THE USE OF PRINCIPAL COMPONENT ANALYSIS FOR REDUCTION OF TRAINING LOAD DATA IN PROFESSIONAL SOCCER

Perry Nosek[1,2], Matthew Andrew[3], Mladen Sormaz[4], Barry Drust[5], and Thomas E. Brownlee[5]

[1]Leicester City Football Club, Leicester, UK
[2]School of Sport and Exercise Science, Liverpool John Moores University, UK
[3]Department of Sport and Exercise Sciences, Institute of Sport, Manchester Metropolitan University, Manchester, UK
[4]777 Partners Football Group, Miami, USA
[5]School of Sport, Exercise and Rehabilitation Sciences, University of Birmingham, UK

Original scientific paper
DOI 10.26582/k.55.2.5

Abstract:

The aim of this study was to explore the use of principal component analysis (PCA) in understanding multivariate relationships in soccer training load data. Training load data were collected from 20 professional male soccer players during a 28-week in-season period. Twelve training load variables (total distance, PlayerLoad™, low-speed running distance, moderate-speed running distance, high-speed running distance, sprint distance, moderate-speed running efforts, high-speed running efforts, sprint efforts, accelerations, decelerations, and changes of direction) were collected during training sessions, with correlation analysis revealing high intercorrelation between most variables (r = 0.04-0.98). Principal component analysis was performed on datasets containing all players and on individual players. On the whole dataset, two principal components were retained explaining a total of 81% of data variance. The first component comprised variables associated with distances in speed zones and the second component changes of direction. Whilst some individual variation existed among players, distances in speed zones were loaded on the first component and inertial movement analysis variables, such as accelerations, decelerations, and changes of direction, were loaded on the second component. These findings evidence the strong relationships between several common training load variables and highlight the risk of data redundancy. By selecting variables from each component, practitioners can reduce the number of variables reported whilst retaining as much of the variation in data as possible.

Key words: *training load, Global Positioning Units (GPS), multivariate analysis, football, feedback*

Introduction

To maximise soccer performance, professional players participate in training sessions (i.e., practice) with the aim to improve technical, tactical, physical, and psychological performance (Morgans, Orme, Anderson, & Drust, 2014; Williams & Reilly, 2000) From a physical perspective, the volume and intensity of these sessions, known collectively as the training load (Impellizzeri et al., 2019), are planned and manipulated to produce this desired response. Utilisation of soccer training load monitoring technologies, such as Global Positioning Systems (GPS), is now considered common in soccer (Akenhead & Nassis, 2015). The success of using such devices lies in their ability to support and influence the decision-making of key stakeholders (e.g., coaches). For example, coaches generally support the usefulness and importance of training data collected by these technologies (Nosek, Brownlee, Drust & Andrew, 2021; Weston, 2018), yet it is still unclear whether this information is impactful in aiding decision-making and may represent the 'translational gap' often suggested within sport science (Eisenmann, 2017; Fullagar, McCall, Impellizzeri, Favero, & Coutts, 2019). Such gaps may be due to coaches perceiving a lack of a common goal with their sport science departments as a barrier towards using training data in coach decision-making (Nosek, et al., 2021).

Feedback on performance related data is a key step in the coaching process (Franks & Goodman,

2008) and involves the collection, analysis, and delivery/communication of data to those responsible for decision-making (e.g., coaches; Buchheit, 2017). Perhaps the most important part of this process may be the communication of this data to coaches as if done poorly, it can create a barrier to data utilisation (Nosek, et al., 2021). Consequently, the delivery format and complexity of chosen data must be considered carefully when designing feedback to ensure efficient communication of key data to stakeholders and decision makers. Adding to feedback complexity is potential disagreements between coaches and sport scientists regarding the information provided. For example, Weston (2018) reported differences between coaches and practitioners (sport scientists; fitness coaches; strength and conditioning coaches) regarding the frequency, timing, and expertise requirements of training load reports. Furthermore, whilst practitioners identify typically using 4-10 variables in reports (Akenhead & Nassis, 2015), coaches have suggested receiving too much information is a barrier towards using training data to inform decision-making (Nosek, et al., 2021). These variables often show strong relationships with each other, with intercorrelations existing between various measures such as total distance, PlayerLoadÔ (sum of accelerations in all planes collected via tri-axial accelerometer), high-speed running, sprinting, session rating of perceived exertion (sRPE), and heart rate (HR)-derived measures during soccer training (Casamichana, Castellano, Calleja-Gonzalez, San Román, & Castagna, 2013; Maughan, MacFarlane, & Swinton, 2021; McLaren, et al., 2018). This high intercorrelation may result in data 'overload' for coaches, and reporting many variables appears unnecessary and may increase complexity during decision-making (Weaving, Beggs, Dalton-Barron, Jones, & Abt, 2019; Weaving, Marshall, Earle, Nevill, & Abt, 2014).

Reducing variable number (dimensionality reduction) can be complex, as simple variable removal can lead to information loss. A popular dimension reduction technique is principal component analysis (PCA), which extracts important information from correlated variables and expresses them as new, uncorrelated compound variables named principal components (PCs) (Jolliffe, 1986; Jollife & Cadima, 2016). In a sporting context, PCA has previously been used to examine technique analysis (Federolf, Reid, Gilgien, Haugen, & Smith, 2014; Gløersen, Myklebust, Hallén, & Federolf, 2018), injury risk (Williams, Trewartha, Cross, Kemp, & Stokes, 2017), performance indicators (Parmar, James, Hearne, & Jones, 2018), and training load (Weaving, et al., 2014, 2018, 2019; Weaving, Jones, Till, Abt, & Beggs, 2017). The PCs produced can then have hypotheses framed around them. For example, Parmar et al. (2018) examined the classification accuracy of PCs against the win/loss probabilities in rugby league and reported a 90% accuracy. Similarly, Williams et al. (2017) examined the relationship between variables in each PC against injury risk in rugby union, reporting that 4-week cumulative load, acute:chronic workload ratio, and daily workload were the measures describing the largest amount of variation in injury risk from the first three PCs, respectively. Taken together, these studies provide a framework to reduce large datasets into groups of variables that can then be explored against outcome measures of interest (Williams, et al., 2017).

Recently, PCA has been applied to training data collected via GPS with the aim of reducing the number of variables used in training load monitoring. For instance, Weaving et al. (2018) identified 60-70% of training load variance in field-based skills training could be provided by a PC containing total distance, sRPE or PlayerLoadÔ, supporting the use of these variables as monitoring measures. Moreover, Weaving et al. (2019) demonstrated how 12 training load variables could be transformed into a 2D scatterplot, allowing for heuristic decision-making such as modifying future training content after comparing present day data to historical benchmarks. In soccer, Maughan et al. (2021) illustrated that multiple measures of subjective and external training load variables could be reduced to two PCs that explained 83% of the variance within the data, one which contained all the variables to represent total training load, and one which contrasted subjective and external measures. Furthermore, during the competitive season, PCA produced only one PC which housed all training load variables excluding sprinting, which suggests all the variables used represent similar underlying information and could theoretically be used interchangeably (Maughan, MacFarlane, & Swinton, 2022). Despite some interesting results here, research utilising PCA within professional soccer is limited.

Whilst most training load PCA studies analyse whole squad data (Parmar, et al., 2018; Ryan, Kempton, & Coutts, 2021; S. Williams, et al., 2017), this may result in individual player characteristics being hidden in the analysis. One possible approach is to perform PCA on each player. Using this approach, Weaving et al. (2018) reported that the same variables were loaded on each PC for all rugby union players. Despite this, some individual variation in loadings were reported meaning that utilising PC scores (i.e., standardized training load data multiplied by PC loadings for each variable), as suggested by the authors, would only allow within-player comparisons and not between-player. Whilst PCA appears a suitable method for dimension reduction of training load data, its application in soccer is not understood. Therefore, the main aim

of the present study was to explore the use of PCA to uncover multivariate relationships within soccer training load data. In line with previous research (Casamichana, et al., 2013; Weaving, et al., 2018, 2019) we hypothesized high intercorrelation would exist within soccer training load data and that these relationships would vary between players when analysed with PCA.

Methods

Participants

Twenty male professional soccer players (age 20.50 ± 1.20 years, height 178.60 ± 6.60 cm [Seca 213, Seca, Hamburg, Germany], weight, 80.10 ± 8.10 kg [Seca 876, Seca, Hamburg, Germany]) were recruited for this study. All players came from the same Professional Development Phase squad at an English Premier League club that was competing in the Premier League 2-Divison 1 (i.e., the highest tier) for the 2019/20 season. Goalkeepers were excluded from the analysis due to their vastly different training and game physical requirements (Moreno-Pérez, et al., 2020). Players represented the following primary positions: centre-back (n = 4), full-back (n = 4), centre-midfield (n = 5), wide-midfield (n = 3) and forward (n = 4). The inclusion criteria were that players must have completed a minimum of 50 available sessions (mean 83 ± 12). Data were provided entirely as part of players' normal daily training routine; thus, no ethical approval was required. The study did, however, conform to the Declaration of Helsinki and Gatekeeper written consent was provided to allow data use.

Experimental design

A longitudinal, observational design was used with training load data collected during 28 weeks of the competitive, in-season period between August 2019 and March 2020. Although the competitive season typically runs to May, the season concluded early due to the COVID-19 pandemic. Prior to the season, microcycle structure was designed with input from coaching and sport science departments to meet the tactical, technical, and physical requirements of the game. During weeks containing one match (MD; n = 18), this involved a recovery session the day after the match (i.e., match day plus one; MD+1), followed by a day off. There would then be three consecutive days of conditioned training before a tactical themed activation session the day before the next match (MD-1; n = 18).

The themes for the three conditioned training days were termed *strength* (MD-4; n = 25); intensive work in small areas of <110 m^2 per player over duration bouts of 45-s to 4-min per activity, with small player numbers (2-12, to overload acceleration, deceleration, change of direction and metabolic demands; *endurance* (MD-3; n = 42); extensive work in large areas of greater than 160 m^2 per player over durations of 4-12 min, with increased player numbers (8-22) to overload high-speed running demands; and *speed* (MD-2; n = 29); work in moderately sized areas of 110-160 m^2 focusing on speed of play (Buchheit, Lacome, Cholley & Simpson, 2018), which involved various numbers of players per activity. These themes typically corresponded to MD-4, MD-3, and MD-2, respectively, though not in all cases. Training was typically conducted in the morning with games in the evening.

Training content was designed to reflect the fitness and physical status of individual players (i.e., number of days between games, acute and chronic loadings, wellbeing responses to training) and therefore, whilst the areas, durations and players numbers outlined were typical of each theme, these were not strict constraints for the sessions. Furthermore, some sessions included individual work and/or physical sessions designed to provide additional stimuli typically to substitutes known as 'top-ups', as required. Training data from rehabilitation, individual, physical fitness and partially completed (i.e., player injured during session) sessions were excluded from the analyses.

Equipment and procedures

During sessions, training load data were recorded using 10 Hz GPS with an embedded 100 Hz tri-axial accelerometer (Catapult Vector, Catapult Sports, Melbourne). Units were worn in a tight-fitting vest placing the unit between the scapulae, with players wearing the same unit for each session to minimise inter-unit variability. All players had previous experience of wearing the vests and units as it was part of routine working practice to collect such data. Prior to training, units were turned on outside for 30 min to allow optimal connectivity with satellites. After each session, data were downloaded into the manufacturer's software (Openfield v2.2) and inspected for artefacts (i.e., unrealistic spikes in velocity). Data were collected and analysed by a member of the sports science department at the club who is a current professional doctorate student and lead author of the study. Data were included if the number of connected satellites was at least six and if the horizontal dilution of precision (HDOP) was <1.5 as per manufacturer guidance. Where these conditions were not met, or full session data were not available due to other errors such as units running out of battery, these data were removed and replaced with positional mean averages for that session (Jaspers, et al., 2018). This method was chosen to represent what had occurred in the current session as opposed to using player previous data of sessions which may have looked vastly different. This resulted in 11 out of 1466 (0.8%) individual player files being replaced.

Data were split into individual drills, exported into a custom Excel spreadsheet and assigned the corresponding session theme. The variables selected for analysis are described in Table 1. These were based primarily on research involving training and match load and associated fitness, fatigue and injury responses, alongside some variables used as part of a club-wide monitoring philosophy (Barrett, Midgley, & Lovell, 2014; Bradley, et al., 2009; Luteberget, Holme, & Spencer, 2018). Accelerations, decelerations, and changes of direction were calculated based on the manufacturer's *inertial movement analysis* (IMA). This uses accelerometer and gyroscope data to count one-step efforts, the magnitude of which is expressed as delta velocity ($m \cdot s^{-1}$) (Luteberget, et al., 2018). Although limited research has shown IMA counts in intensity bands to possess moderate reliability, this is shown to improve when multiple bands, such as medium and high intensity, are aggregated (Luteberget, et al., 2018) as in the present study. Further, it is hoped that improvements in effort detection algorithms through updates in the manufacturer's software have improved this. Dwell time was set at 0.5s for GPS variables.

Though HR-derived variables were collected, due to a change in measurement devices mid-season and data recording issues that arose from this change, these variables were removed from the analyses.

Data reduction and analysis

Prior to PCA, the dataset, which consisted of 1466 rows of data, was explored for missing/erroneous data, which may have occurred due to units not being worn or poor satellite connectivity. These data were replaced with the session positional averages. Redundancy of the dataset was examined using repeated-measures correlation to view the strength of relationships between the variables (Bakdash & Marusich, 2017). The qualitative descriptors for the magnitude of the correlations were: <0.1 trivial; 0.1 to 0.3 small; 0.3 to 0.5 moderate; 0.5 to 0.7 large; 0.7 to 0.9 very large; 0.9 to 1.0 almost perfect (Hopkins, 2010). Additionally, the Bartlett test of sphericity and the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy were performed to assess the suitability of the data for PCA. The Bartlett test of sphericity was significant ($p<.01$) with a KMO of 0.7, with a KMO above 0.5 suggesting data were suitable for PCA (Williams, Onsman, & Brown, 2010).

Data were mean-centred and scaled to unit variance allowing equal weighting across variables with differing measurement units such as distances and counts. PCA was performed on the whole dataset using the singular value decomposition method where components with an eigenvalue of >1 were retained for analysis and indicated the PC accounted for more variance than a single original variable alone (Kaiser, 1960). Visual inspection of the scree plot to identify the 'elbow' of the eigenvalues was also utilised to support this decision (Nguyen & Holmes, 2019). Factor loadings, the strength of a variable's relationship with the PC, were considered meaningful if it exceeded 0.7 (Rojas-Valverde, Pino-Ortega, Gómez-Carmona, & Rico-González, 2020). Subsequently, the same analysis was performed on each individual players' dataset using the same method. All analysis was performed in R (Version 4.0.0) using the FactoMineR package for PCA (Lê, Josse, Rennes, & Husson, 2008).

Table 1. Description of GPS training load variables

Variable (Abbreviation)	Description
Total Distance (TD)	Total distance covered during the session.
Player Load™ (PL)	Accumulated accelerometer data across vertical, medio-lateral, and anterior-posterior planes, divided by a scaling factor of 100.
Low-Speed Running Distance (LSRD)	Distance covered between 0 and 4.5$m \cdot s^{-1}$.
Moderate-Speed Running Distance (MSRD)	Distance covered between 4.5 and 5.5$m \cdot s^{-1}$.
High-Speed Running Distance (HSRD)	Distance covered between 5.5 and 7$m \cdot s^{-1}$.
Sprint Distance (SD)	Distance covered above 7$m \cdot s^{-1}$.
Moderate-Speed Running Efforts (MSRE)	Number of efforts between 4.5 and 5.5$m \cdot s^{-1}$.
High-Speed Running Efforts (HSRE)	Number of efforts between 5.5 and 7$m \cdot s^{-1}$.
Sprint Efforts (SE)	Number of efforts above 7$m \cdot s^{-1}$.
Accelerations (ACC)	Number of medium and high intensity accelerations above 2.5$m \cdot s^{-1}$ derived from *inertial movement analysis* (IMA) using accelerometer and gyroscope data.
Decelerations (DEC)	Number of medium and high intensity decelerations above 2.5$m \cdot s^{-1}$ derived from IMA using accelerometer and gyroscope data.
Changes of Direction (COD)	Number of medium and high intensity changes of direction to the left above 2.5$m \cdot s^{-1}$ derived from IMA using accelerometer and gyroscope data.

Results

The repeated-measures correlation matrix is shown in Table 2. All correlations were significant ($p<.01$) except for between sprint efforts (SE) and changes of direction (COD) ($p=.12$).

Principal component analysis of the whole dataset extracted two PCs as having an eigenvalue of >1 with PC1 (65%) and PC2 (16%) accounting accumulatively for 81% of the variance in training load data (Table 3). The correlation between each training load variable and each PC is also shown in Table 3. Whilst all variables were loaded somewhat on PC1, those relating to overall volume and running in different speed zones were loaded above the meaningful threshold (Figure 1). IMA variables were highest loaded on PC2, however, only COD were above the meaningful threshold (Table 3).

When individual players were analysed, all players produced two PCs with eigenvalues >1 (PC1 eigenvalue range 6.32-9.13; PC2 eigenvalue range 1.3-2.65). PC1 accounted for 53-76% of the variance, whilst PC2 accounted for a further 11-22%. Loadings between the variables and PCs for each player are shown in Table 4 and Table 5.

Discussion and conclusions

The primary aim of the present study was to explore the use of PCA to uncover multivariate relationships within soccer training load data to reduce reporting redundant data to coaches. Correlation analysis revealed strong relationships between multiple training load variables with subsequent PCA identifying two PCs explaining a combined 81% of the variance in training load data. Although similar within-player results were produced, some variation in variable loadings on each PC existed suggesting that some individual player characteristics may be hidden by analysing the full dataset.

Correlation analysis revealed most variables had a moderate to almost perfect relationship, although some correlations between IMA-based variables and running variables were trivial or small (Table 2). This highlights the redundancy in the dataset in that many of the variables will change at similar rates across the training period and supports the need to perform data reduction techniques to explore these relationships further. These results are similar to others, which have reported very large to almost perfect correlations between total distance (TD)

Table 2. Correlation matrix (95% confidence intervals) for each training load variable during soccer training

	TD	PL	LSRD	MSRD	HSRD	SD	MSRE	HSRE	SE	ACC	DEC	COD
TD	1											
PL	0.98 (0.98-0.99) AP	1										
LSRD	0.99 (0.99-1.00) AP	0.98 (0.98-0.98) AP	1									
MSRD	0.93 (0.92-0.94) AP	0.90 (0.89-0.91) AP	0.90 (0.88-0.91) AP	1								
HSRD	0.88 (0.86-0.89) VL	0.85 (0.83-0.86) VL	0.83 (0.81-0.85) VL	0.89 (0.87-0.90) VL	1							
SD	0.67 (0.63-0.70) L	0.65 (0.61-0.69) L	0.63 (0.59-0.67) L	0.60 (0.56-0.64) L	0.75 (0.72-0.79) VL	1						
MSRE	0.93 (0.92-0.93) AP	0.89 (0.88-0.90) VL	0.88 (0.86-0.89) VL	0.96 (0.95-0.96) AP	0.84 (0.83-0.86) VL	0.63 (0.60-0.66) L	1					
HSRE	0.85 (0.83-0.86) VL	0.80 (0.78-0.82) VL	0.78 (0.76-0.80) VL	0.89 (0.88-0.90) VL	0.94 (0.94-0.95) AP	0.75 (0.72-0.77) VL	0.92 (0.91-0.92) AP	1				
SE	0.67 (0.65-0.70) L	0.63 (0.60-0.66) L	0.61 (0.58-0.64) L	0.66 (0.63-0.69) L	0.79 (0.77-0.80) VL	0.93 (0.93-0.94) AP	0.69 (0.66-0.72) L	0.81 (0.80-0.83) VL	1			
ACC	0.37 (0.32-0.43) M	0.45 (0.40-0.50) M	0.40 (0.35-0.45) M	0.24 (0.18-0.30) S	0.24 (0.18-0.30) S	0.26 (0.20-0.32) S	0.24 (0.19-0.29) S	0.15 (0.10-0.20) S	0.13 (0.08-0.18) S	1		
DEC	0.46 (0.41-0.51) M	0.49 (0.44-0.54) M	0.49 (0.44-0.54) M	0.34 (0.28-0.39) M	0.30 (0.24-0.36) M	0.21 (0.15-0.27) S	0.26 (0.21-0.30) S	0.15 (0.09-0.20) S	0.06 (0.01-0.11) T	0.35 (0.30-0.41) M	1	
COD	0.48 (0.43-0.53) M	0.55 (0.50-0.59) L	0.51 (0.46-0.57) L	0.36 (0.30-0.41) M	0.31 (0.25-0.37) M	0.22 (0.16-0.28) S	0.28 (0.23-0.33) S	0.13 (0.08-0.19) S	0.04 (-0.01-0.09) T	0.58 (0.54-0.62) L	0.65 (0.62-0.69) L	1

Note. AP – almost perfect, VL – very large, L – large, M – moderate, S – small, T – trivial; TD – Total Distance, PL – PlayerLoad™, MSRD – Moderate-Speed Running Distance, HSRD – High-Speed Running Distance, SD – Sprint Distance, MSRE – Moderate-Speed Running Efforts, HSRE – High-Speed Running Efforts, SE – Sprint Efforts, ACC – Accelerations, DEC – Decelerations, COD – Changes of Direction.

Table 3. Principal component analysis results for the training load data showing eigenvalues, percentage of variance explained, cumulative variance explained, and component loadings for the first two PCs

	PC1	PC2
Eigenvalue	7.78	1.96
% of total variance explained	64.85	16.35
Cumulative % of total variance explained	64.85	81.19
Variable loadings		
TD	**0.96**	0.08
PL	**0.95**	0.15
LSRD	**0.93**	0.17
MSRD	**0.91**	-0.15
HSRD	**0.88**	-0.33
SD	**0.74**	-0.35
MSRE	**0.95**	-0.07
HSRE	**0.93**	-0.25
SE	**0.79**	-0.37
ACC	0.37	0.69
DEC	0.45	0.64
COD	0.47	**0.75**

Note. TD – Total Distance, PL – PlayerLoad™, MSRD – Moderate-Speed Running Distance, HSRD – High-Speed Running Distance, SD – Sprint Distance, MSRE – Moderate-Speed Running Efforts, HSRE – High-Speed Running Efforts, SE – Sprint Efforts, ACC – Accelerations, DEC – Decelerations, COD – Changes of Direction.
Loadings that met interpretation criteria (≥ 0.7) are highlighted in bold

Table 4. Correlations between variables and PC1 for individual players

Player	TD	PL	LSRD	MSRD	HSRD	SD	MSRE	HSRE	SE	ACC	DEC	COD
1	**0.98**	**0.98**	**0.97**	**0.96**	**0.96**	**0.88**	**0.97**	**0.97**	**0.92**	0.28	0.67	0.63
2	**0.98**	**0.98**	**0.97**	**0.96**	**0.96**	**0.88**	**0.97**	**0.97**	**0.92**	0.37	0.40	0.30
3	**0.96**	**0.94**	**0.93**	**0.81**	**0.84**	0.65	**0.95**	**0.91**	**0.75**	0.28	0.57	0.55
4	**0.95**	**0.94**	**0.90**	**0.84**	**0.54**	0.60	**0.96**	**0.88**	0.56	0.69	0.59	**0.76**
5	**0.98**	**0.96**	**0.96**	**0.95**	**0.95**	**0.86**	**0.97**	**0.96**	**0.90**	0.47	0.50	0.39
6	**0.98**	**0.98**	**0.97**	**0.95**	**0.91**	**0.82**	**0.96**	**0.94**	**0.85**	0.35	0.42	0.64
7	**0.96**	**0.95**	**0.94**	**0.93**	**0.86**	**0.74**	**0.95**	**0.94**	**0.81**	0.57	0.50	0.57
8	**0.96**	**0.94**	**0.92**	**0.90**	**0.87**	**0.78**	**0.95**	**0.91**	**0.80**	0.27	0.43	0.25
9	**0.93**	**0.87**	**0.85**	**0.79**	**0.70**	**0.77**	**0.91**	**0.88**	**0.78**	0.17	0.16	0.18
10	**0.95**	**0.94**	**0.91**	**0.80**	**0.71**	**0.80**	**0.89**	**0.85**	**0.77**	0.41	0.51	0.61
11	**0.98**	**0.98**	**0.96**	**0.96**	**0.92**	**0.85**	**0.96**	**0.95**	**0.90**	0.64	0.60	0.51
12	**0.97**	**0.98**	**0.96**	**0.94**	**0.94**	0.66	**0.96**	**0.95**	**0.80**	0.60	0.52	**0.71**
13	**0.97**	**0.97**	**0.94**	**0.93**	**0.89**	**0.85**	**0.96**	**0.93**	**0.90**	0.55	0.49	0.60
14	**0.96**	**0.91**	**0.90**	**0.84**	**0.84**	**0.80**	**0.94**	**0.91**	**0.79**	0.40	0.39	0.34
15	**0.98**	**0.98**	**0.97**	**0.92**	**0.95**	0.52	**0.95**	**0.95**	0.64	0.34	0.68	**0.79**
16	**0.97**	**0.95**	**0.92**	**0.93**	**0.85**	0.54	**0.96**	**0.93**	0.66	0.11	0.48	0.51
17	**0.98**	**0.97**	**0.96**	**0.93**	**0.95**	**0.71**	**0.96**	**0.96**	**0.78**	0.58	0.38	0.50
18	**0.98**	**0.97**	**0.97**	**0.98**	**0.96**	**0.85**	**0.97**	**0.98**	**0.85**	0.41	0.18	0.12
19	**0.97**	**0.94**	**0.93**	**0.89**	**0.82**	**0.82**	**0.94**	**0.91**	**0.88**	0.43	0.53	0.49
20	**0.97**	**0.93**	**0.93**	**0.84**	**0.76**	**0.75**	**0.94**	**0.91**	**0.81**	0.39	0.54	0.52

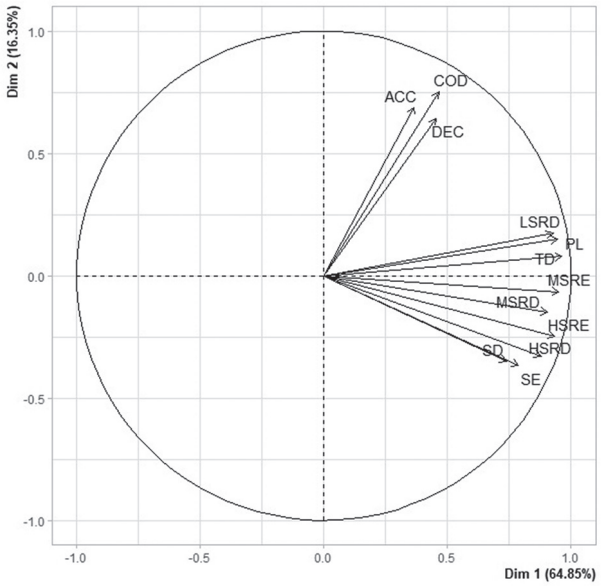
Loadings that met interpretation criteria (≥ 0.7) are highlighted in bold

and PlayerLoad™ (PL) in soccer (Casamichana, et al., 2013; Maughan, et al., 2021; Scantlebury, et al., 2020), which is likely due to these measures being functions of the duration of the session and accumulations of all other variables. In contrast, however, whilst this study showed a very large relationship between TD and high-speed running distance (HSRD), Scantlebury et al. (2020) reported only a

Table 5. Correlations between variables and PC2 for individual players

Player	TD	PL	LSRD	MSRD	HSRD	SD	MSRE	HSRE	SE	ACC	DEC	COD
1	-0.03	0.01	0.01	-0.16	-0.20	-0.11	-0.12	-0.15	-0.22	**0.84**	0.36	0.69
2	0.17	0.32	0.33	-0.28	-0.41	-0.42	-0.03	-0.29	-0.46	**0.71**	0.66	**0.86**
3	0.03	0.24	0.14	-0.34	-0.31	-0.28	-0.11	-0.24	-0.32	**0.85**	0.48	**0.71**
4	-0.23	-0.30	-0.37	0.29	**0.70**	0.64	0.07	0.41	0.68	-0.49	-0.43	-0.52
5	-0.02	0.11	0.04	-0.17	-0.22	-0.27	-0.09	-0.18	-0.25	0.64	0.68	**0.82**
6	0.03	0.11	0.12	-0.17	-0.29	-0.27	-0.09	-0.23	-0.32	**0.72**	0.69	0.68
7	0.06	0.13	0.12	-0.07	-0.33	-0.44	-0.02	-0.22	-0.39	0.35	0.69	0.68
8	0.13	0.26	0.23	-0.15	-0.40	-0.32	-0.02	-0.26	-0.43	**0.81**	**0.76**	**0.84**
9	0.20	0.37	0.33	-0.22	-0.47	-0.23	-0.05	-0.26	-0.28	0.59	**0.82**	**0.87**
10	0.10	0.26	0.23	-0.41	-0.53	-0.18	-0.28	-0.42	-0.19	**0.75**	0.66	0.68
11	0.04	0.09	0.10	-0.15	-0.20	-0.31	-0.12	-0.22	-0.28	0.38	0.56	**0.72**
12	-0.06	-0.02	-0.01	-0.23	-0.25	-0.09	-0.14	-0.21	-0.17	0.55	**0.72**	0.52
13	0.07	0.10	0.18	-0.25	-0.36	-0.26	-0.13	-0.28	-0.28	0.60	0.67	0.69
14	0.14	0.29	0.28	-0.28	-0.40	-0.34	-0.11	-0.28	-0.33	0.63	**0.72**	**0.82**
15	-0.10	-0.10	-0.13	-0.14	0.15	**0.80**	-0.11	0.11	**0.72**	0.11	-0.50	-0.34
16	0.15	0.26	0.30	-0.16	-0.41	-0.53	-0.01	-0.23	-0.56	0.50	0.60	**0.72**
17	0.02	0.07	0.09	-0.15	-0.21	-0.32	-0.09	-0.18	-0.33	0.40	**0.75**	**0.77**
18	0.01	0.10	0.04	-0.02	-0.12	-0.32	0.07	-0.01	-0.35	0.60	**0.81**	**0.91**
19	0.12	0.27	0.25	-0.29	-0.41	-0.34	-0.15	-0.32	-0.30	**0.70**	0.56	**0.76**
20	0.09	0.29	0.23	-0.30	-0.54	-0.27	-0.13	-0.34	-0.35	**0.75**	0.62	**0.75**

Loadings that met interpretation criteria (≥ 0.7) are highlighted in bold



Note. TD – Total Distance, PL – PlayerLoad™, MSRD – Moderate-Speed Running Distance, HSRD – High-Speed Running Distance, SD – Sprint Distance, MSRE – Moderate-Speed Running Efforts, HSRE – High-Speed Running Efforts, SE – Sprint Efforts, ACC – Accelerations, DEC – Decelerations, COD – Changes of Direction.

Figure 1. PCA loading plot for the two extracted principal components.

moderate relationship. Furthermore, Scantlebury et al. (2020) reported only a trivial correlation between PL and HSRD compared with a very large relationship in the present study. These differences may be due to the method used to define HSRD (distance covered above 61% of a player's maximum velocity compared with an arbitrary threshold of 5.5m·s^{-1} in the present study) or the large differences in training output and methodology used by semi-professional players used by Scantlebury et al (2020). Whilst correlations between running-based variables were strong, those between IMA variables and running-based variables were weaker. This could be due to the different physical qualities being captured by IMA variables since the one-step actions of accelerations (ACC), decelerations (DEC) and changes of direction (COD) are highly independent of any running effort in different speed zones. For example, to produce an effort in the sprint speed zone, a player must travel through all previous zones, thus accumulating TD, low-speed running distance (LSRD), moderate-speed running distance (MSRD), HSRD and sprint distance (SD), which can be achieved without registering any ACC, DEC or COD. Taken together, the reported high intercorrelation suggests that all these measures may lead to data redundancy and unnecessary data overload for coaches.

Performing PCA on the whole dataset uncovered two PCs identified as having eigenvalues >1 indicating these new composite variables account for more variance in training load data than a single original variable (Kaiser, 1960). The first PC accounted for 65% of the total variance in the

dataset with running- and volume-based variables showing meaningful relationships with the component (Table 3). The second PC accounted for 16% of the total variance where COD was the only variable with a meaningful relationship with the component, although the other IMA variables loaded just short of the threshold (Table 3) yet within other reported acceptable thresholds (Rojas-Valverde, et al., 2020). This demonstrates the independent information that running-based metrics and explosive actions provide which is unsurprising given their differing physical requirements. These results are dissimilar to Maughan et al. (2021) who reported that accelerations and decelerations measured by GPS were loaded on the first component with other running-based metrics such as TD, PL, HSRD and subjective measures of session rating of perceived exertion (sRPE). The differences here may highlight the benefit of using IMA-derived accelerations and decelerations over those produced by GPS, as these provide additional information to running-based GPS variables. The results here show some similarities to those by Scantlebury et al. (2020) who also showed that TD and PL were heavily weighted on PC1, however, in contrast to this study, they reported high loadings for HSRD on PC2. This is likely due to differences in variables collected as no IMA data were reported (Scantlebury, et al., 2020). This shows how the results of PCA are dependent on the variables used, meaning practitioners should perform PCA on their own dataset and not use results from research studies.

When PCA was performed on individual players, similar results were found; however, some individual differences were observed. For example, COD was highly associated with PC1 for three players (Table 4) and on PC2 there were various IMA variables (Table 5) that were meaningfully loaded for each player. These results contrast the main findings of Weaving et al. (2018) who showed that all rugby union players had sRPE, TD and PL meaningfully loaded on PC1 and HSRD loaded on PC2. These differences could be attributed to the type of dataset analysed, with only skills training used in Weaving et al. (2018), which will likely yield different results due to the differences in relationships between training load variables in different modes of training (Lovell, Sirotic, Impellizzeri, & Coutts, 2013; Weaving et al., 2014). Whilst the individual variation highlights the unique training outputs between players, selecting different variables for different players would seem nonsensical and would make the evaluation of training sessions difficult for coaches.

To make use of PCA results, practitioners have several options. By multiplying the standardized training data by the loadings on each PC, PC scores can be produced giving a single score for each PC (Weaving, et al., 2019). Whilst this may seem ideal in that it will reduce 12 training load variables down to just two yet still retaining 81% of the variance, the new variables, reported in arbitrary units, may increase the complexity of the report and misunderstanding of coaches. For example, understanding what an increase of 1AU for PC1 and making inferences around this is very difficult. Applying this method to individually analysed datasets would further reduce interpretability if different variables for different players were used to create the PC scores which would make between-player comparisons impossible. Furthermore, the repeated-measures nature of training load monitoring would mean the standardized data is updated with each training session, thus adjusting the PCA model and rendering comparisons between sessions using PC scores as unworthwhile.

An alternative approach to using the PCA information is to simply select variables from each PC that are highly loaded. In its simplest form, this may be selecting a single variable from each PC, such as those that have the highest validity and reliability or practicality (Ryan et al., 2021). Although many variables are available to choose from PC1, TD has been shown to be a valid and reliable measure that is easily understood by coaches (Johnston, Watsford, Kelly, Pine, & Spurrs, 2014). Yet this measure does not account for any distances in higher speed zones which are an important aspect of soccer match play (Barnes, Archer, Hogg, Bush, & Bradley, 2014).

Selecting a single variable from PC2 may be more difficult as the IMA variables show similar levels of validity and reliability as well as interpretability (Luteberget, et al., 2018). In this regard, practitioners may aggregate variables that are both loaded above the threshold on the PC and share the same measurement units. For example, HSRD and SD could be grouped from the first PC, whilst ACC, DEC and COD could be grouped from the second PC. Using these two new aggregate variables would give a coach an idea of the volume of high-speed running and the number of intense actions in the session, with both giving unique information and expressed in interpretable units as opposed to the arbitrary units of PC scores. It could be suggested that sport scientists utilise PCA results, their domain-specific expertise, and input from other key stakeholders such as coaches to co-create an impactful feedback tool (Richter, O'Reilly, & Delahunt, 2021).

Despite the potential application of PCA seen here, several limitations exist. No internal training load variables such as HR-based measures and sRPE were used in this study, which may provide additional information due to the individual internal response to external training load (Impellizzeri, et al., 2019) and the differences in internal load between different training formats (Owen, Wong, Mckenna, & Dellal, 2011). Furthermore, only one

soccer team was studied, meaning the results likely depend on the periodisation model and training strategy of the team, and thus may not be generalisable to other soccer teams or team sports. The team studied showed a distinct variation in loadings throughout the training week whereby different physical qualities were targeted on each day, which may in part explain the correlations between variables and PCA loadings. It is recommended that sport scientists perform PCA on their own data set to provide insights relating to their team's specific periodisation and loading strategy. Finally, speed zones analysed were arbitrary, which fails to account for differences in fitness and athleticism across the squad (Hunter, et al., 2014). Future research may look to perform PCA using individualised speed and acceleration zones as well as comparing variable relationships between positions and competitive level.

In sum, the present study aimed to address concerns from soccer coaching staff that too much information is fed back to them regarding GPS-derived training load data (Nosek, et al., 2021). Data reduction was undertaken using PCA which identified two PC's, suggesting a multivariate approach is needed when utilising training data. Results from both the whole dataset and individual analysis demonstrated how PCA can be used to uncover multivariate relationships between twelve training load variables, with variables relating to volume and running distances in speed zones associated with the first PC and IMA-derived intensive effort variables mostly associated with PC2. Practitioners can therefore be confident that by reporting variables from each PC they capture unique information compared to using multiple variables from a single PC. The impact of these results, however, relies on the collaboration between sport science practitioners and coaches to select variables that help answer coach questions, such as those pertinent to planning and evaluating training.

References

Akenhead, R., & Nassis, G. P. (2015). Training load and player monitoring in high-level football: Current practice and perceptions. *International Journal of Sports Physiology and Performance*, *11*(5), 587–593. https://doi.org/10.1123/ijspp.2015-0331

Bakdash, J. Z., & Marusich, L. R. (2017). Repeated measures correlation. *Frontiers in Psychology*, *8*, 1–13. https://doi.org/10.3389/fpsyg.2017.00456

Barnes, C., Archer, D., Hogg, B., Bush, M., & Bradley, P. (2014). The evolution of physical and technical performance parameters in the English Premier League. *International Journal of Sports Medicine*, *35*(13), 1095–1100. https://doi.org/10.1055/s-0034-1375695

Barrett, S., Midgley, A., & Lovell, R. (2014). PlayerLoad™: Reliability, convergent validity, and influence of unit position during treadmill running. *International Journal of Sports Physiology and Performance*, *9*(6), 945–952. https://doi.org/10.1123/IJSPP.2013-0418

Bradley, P. S., Sheldon, W., Wooster, B., Olsen, P., Boanas, P., & Krustrup, P. (2009). High-intensity running in English FA Premier League soccer matches. *Journal of Sports Sciences*, *27*(2), 159–168. https://doi.org/10.1080/02640410802512775

Buchheit, M. (2017). Want to see my report coach. *Aspetar Sports Medicine Journal*, *6*, 36–43.

Buchheit, M., Lacome, M., Cholley, Y., & Simpson, B. M. (2018). Neuromuscular responses to conditioned soccer sessions assessed via GPS-Embedded accelerometers: Insights into tactical periodization. *International Journal of Sports Physiology and Performance*, *13*(5), 577–583. https://doi.org/10.1123/ijspp.2017-0045

Casamichana, D., Castellano, J., Calleja-Gonzalez, J., San Román, J., & Castagna, C. (2013). Relationship between Indicators of training load in soccer players. *Journal of Strength and Conditioning Research*, *27*(2), 369–374. https://doi.org/10.1519/JSC.0b013e3182548af1

Eisenmann, J. (2017). Translational gap between laboratory and playing field: new era to solve old problems in sports science. *Translational Journal of the American College of Sports Medicine*, *2*(8), 37–43.

Federolf, P., Reid, R., Gilgien, M., Haugen, P., & Smith, G. (2014). The application of principal component analysis to quantify technique in sports. *Scandinavian Journal of Medicine & Science in Sports*, *24*(3), 491–499. https://doi.org/10.1111/j.1600-0838.2012.01455.x

Franks, I. M., & Goodman, D. (2008). A systematic approach to analysing sports performance. *Journal of Sports Sciences*, *4*(1), 49–59. https://doi.org/10.1080/02640418608732098

Fullagar, H. H. K., McCall, A., Impellizzeri, F. M., Favero, T., & Coutts, A. J. (2019). The translation of sport science research to the field: A current opinion and overview on the perceptions of practitioners, researchers and coaches. *Sports Medicine, 49*(12), 1817–1824. https://doi.org/10.1007/s40279-019-01139-0

Gløersen, Ø., Myklebust, H., Hallén, J., & Federolf, P. (2018). Technique analysis in elite athletes using principal component analysis. *Journal of Sports Sciences*, *36*(2), 229–237. https://doi.org/10.1080/02640414.2017.1298826

Hopkins, W. G. (2010). Linear models and effect magnitudes for research, clinical and practical applications. *Sportscience*, *14*, 49–56. https://doi.org/sportsci.org/2010/wghlinmod.htm

Hunter, F., Bray, J., Towlson, C., Smith, M., Barrett, S., Madden, J., Abt, G., & Lovell, R. (2014). Individualisation of time-motion analysis: A method comparison and case report series. *International Journal of Sports Medicine*, *36*(01), 41–48. https://doi.org/10.1055/s-0034-1384547

Impellizzeri, F. M., Marcora, S. M., & Coutts, A. J. (2019). Internal and external training load: 15 years on. *International Journal of Sports Physiology and Performance*, *14*(2), 270–273. https://doi.org/10.1123/ijspp.2018-0935

Jaspers, A., Kuyvenhoven, J. P., Staes, F., Frencken, W. G. P., Helsen, W. F., & Brink, M. S. (2018). Examination of the external and internal load indicators' association with overuse injuries in professional soccer players. *Journal of Science and Medicine in Sport*, *21*(6), 579–585. https://doi.org/10.1016/J.JSAMS.2017.10.005

Johnston, R. J., Watsford, M. L., Kelly, S. J., Pine, M. J., & Spurrs, R. W. (2014). Validity and interunit reliability of 10 Hz and 15 Hz GPS units for assessing athlete movement demands. *Journal of Strength and Conditioning Research*, *28*(6), 1649–1655. https://doi.org/10.1519/JSC.0000000000000323

Jolliffe, I. T. (1986). Principal components in regression analysis. In *Principal Component Analysis* (pp. 129–155). https://doi.org/10.1007/978-1-4757-1904-8_8

Jollife, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. In *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* (Vol. 374, Issue 2065). Royal Society of London. https://doi.org/10.1098/rsta.2015.0202

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*(1), 141–151. https://doi.org/10.1177/001316446002000116

Lê, S., Josse, J., Rennes, A., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. In *JSS Journal of Statistical Software* (Vol. 25). http://www.jstatsoft.org/

Lovell, T. W. J., Sirotic, A. C., Impellizzeri, F. M., & Coutts, A. J. (2013). Factors affecting perception of effort (session rating of perceived exertion) during rugby league training. *International Journal of Sports Physiology and Performance*, *8*(1), 62–69. https://doi.org/10.1123/IJSPP.8.1.62

Luteberget, L. S., Holme, B. R., & Spencer, M. (2018). Reliability of wearable inertial measurement units to measure physical activity in team handball. *International Journal of Sports Physiology and Performance*, *13*(4), 467–473. https://doi.org/10.1123/ijspp.2017-0036

Maughan, P. C., MacFarlane, N. G., & Swinton, P. A. (2021). Relationship between subjective and external training load variables in youth soccer players. *International Journal of Sports Physiology and Performance*, *16*(8), 1127–1133. https://doi.org/10.1123/IJSPP.2019-0956

Maughan, P. C., MacFarlane, N. G., & Swinton, P. A. (2022). The influence of season phase on multivariate load relationships in professional youth soccer. *Journal of Sports Sciences*, *40*(3), 345–350. https://doi.org/10.1080/02640414.2021.1993642

McLaren, S. J., Macpherson, T. W., Coutts, A. J., Hurst, C., Spears, I. R., & Weston, M. (2018). The relationships between internal and external measures of training load and intensity in team sports: A meta-analysis. In *Sports Medicine* (Vol. 48, Issue 3, pp. 641–658). Springer International Publishing. https://doi.org/10.1007/s40279-017-0830-z

Moreno-Pérez, V., Malone, S., Sala-Pérez, L., Lapuente-Sagarra, M., Campos-Vazquez, M. A., & Del Coso, J. (2020). Activity monitoring in professional soccer goalkeepers during training and match play. *International Journal of Performance Analysis in Sport*, *20*(1), 19–30. https://doi.org/10.1080/24748668.2019.1699386

Morgans, R., Orme, P., Anderson, L., & Drust, B. (2014). Principles and practices of training for soccer. In *Journal of Sport and Health Science* (Vol. 3, Issue 4, pp. 251–257). Elsevier. https://doi.org/10.1016/j.jshs.2014.07.002

Nguyen, L. H., & Holmes, S. (2019). Ten quick tips for effective dimensionality reduction. *PLOS Computational Biology*, *15*(6), e1006907. https://doi.org/10.1371/journal.pcbi.1006907

Nosek, P., Brownlee, T. E., Drust, B., & Andrew, M. (2021). Feedback of GPS training data within professional English soccer: A comparison of decision making and perceptions between coaches, players and performance staff. *Science and Medicine in Football*, 24733938.2020.1770320. https://doi.org/10.1080/24733938.2020.1770320

Owen, A. L., Wong, D. P., Mckenna, M., & Dellal, A. (2011). Heart rate responses and technical comparison between small-vs. large-sided games in elite professional soccer. *Journal of Strength and Conditioning Research*, *25*(8), 2104–2110. https://doi.org/10.1519/JSC.0b013e3181f0a8a3

Parmar, N., James, N., Hearne, G., & Jones, B. (2018). Using principal component analysis to develop performance indicators in professional rugby league. *International Journal of Performance Analysis in Sport*, *18*(6), 938–949. https://doi.org/10.1080/24748668.2018.1528525

Richter, C., O'Reilly, M., & Delahunt, E. (2021). Machine learning in sports science: challenges and opportunities. *Sports Biomechanics*, 1–7. https://doi.org/10.1080/14763141.2021.1910334

Rojas-Valverde, D., Pino-Ortega, J., Gómez-Carmona, C. D., & Rico-González, M. (2020). A systematic review of methods and criteria standard proposal for the use of principal component analysis in team's sports science. *International Journal of Environmental Research and Public Health*, *17*(23), 8712. https://doi.org/10.3390/IJERPH17238712

Ryan, S., Kempton, T., & Coutts, A. J. (2021). Data reduction approaches to athlete monitoring in professional Australian football. *International Journal of Sports Physiology and Performance*, *16*(1), 59–65. https://doi.org/10.1123/IJSPP.2020-0083

Scantlebury, S., Till, K., Beggs, C., Dalton-Barron, N., Weaving, D., Sawczuk, T., & Jones, B. (2020). Achieving a desired training intensity through the prescription of external training load variables in youth sport: More pieces to the puzzle required. *Journal of Sports Sciences*. https://doi.org/10.1080/02640414.2020.1743047

Weaving, D., Beggs, C., Dalton-Barron, N., Jones, B., & Abt, G. (2019). Visualizing the complexity of the athlete-monitoring cycle through principal-component analysis. *International Journal of Sports Physiology and Performance*, *14*(9), 1304–1310. https://doi.org/10.1123/ijspp.2019-0045

Weaving, D., Dalton, N. E., Black, C., Darrall-Jones, J., Phibbs, P. J., Gray, M., Jones, B., & Roe, G. A. B. (2018). The same story or a unique novel? within-participant principal-component analysis of measures of training load in professional rugby union skills training. *International Journal of Sports Physiology and Performance*, *13*(9), 1175–1181. https://doi.org/10.1123/ijspp.2017-0565

Weaving, D., Jones, B., Till, K., Abt, G., & Beggs, C. (2017). The case for adopting a multivariate approach to optimize training load quantification in team sports. *Frontiers in Physiology*, *8*. https://doi.org/10.3389/fphys.2017.01024

Weaving, D., Marshall, P., Earle, K., Nevill, A., & Abt, G. (2014). Combining internal- and external-training-load measures in professional rugby league. *International Journal of Sports Physiology and Performance*, *9*(6), 905–912. https://doi.org/10.1123/ijspp.2013-0444

Weston, M. (2018). Training load monitoring in elite English soccer: a comparison of practices and perceptions between coaches and practitioners. *Science and Medicine in Football*, *2*(3), 216–224. https://doi.org/10.1080/24733938.2018.1427883

Williams, A. M., & Reilly, T. (2000). Talent identification and development in soccer. *Journal of Sports Sciences, 18*(9), 657–667. https://doi.org/10.1080/02640410050120041

Williams, B., Onsman, A., Brown, T. (2010). Exploratory factor analysis: A five-step guide for novices. *Journal of Emergency Primary Healthcare*, *8*(3), 1–13. https://doi.org/10.33151/AJP.8.3.93

Williams, S., Trewartha, G., Cross, M. J., Kemp, S. P. T., & Stokes, K. A. (2017). Monitoring what matters: A systematic process for selecting training-load measures. *International Journal of Sports Physiology and Performance*, *12*(Suppl 2), S2-101-S2-106. https://doi.org/10.1123/ijspp.2016-0337

Submitted: January 31, 2022
Accepted: May 29, 2023
Published Online First: November 8, 2023

Correspondence to:
Tom Brownlee, Ph.D.
School of Sport, Exercise and Rehabilitation Sciences
University of Birmingham
Birmingham, B15 2TT UK
Phone: +44 7775 333932
E-mail: t.brownlee@bham.ac.uk