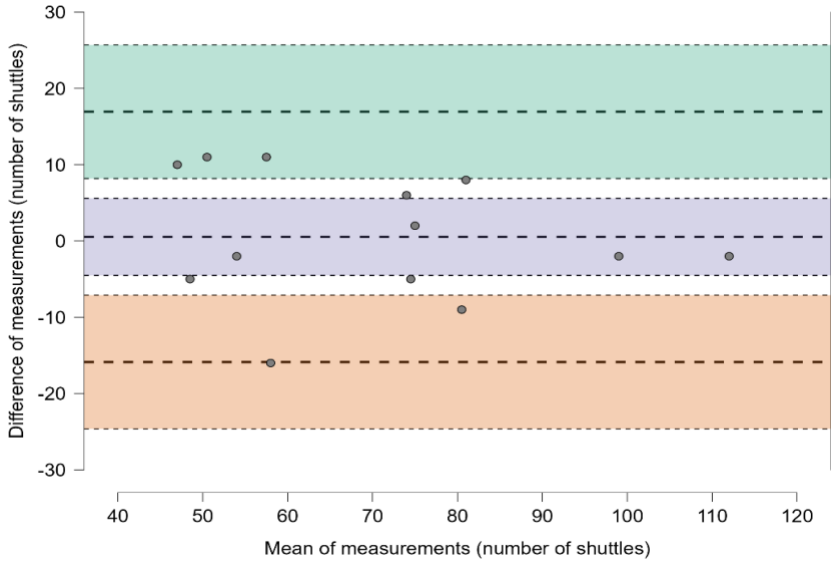

Figure 2. Bland-Altman plot indicating limits of agreement between the two trials. Figure shows 95% confidence intervals for mean (purple shaded area) and limits of agreement (upper in green and lower in orange shaded area).

Discussion

The aim of this study was to investigate the test-retest reliability of the 20-m MSFT in a population of police trainees. The results of the study showed no statistically significant difference between mean scores from two trials, indicating strong test-retest reliability. The ICC (reliability coefficient) of 0.922, which is considered to indicate an 'excellent' level of agreement between paired scores at the two testing timepoints (Koo & Li, 2016), supports such finding.

The test-retest reliability for the 20-m MSFT identified in the current study was consistent to the reliability reported in the literature. Recent systematic review by Cuenca-Garcia et al. (2022), concluded that the 20-m MSFT showed strong test re-test reliability, with seven out of nine high-quality studies reporting ICCs between 0.93–0.96 and correlation coefficients between 0.85–0.96. Similar strong reliability of the 20-MSFT was demonstrated in studies involving tactical populations. A US military report by Knapik et al (2004) documented test-retest reliability coefficients ranging from 0.87 to 0.98, which were higher than those for the 1-mile (1.6 km) and 2-mile (3.2 km) run tests (0.82–0.92). Similarly, Aandstad et al.,(2011) described a reliability coefficient of 0.95 for the 20-m MSFT in a study involving US soldiers.

While the mean difference in scores between trials was small (0.55 shuttles), the standard deviation of 8.4 shuttles indicate variability across trainees. This variability may be partly due to the small sample size or the short interval between trials (48 hours). Similar studies assessing test–retest reliability of the 20-m MSFT have reported intervals ranging from 2-4 days (Aandstad et al., 2011; Metsios et al., 2008) to 1-4 weeks (Cooper et al., 2005; Kim et al., 2011; Lamb & Rogers, 2007; Sproule et al., 1993), yet consistently reporting high reliability. In the present study, the LoAs, based on the SD, were approximately 16 shuttles. Despite the small sample size, these results are comparable to studies with larger sample sizes: Lamb and Rogers (2007) reported LoAs of 18 shuttles when assessing 35 university students, while Aandstad et al (2011) reported LoAs of ~10 shuttles for 41 Home Guard soldiers.

Although the 20-m MSFT demonstrated high group-level reliability, individual differences were evident, with six trainees performing better in Trial 1 and seven in Trial 2. Several factors may explain this variability. Familiarization with the pacing and auditory cues of the MSFT can improve performance on a second attempt, and prior studies have shown that one practice session is often sufficient to alter performance (Ramsbottom et al., 1988; Stickland et al., 2003). Conversely, poorer performance in Trial 2 may have been due to residual fatigue or incomplete recovery, as other studies have reported longer rest times between trials (Cooper et al., 2005; Kim et al., 2011; Lamb & Rogers, 2007; Sproule et al., 1993). Motivation and psychological factors may also have contributed to the variability observed as some trainees may have been more motivated during the first attempt, whereas others may have been more driven in the second trial by the opportunity to surpass their prior score (Neto et al., 2015) . Finally, mental stress has been shown to impair endurance performance by increasing perceived exertion and reducing pacing efficiency (Van Cutsem et al., 2017) , which could have contributed to lower shuttle scores in some individuals. Collectively, these findings highlight that while the 20-m MSFT is reliable at the group level, substantial within-subject variability may occur, and may be more prominent in studies with small sample size.

Several limitations of the present study should be acknowledged. First, the small sample size may have contributed to the wide 95% confidence interval (0.766–0.975) and amplified the influence of individual performance variability, as reflected by the standard deviation of 8.4 shuttles. Second, police academy training could not be suspended during the study period, meaning trainees were exposed to occupational stressors and cognitively demanding training activities (i.e., not physical), which are known to negatively affect 20-m MSFT performance and recovery (Macmahon et al., 2019; Slimani et al., 2018). Finally, although factors such as diurnal variation, physical fatigue, and rater consistency were controlled, these external

stressors may have introduced additional variability in participants' physical and mental states between trials.

Conclusion

The findings of this study suggest that while there will be some variability in performance given a variety of individual factors (e.g., fatigue, etc.), the 20-m MSFT has excellent test-retest reliability. However, there were some small variations in scores and the associated between-trial variation could have implications for gateway testing. Thus, if the 20-m MSFT was to be used as a decision-making tool for training entry or completion, retesting of candidates or trainees who fail to meet any discriminatory standard by a small number of shuttles could be considered, with repeat testing within a few days feasible as the candidate or trainee could have the capacity to pass the standard on a different day.

Practical Implications

The test-retest reliability of the 20-m MSFT across two trials spaced 48 hours apart was 'excellent', with LoAs of approximately ±16 shuttles. The findings of this study demonstrate that the 20-m MSFT is generally test-retest reliable in this population. However, given there was some variability in individual results, re-testing of candidates who fail to meet any discriminatory standard by a small margin can be considered within a short timeframe (e.g., 48 hours), if allowed full rest, to ensure application of test results in decision-making is fair.

Acknowledgements: The authors acknowledge no conflicts of interest. This study was funded by the New Zealand Police Force and the authors would like to thank and acknowledge those personnel who volunteered to take part in this study.

References

Aandstad, A., Holme, I., Berntsen, S., & Anderssen, S. (2011). Validity and reliability of the 20 meter shuttle run test in military personnel. Mil Med, 176(5), 513-518. https://doi.org/10.7205/MILMED-D-10-00373

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. Stat Methods Med Res, 8(2), 135-160.

Cohen, J. (2013). Statistical power analysis for the behavioral sciences. Routledge.

Cooper, S. M., Baker, J. S., Tong, R. J., Roberts, E., & Hanford, M. (2005). The repeatability and criterion related validity of the 20 m multistage fitness test as a predictor of maximal oxygen uptake in active young men. Br J Soc Med, 39(4), e19-e19. https://bjsm.bmj.com/content/bjsports/39/4/e19.full.pdf

Cuenca-Garcia, M., Marin-Jimenez, N., Perez-Bey, A., Sanchez-Oliva, D., Camiletti-Moiron, D., Alvarez-Gallardo, I. C., Ortega, F. B., & Castro-Pinero, J. (2022). Reliability of field-based fitness tests in adults: a systematic review. Sports Med, 52(8), 1961-1979. https://link.springer.com/article/10.1007/s40279-021-01635-2

Dawes, J. J., Lockie, R. G., Orr, R. M., Kornhauser, C., & Holmes, R. J. (2019). Initial fitness testing scores as a predictor of police academy graduation. J Aust Strength Cond, 27(4), 30-37.

Dawes, J. J., Orr, R. M., Flores, R. R., Lockie, R. G., Kornhauser, C., & Holmes, R. (2017). A physical fitness profile of state highway patrol officers by gender and age. Annals of Occupational and Environmental Medicine, 29(16), 16. https://doi.org/10.1186/s40557-017-0173-0

French, D., & Ronda, L. T. (2021). NSCA's Essentials of Sport Science. Human Kinetics.

Hopkins, W. G. (2000). Measures of reliability in sports medicine and science. Sports Med, 30, 1-15.

Hopkins, W. G., Marshall, S. W., Batterham, A. M., & Hanin, J. (2009). Progressive statistics for studies in sports medicine and exercise science. Med Sci Sports Exerc, 41(1), 3-13. https://doi.org/10.1249/MSS.0b013e31818cb278

Kim, J., Jung, S., & Cho, H.-C. (2011). Validity and reliability of shuttle-run test in Korean adults. Int J Sports Med, 32(08), 580-585.

Knapik, J. J., Jones, B. H., Sharp, M. A., Darakjy, S., & Jones, S. (2004). The case for pre-enlistment physical fitness testing: Research and recommendations.

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

Lamb, K. L., & Rogers, L. (2007). A re-appraisal of the reliability of the 20 m multi-stage shuttle run test. Eur J Appl Physiol, 100(3), 287-292. https://link.springer.com/article/10.1007/s00421-007-0432-9

Lockie, R., Balfany, K., Bloodgood, A. M., Moreno, M. R., Cesario, K. A., Dulla, J. M., Dawes, J. J., & Orr, R. M. (2019). The influence of physical fitness on reasons for academy separation in law enforcement recruits. Int J Env Res Pub He, 16(3), 372. https://mdpi-res.com/d_attachment/ijerph/ijerph-16-00372/article_deploy/ijerph-16-00372.pdf?version=1548735319

Lockie, R., Dawes, J. J., Moreno, M. R., Cesario, K. A., Balfany, K., Stierli, M., Dulla, J. M., & Orr, R. M. (2021). Relationship Between the 20-m Multistage Fitness Test and 2.4-km Run in Law Enforcement Recruits. J Strength Cond, 35(10), 2756-2761. https://doi.org/10.1519/JSC.0000000000003217

Lockie, R., Ruvalcaba, T. R., Stierli, M., Dulla, J. M., Dawes, J. J., & Orr, R. M. (2020). Waist circumference and waist-to-hip ratio in law enforcement agency recruits: Relationship to performance in physical fitness tests. J Strength Cond, 34(6), 1666-1675. https://doi.org/10.1519/jsc.0000000000002825

Macmahon, C., Hawkins, Z., & Schuecker, L. (2019). Beep test performance is influenced by 30 minutes of cognitive work. Med Sci Sports Exerc, 51(9), 1928.

Metsios, G. S., Flouris, A. D., Koutedakis, Y., & Nevill, A. (2008). Criterion-related validity and test–retest reliability of the 20 m square shuttle test. J Sci Med Sport, 11(2), 214-217.

Neto, J. M. D., Silva, F. B., De Oliveira, A. L. B., Couto, N. L., Dantas, E. H. M., & de Luca Nascimento, M. A. (2015). Effects of verbal encouragement on performance of the multistage 20 m shuttle run. Acta Scientiarum. Health Sciences, 37(1), 25.

Orr, R., Ferguson, D., Schram, B., Dawes, J. J., Lockie, R., & Pope, R. (2020). The relationship between aerobic test performance and injuries in police recruits. Int J Exerc Sci, 13(4), 1052. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7449329/pdf/ijes-13-4-1052.pdf

Orr, R., Knapik, J. J., & Pope, R. (2016). Avoiding program-induced cumulative overload (PICO). J Spec Oper Med : Peer Rev J SOF Med Prof, 16(2), 91-95.

Orr, R. M., Lockie, R., Milligan, G., Lim, C., & Dawes, J. (2022). Use of physical fitness assessments in tactical populations. Strength Cond J, 44(2), 106-113.

Ramsbottom, R., Brewer, J., & Williams, C. (1988). A progressive shuttle run test to estimate maximal oxygen uptake. Br J Sports Med, 22(4), 141-144.

Shusko, M., Benedetti, L., Korre, M., Eshleman, E. J., Farioli, A., Christophi, C. A., & Kales, S. N. (2017). Recruit Fitness as a Predictor of Police Academy Graduation. Occup Med (Lond), 67(7), 555-561. https://doi.org/10.1093/occmed/kqx127

Slimani, M., Znazen, H., Bragazzi, N. L., Zguira, M. S., & Tod, D. (2018). The effect of mental fatigue on cognitive and aerobic performance in adolescent active endurance athletes: insights from a randomized counterbalanced, cross-over trial. Journal of clinical medicine, 7(12), 510.

Sproule, J., Kunalan, C., McNeill, M., & Wright, H. (1993). Validity of 20-MST for predicting VO2max of adult Singaporean athletes. Br J Soc Med, 27(3), 202-204. https://bjsm.bmj.com/content/bjsports/27/3/202.full.pdf

Stickland, M. K., Petersen, S. R., & Bouffard, M. (2003). Prediction of maximal aerobic power from the 20-m multi-stage shuttle run test. Can J Appl Physiol, 28(2), 272-282.

Tomes, C. D., Sawyer, S., Orr, R., & Schram, B. (2020). Ability of fitness testing to predict injury risk during initial tactical training: a systematic review and meta-analysis. Injury Prev, 26(1), 67-81. https://injuryprevention.bmj.com/content/26/1/67.long

Van Cutsem, J., Marcora, S., De Pauw, K., Bailey, S., Meeusen, R., & Roelands, B. (2017). The effects of mental fatigue on physical performance: a systematic review. Sports Med, 47(8), 1569-1588.

Zulfiqar, M. M., Wooland, J., Schram, B., Dawes, J. J., Lockie, R., & Orr, R. (2021). Battery fitness testing in law enforcement: A critical review of the literature. Int J Exerc Sci, 14(4), 613.