European Journal of Human Movement

OPEN ACCESS

Technical note

Between-week reliability of motion tracking screening: A preliminary study with youth male football players

Mihkel M. Laas [1,2]*, Matthew D. Wright [1], Shaun J. McLaren [3,4], Matthew D. Portas [1,5], Guy Parkin [2] and Daniel L. Eaves [1]

[1] School of Health and Life Sciences, Teesside University, Middlesbrough, UK
[2] Pro Sport Support Ltd, Huddersfield, UK
[3] Newcastle Falcons Rugby Club, Newcastle upon Tyne, UK
[4] Department of Sport and Exercise Sciences, Durham University, Durham, UK
[5] Technical Directorate, The Football Association, St. George's Park, Burton upon Trent, UK

* Correspondence: (MML) MM.Laas@tees.ac.uk ORCID ID 0000-0002-2401-1092

Received: 20/02/2021; Accepted: 31/05/2021; Published: 30/06/2021

Abstract: We investigated the reliability of fundamental movements in thirteen youth football players (mean age = 16.8 ± 0.6 y). Following a habituation warm-up, players performed three trials of stride-for-distance and bodyweight squats between two weeks. A motion tracking device was used to measure stride distance and squat depth. The weekly mean changes in mean and maximum performance were moderate for the stride (2.8%; 90% confidence interval: 1.1 to 4.5 and 3.6%; 2.1 to 5.2, respectively) and small for the squat (-2.7%; -12.3 to 7.9 and 3.3%; -5.2 to 12.6). ICCs for stride mean and maximum performance were moderate (0.74; 0.43 to 0.90) and high (0.76; 0.46 to 0.90), respectively, and low for the squat (0.22; -0.27 to 0.61 and 0.42; -0.04 to 0.74, respectively). Typical errors for mean and maximum performance were moderate for the stride (2.4%; 1.8 to 3.6 and 2.1%; 1.6 to 3.3, respectively) and large for the squat (15.9%; 11.8 to 25.1 and 13.1%; 9.7 to 20.5, respectively). The motion tracking reliability was encouraging in the stride. This finding warrants further investigation and consideration of the stride test for use in applied practice with a group of youth footballers.

Keywords: movement screening, fundamental movement skills, biomechanics, test retest reliability, motion tracking

1. Introduction

Fundamental movements are a crucial component in the athletic development of adolescents (Lloyd & Oliver, 2012). They have been previously assessed in boys` football academies using the Functional Movement Screen (FMS™) (Portas et al., 2016; Newton et al., 2017). A number of limitations to the FMS™ have been acknowledged, including the subjective and categorical scoring system which reduces the likelihood of determining minimum practically important changes in performance (Wright & Chesterton, 2019). To address the known issues in categorical scale movement assessment tools, a novel

testing system (AMAT Performance) was developed.

The AMAT system has demonstrated good criterion validity for dynamic athletic movements against manual measurements (Wijnbergen, 2019). The system has also produced high static reliability using three-dimensional video and skeletal tracking algorithms (Wijnbergen, 2019). However, large within-session variation in performance of fundamental movements has been observed in academy footballers (Laas et al., 2020). Laas et al., (2020) recommended employing habituation in future studies to reduce the variability in the players` movement performance scores. Short-term, between-session reliability of test performance, which is a necessity for monitoring changes in individual players (Hopkins, 2000; Hurst et al., 2018), has not previously been investigated using AMAT.

Reliability refers to the repeatability or reproducibility of a measure (Hopkins, 2000). Common methods of testing reliability are based on measures of relative and absolute reliability (Atkinson & Nevill, 1998). Physical performance tests are associated with measurement errors comprising of two primary sources, systematic bias and random error (Atkinson & Nevill, 1998). Quantifying measurement error can enable practitioner's to make better informed decisions about observed changes in an individual`s performance, upon which future training recommendations can be based. Practitioners could therefore use short-term (between-session) reliability testing data as a method for identifying the athletes` individual performance changes. In the current study, we aimed to establish the between-week reliability of two fundamental movements (stride and back squat) and to demonstrate a method of using reliability data to report between-week individual changes in the movement scores using motion tracking in youth male footballers.

2. Materials and Methods

Subjects — Thirteen under-18 football players (age: 16.8 ± 0.6 y; stature: 180.5 ± 4.2 cm; body mass: 69.9 ± 8.1 kg) from a category three club within the English Elite Player Performance Plan (EPPP) system participated in this test-retest study. All players were injury-free and medically cleared to participate in training by the club`s medical staff. Ethical release was obtained from Teesside University's Research Ethics committee to use anonymised data provided by Pro Sport Support Ltd. The study was conducted in accordance with the declaration of Helsinki.

Methodology — After an equipment-free habituation warm-up of three trials, the players completed three trials of the stride-for-distance (left to right and right to left leg) and the back squat test on the AMAT system (Figure 1) during the 2017/18 pre-season for two consecutive weeks. The testing took place on Monday afternoons during a gym session where the players were split into two groups (first and second year scholars). One group engaged in an active recovery session while they were individually asked to undertake the movement testing, the other group undertook a light strength training session. Afterwards the two groups swapped activities. The order of activities for the groups was switched the following week.

Players wore their normal indoor training footwear and reflective markers were attached to the middle of their shoelaces and above the patella, as per the manufacturer's user guidelines (Figure 1). The tests started and finished with an auditory cue from the testing system, which was explained to the players beforehand. The maximum movement performance outcome scores were tracked by the 30 Hz depth sensor camera (Kinect™ V2, Microsoft, USA) (Figure 1). The stride performance was measured as the front position of the landing foot from the start position (mm, measured in the anterior-posterior axis). The squat performance (depth) was measured by tracking the squat

saddle marker, as the marker`s maximum displacement of the movement from the start to the bottom position (mm, measured in the superior-inferior axis).

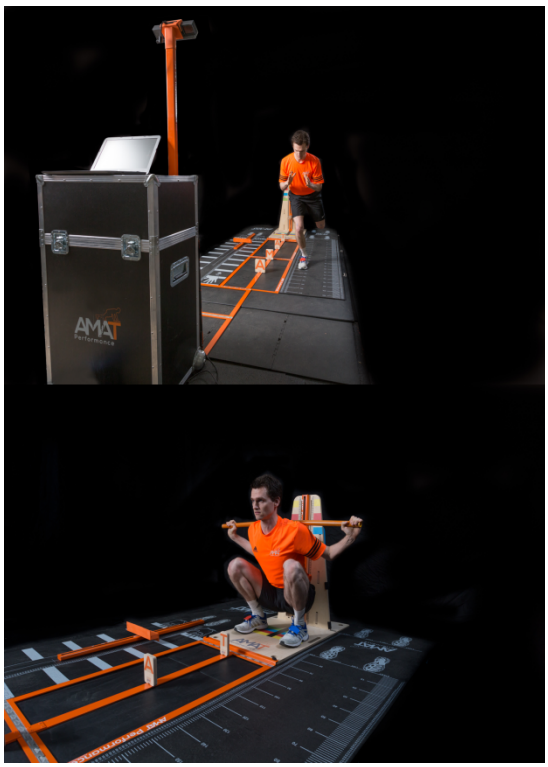

Figure 1. The stride (above) and squat (below) testing on the AMAT system. In the stride test, the players pushed off with one leg, hopped as far as possible and landed on the opposite leg (left to right leg and right to left leg). In the squat, the players were instructed to squat as deep as possible on the squat saddle with control.

Statistical Analysis — The mean and the maximum movement scores from the three trials were assessed for reliability. In addition, the stride of both sides (left to right and right to left leg) were combined to generate a composite performance indicator. All data were log transformed and subsequently back transformed to obtain a percentage change for movement scores during both weeks of testing. Paired samples t-tests were used to evaluate the systematic bias as the mean differences between weeks. Random error was assessed using the between-week typical error (i.e. within-player variation), which was estimated as the standard deviation of change scores divided by the square root of 2 (Hopkins, 2000). Random error was also assessed using the intraclass correlation coefficient (ICC; i.e. reproducibility of the rank order of players) and calculated from a 2-way mixed-effects model ($ICC_{3,1}$) (Shrout & Fleiss, 1979). Uncertainty in all estimates were expressed as 90% confidence intervals (CI).

The magnitude of the between-week changes was interpreted using standardised thresholds for trivial, small, moderate, large, very large differences, calculated as <0.2, 0.2, 0.6, 1.2 and 2.0 of the pooled between-player standard deviations (Hopkins et al., 2009). To assess the magnitude of typical error, these thresholds were halved (Smith & Hopkins, 2011). The qualitative inference for the $ICC_{3,1}$ was based on the following thresholds: >0.99, extremely high; 0.99–0.90, very high; 0.90–0.75, high; 0.75–0.50, moderate; 0.50–0.20, low; <0.20, very low (Malcata et al., 2014).

To establish reference values for the likely range of a "true" between-week change in stride and squat performance (i.e., free from noise), we calculated 80%, 90%, 95% CI as the product of the inverse Student's t-distribution (two-tailed) and the standard deviation of test-retest change scores. An individual change was determined as both "true" and substantial when the test 2–1 difference was greater than or equal to the CI plus the typical error and the threshold for a small effect (i.e., 0.2 between-player SDs) (Hopkins, 2004).

3. Results

The weekly mean change in mean and maximum performance scores were moderate for the stride and small for squats (Table 1). This indicates that the potential between-week learning effect was relatively low for both tests. The ICCs for stride mean and maximum performance were moderate and high respectively, low for the squats (Table 1). Thus, there was moderate to high reproducibility of the rank order through both testing sessions in the stride, but not the squat. Typical errors for mean and maximum performance were moderate for the stride and large for the squat (Table 1).

These represent the random between-week, within-player variability. The weekly individual "true" changes were lower (Table 2) and more frequent (Figure 2) for the stride.

Table 1. Between-week reliability for the stride and squat mean and maximum scores (90% CI in brackets).

Mean score	Stride	Squat
Mean change		
Percent (%)	2.8 (1.1 to 4.5)	-2.7 (-12.3 to 7.9)
cm	6.1 (2.4 to 9.7)	-1.2 (-5.6 to 3.2)
Qualitative Inference	Moderate	Small
Typical error		
Percent (%)	2.4 (1.8 to 3.6)	15.9 (11.8 to 25.1)
cm	5.2 (4.0 to 7.9)	6.3 (4.7 to 9.5)
Qualitative Inference	Moderate	Large
ICC	0.74 (0.43 to 0.90)	0.22 (-0.27 to 0.61)
Qualitative Inference	Moderate	Low
Maximum score	**Stride**	**Squat**
Mean change		
Percent (%)	3.6 (2.1 to 5.2)	3.3 (-5.2 to 12.6)
cm	8.3 (4.8 to 11.8)	1.4 (-2.6 to 5.5)
Qualitative Inference	Moderate	Small
Typical error		
Percent (%)	2.1 (1.6 to 3.3)	13.1 (9.7 to 20.5)
cm	5.0 (3.8 to 7.6)	5.8 (4.4 to 8.8)
Qualitative Inference	Moderate	Large
ICC	0.76 (0.46 to 0.90)	0.42 (-0.04 to 0.74)
Qualitative Inference	High	Low

ICC = Intraclass correlation coefficient; CI = confidence interval

Table 2. Between-week "true" changes in the stride and squat at three confidence limits (80%, 90%, 95%).

		Confidence limits for an individual change		
Performance measure		**± 80%**	**± 90%**	**± 95%**
Stride (%)	Mean	4.6	6.0	7.4
	Maximum	4.0	5.3	6.5
Stride (cm)	Mean	10.0	13.1	16.0
	Maximum	9.6	12.6	15.4
Squat (%)	Mean	30.5	40.1	49.0
	Maximum	25.1	33.0	40.4
Squat (cm)	Mean	12.1	15.9	19.4
	Maximum	11.1	14.6	17.9

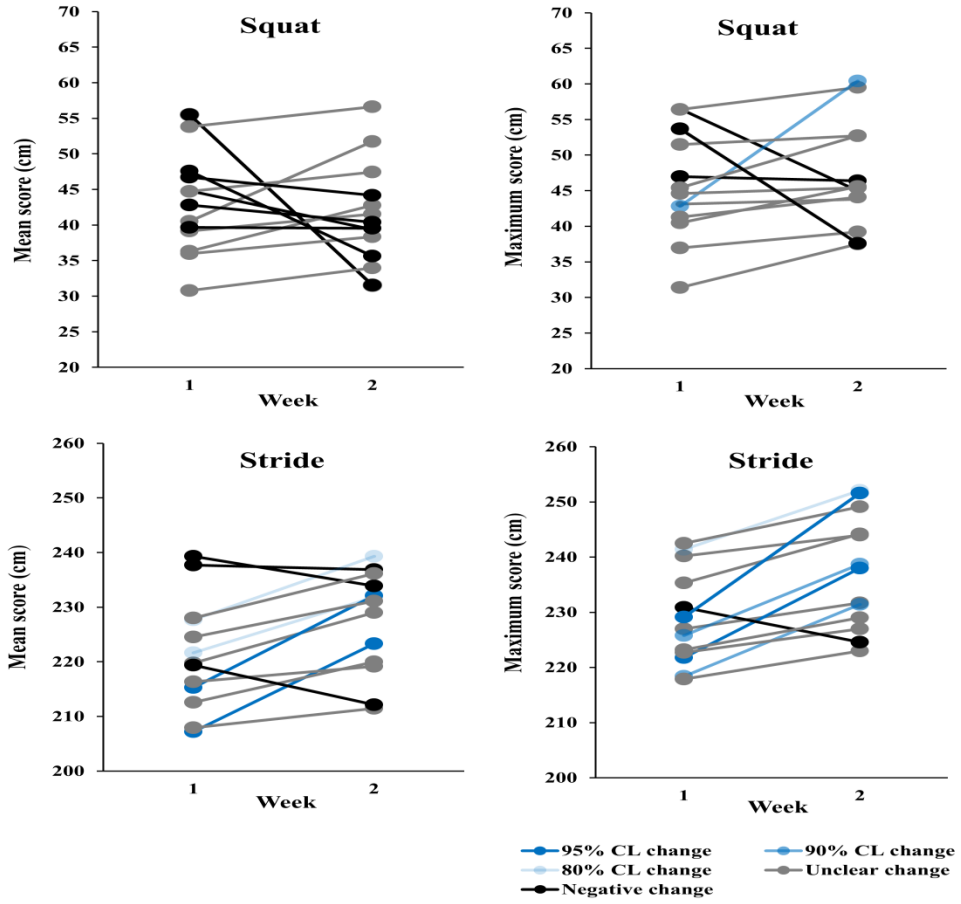

Figure 2. Individual player mean and maximum performance scores of three trials (cm) between week 1 and 2 for the squat and stride. "True" changes were defined when the weekly score difference was greater or equal to the confidence interval plus the typical error and threshold of the smallest effect (0.2 between-player SDs). The "true" changes were observed at different levels of confidence limits using the transparency in blue horizontal lines.

4. Discussion

We aimed to establish the between-week reliability and report individual changes of two fundamental movement tests using an objective motion tracking system (Wijnbergen, 2019; Laas et al., 2020) in a group of youth male footballers. The main finding of our study was that the stride showed high reliability (ICC, 0.76), yet this was not evident for the squat.

We reported moderate and small mean systematic change in the stride and squat scores (Table 1). The point estimate ICC (relative reliability) in the maximum stride score showed high reliability (0.76). The width of the confidence interval ranged from low to very high reliability (0.46 to 0.90; Table 1), which was influenced by a small sample size and serves as a limitation in this study. We also reported moderate typical errors in the stride (2.1 to 2.4%; absolute reliability, Table 1), which were less than previous research in hopping (3.5%) (Reid et al., 2007). The relative and absolute reliability for the squat were low (Table 1). The squat is dependent on several upper body, lower body and movement mechanics factors (Myer et al., 2014) where inconsistency in any of them during different trials could have impacted the reliability in the measured depth in our study. The reliability was therefore reassuring for the stride. The variation we observed in this study was likely due to biological factors, given the high reliability of the technology we employed to measure performance (Wijnbergen, 2019).

We calculated several confidence limits with the degrees of freedom of our sample size and found more individual "true" changes (i.e., noise free) with the stride than squat test (Figure 2). The "true" changes in the stride ranged from ~4 to ~7% (Table 2). These were lower than have been

previously reported in hop tests (8.1%) (Reid et al., 2007). With this method we demonstrated how reliability data can be incorporated within athlete monitoring to assist practitioners in making informed decisions about individual changes in performance.

We had some limitations with this study. Firstly, we could not conduct player habituation on the equipment due to the time-restricted testing session we were afforded. However, to help overcome this issue the same movements and standardised coaching cues were used during the habituation warm-up as in the testing. Secondly, this study of thirteen participants was not sufficiently powered (Hopkins, 2000) to make substantive claims about populations other than the one tested in this study, although it has provided some valuable preliminary insights. An increased sample size from multiple clubs would have been preferable to increase the statistical power of the study and improve the generalisability of these results for practitioners.

5. Practical Applications.

Practitioners should look to assess reliability in their environment and can follow the methods outlined in this study. To increase the reliability of the stride with a group of players in future studies, practitioners are recommended to incorporate habituation ideally on the testing equipment (Hurst et al., 2018) and ensure sufficient recovery between trials and tests (Reid et al., 2007). Following the process highlighted in this study, practitioners can track and determine the players` "true" changes of performance to the degree of chosen confidence.

6. Conclusions

The AMAT system showed highly reproducible performance for the stride and the reliability of the squat (i.e., depth) was low. We also presented a method for using reliability data to track individual changes in test performance. The high reliability (ICC, 0.76) of the stride test using the motion tracking system serves as preliminary data, and we recommend a larger sample size of approximately 50 participants (Hopkins, 2000) in future, for conducting a definitive trial.

Funding: The project received government funding from a Knowledge Transfer Partnership (Innovate UK) to Pro Sport Support Ltd and Teesside University (KTP 009965. Project title: To develop a specialist technology enhanced Adolescent Movement Analysis Tool and associated training intervention curriculum, exergaming and CPD offers underpinned by leading biomechanical research to improve the physicality of elite youth athletes).

Acknowledgments: We would like to express our gratitude to Professor Alan M. Batterham for his invaluable statistical advice throughout the completion of this research. The authors acknowledge the help of Pro Sport Support Ltd for the provision of the anonymised data, which was used for this study.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

Atkinson, G., Nevill, A. M. (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports Medicine, 26(4), 217–238. doi:10.2165/00007256-199826040-00002

Hopkins, W. G. (2000). Measures of reliability in sports medicine and science. *Sports Medicine*, 30(1), 1–15. doi:10.2165/00007256-200030010-00001

Hopkins, W. G. (2004). How to interpret changes in an athletic performance test. *Sportscience*, 8, 1–7.

Hopkins, W. G., Marshall, S. W., Batterham, A. M., Hanin, J. (2009). Progressive statistics for studies in sports medicine and exercise science. *Medicine and Science in Sports and Exercise*, 41(1), 3–13. doi:10.1249/mss.0b013e31818cb278

Hurst, C., Batterham, A. M., Weston, K.L., Weston, M. (2018). Short- and long-term reliability of leg extensor power measurement in middle-aged and older

adults. *Journal of Sports Sciences*, 36(9), 970–977. doi:10.1080/02640414.2017.1346820

Laas, M. M., Wright, M. D., McLaren, S. J., Eaves, D. L., Parkin, G., Portas, M. D. (2020). Motion tracking in young male football players: a preliminary study of within-session movement reliability. *Science and Medicine in Football*, 4(3), 203–210. doi:10.1080/24733938.2020.1737329

Lloyd, R. S., Oliver, J. L. (2012). The Youth Physical Development Model: A New Approach to Long-Term Athletic Development. *Strength and Conditioning Journal*, 34(3), 61–72. doi:10.1519/SSC.0b013e31825760ea

Malcata, R. M., Vandenbogaerde, T. J., Hopkins, W. G. (2014). Using athletes' World rankings to assess countries' Performance. *International Journal of Sports Physiology and Performance*, 9(1), 133–138. doi:10.1123/ijspp.2013-0014

Myer, G. D., Kushner, A. M., Jensen, B. L., Schoenfeld, B. J., Hugentobler, J., Lloyd, R. S., Vermeil, A., Chu, D. A., Harbin, J., McGill, S. M. (2014). The Back Squat: A Proposed Assessment of Functional Deficits and Technical Factors That Limit Performance. *Strength and Conditioning Journal*, 36(6), 4–27. doi:10.1519/SSC.0000000000000103

Newton, F., McCall, A., Ryan, D., Blackburne, C., aus der Fünten, K., Meyer, T., Lewin, C., McCunn, R. (2017). Functional Movement Screen (FMS™) score does not predict injury in English Premier League youth academy football players. *Science and Medicine in Football*, 1(2), 102–106. doi:10.1080/24733938.2017.1283436

Portas, M. D., Parkin, G., Roberts, J., Batterham, A. M. (2016). Maturational effect on Functional Movement Screen™ score in adolescent soccer players. *Journal of Science and Medicine in Sport*, 19(10), 854–858. doi:10.1016/j.jsams.2015.12.001

Reid, A., Birmingham, T. B., Stratford, P. W., Alcock, G. K., Giffin, J. R. (2007). Hop testing provides a reliable and valid outcome measure during rehabilitation after anterior cruciate ligament reconstruction. *Physical Therapy*, 87(3), 337–349. doi:10.2522/ptj.20060143

Shrout, P. E., Fleiss, J. L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin*, 86(2), 420–428. doi:10.1037/0033-2909.86.2.420

Smith, T. B., Hopkins, W. G. (2011). Variability and predictability of finals times of elite rowers. *Medicine and Science in Sports and Exercise*, 43(11), 2155–2160. doi:10.1249/mss.0b013e31821d3f8e

Wijnbergen, M. (2019). Novel algorithms to capture kinematic variables with depth-sensing technology. The development of a reliable, valid and practical movement assessment tool (Doctoral dissertation, Teesside University). Middlesbrough, UK: Teesside University.

Wright, M. D., Chesterton, P. (2019). Functional movement screen™ total score does not present a gestalt measure of movement quality in youth athletes. *Journal of Sports Sciences*, 37(12), 1393–1402. doi:10.1080/02640414.2018.1559980