DEPARTMENT OF SPORT AND EXERCISE SCIENCES
ASSESSMENT NUMBER: J116658
Student Number: 2403329
MODULE CODE: SS7301
ASSIGNMENT: Research Proposal

Rationale & Aim

Understanding the relationship between performance metrics and their impacts on a sporting event is a fundamental aim of performance analysts (Hughes & Bartlett, 2002). Highlighting important metrics within sport can provide coaches and athletes with important benchmarks that can be utilized for peer-to-peer comparisons and training strategies. Performance metrics in swimming are unique in their practical utility but are often interrelated based on their definition (Staunton et al., 2024a). For example, free swimming speed (FSS) consists of distance per stroke (DPS) and stroke rate (SR) during the middle section of the pool, so any change in DPS or SR would impact FSS. The dimensionality of swim race data creates a multivariate dataset that makes it difficult to differentiate relative value in terms of overall performance. This characteristic makes dimensionality techniques such as principal component analysis (PCA) appropriate as they are able reduce dimensionality while preserving as much relevant information as possible (Sewell, 2007). This aids coaches and athletes to better understand and interpret key relationships that are often complex (Weaving et al., 2019).

Previous research has been done on international level short-course races in both men and women using PCA as a method of data reduction, so any additional research on long-course or relay events would be novel (Staunton et al., 2024a, 2024b). Previous prediction models formed for the 4 x 200 freestyle event have used qualitative variables such as world ranking, pacing strategy, start lap strategy, and relay leg order as indicators of overall performance (Wu et al., 2021). While this information is valuable is certain circumstances, predicting a relay team's placing based off their current world standings provides no insight into how they achieved their world standings in the first place or how they could improve their performance.

The men's 4x100 freestyle relay is the fastest of all relays and is well known for its close finishes. The 2024 Olympic relay final had an average of 0.55 seconds in-between placings, so it is easy to understand that every millisecond matters, especially when any difference has the

possibility to be compounded by a team of 4. Given the context this events importance, as well as the limited amount of research in the area. This proposed study would aim to record relevant continuous variables within the men's 4 x 100 freestyle event, conduct a statistical analysis using PCA and a stepwise linear regression to produce a prediction model from the b coefficients. The prediction model will then be tested via cross-fold validation and against out of sample data. The aim is to create a prediction model that can then be used as a training tool for coaches and athletes, similar to that of previous research (Staunton et al., 2024b, 2024a; Wu et al., 2021).

Statistical Design

This retrospective analysis starts by analysing video data from 2 men's 4x100m freestyle finals heats from international level meets. The video(s) and analysis software are being provided by Aquatics GB.

Data collection starts by analysing the provided video(s) in the bespoke program exclusive to Aquatics GB, named NEMO. It should be noted that only analysts certified in the program are able to complete this process. Additionally, NEMO has an intra-analyst error flagging system that compares newly entered data to data of its peer group based on competition level. Any metric that is outside the groups' mean by more than 1 standard deviation is flagged for an analyst to review given the context of the athlete's position within the competition level. 12 variables will be gleaned from the video analysis and are defined in the *Video Analysis Definitions* section of the appendix. Location-time variables are recorded when the centre of an athletes' head passes the appropriate distance. Using a calibrated map of the pool, distances are calculated via the coloured rungs of the lane ropes to approximately 0.1m, detailed examples can be found in the *Video Analysis Definitions* section of the appendix.

Statistical design consists of 5 steps, with the main analyses being Principal Component Analysis (PCA) and stepwise multiple linear regression. A procedural list: *Statistical Procedures* is in the appendix. Statistical procedures will be performed in R. First, a correlation matrix will be conducted and any variables that have low correlations will be removed from the dataset. Second, PCA will then be conducted, keeping principal components with eigenvalues that are greater than 1. A rotated component matrix with Varimax rotation will be used to determine the loading for each of the principal components, similar to that of other research (Colyer et al., 2017; Jensen et al., 2023, 2023; Staunton et al., 2024b, 2024a). The variable(s) with the strongest loading on each of the principal component(s) will be retained for subsequent analysis. Third, a stepwise multiple linear regression will be conducted. The regression equation(s) using the variable(s) identified in the PCA will then be retained. Standard estimate of error (SEE) will be calculated for the regression equation. The regression equation will then undergo a k-fold cross analysis to assess the models accuracy and stability (Colyer et al., 2017). SEE will be calculated for the k-fold model and compared to that of the original. The correlation between k-fold predicted time and actual time will be calculated and compared with the R^2 value of the original multiple regression, R^2 differences should not exceed 0.10 (Kleinbaum et al., 2013; Rencher, 2002). Lastly, 95% limits of agreement and 95% confidence interval (CI) will be calculated for the two models. To check the accuracy of the model, predicted times will then be calculated from out-of-sample data and compared to the out-of-sample results. This will determine the model's suitability to be used as a tool for end users.

Participants

This study will be selecting male athletes who have competed in the 4x100m freestyle relay finals at the 2023 World Aquatic Championships and 2024 Olympic Games. The

starting athlete in a relay team is the only team member with a standard start position. Each subsequent athlete after the starter, implements a flying start, also known as a relay takeover. Previous research has shown differences in emersion time, reaction time, and start15 time between an athletes' relay takeover start and their individual start (Qiu & Calvo, 2020). Considering these differences, the sample will be separated into two groups "starters" and "non-starters". The rule-of-thumb for sample size requirements in PCA has been a sample-to-item ratio of 10:1 (Costello & Osborne, 2005). However, strict rules have diminished over the years, and it is generally recognized that a sample can be adequate if it has high communalities that are uniform, does not have cross loadings, and has several variables loaded strongly on each factor (Costello & Osborne, 2005). Similar studies have used sample sizes of 29-74 (Staunton et al., 2024a, 2024b). Authors also reported R^2 values, which allows for the calculation of Cohen's f^2 (Selya et al., 2012; Staunton et al., 2024a). Given the effect size calculation and the known variables that will be used in this study, we are able to estimate an adequate sample size using G*power (version 3.1.9.7**)**. 5 R^2 values were reported, and G*power sample size estimates range from 16-23, more information is available in the *G*power Calculations* section of the appendix.

The two events that meet the inclusion criteria would provide a total of 64 participants (*starters* = 16, *non-starters* = 48). The proposed sample meets the recommendations mentioned even with the maximum variables, however, a KMO test of sampling adequacy will still be performed to determine if the sample size is reliable.

Ethical Issues

The proposed study plans to access two different sources of data: official meet data and video of the events. Official meet data are publicly available through the World Aquatics website. Access to the video data in question is in the ownership of Aquatics GB, and is

maintained safely on a private, protected database. A representative of Aquatics GB indicated that a formal request is required for access to the video data. Once the request is approved, it will be included in the appendix. Data files produced during the proposed study will be secured on a private, protected device in compliance with the UK Data Protection Act. The analysis program, NEMO, contains information that is accessory to that of the proposed study. However, the program is password-protected, not publicly available, and only users approved by the organisation may be able to access any information. Names or nationality of athletes will not be included in the finalised dataset and will instead be given a pseudonym in the form of a number.

Disclosure Statement

There are no potential conflicts of interest to report.

Potential Benefits

The proposed study aims to provide key stakeholders with relevant KPIs for the men's 4 x 100 freestyle relay and a practical formula that may help with predicting performances based off the relevant KPIs. Effectively, this tool would show the impact of the selected KPIs on finish time along with 95% confidence intervals. Additionally, by comparing the sample groups of "starters" and "non-starters", we would be able to better understand the underlying variables that separate the roles. Lastly, similar research to this has only been done in short-course individual races, so novel insights in long-course relay swimming may be gained.

References

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, *8*(2), 135–160. https://doi.org/10.1177/096228029900800204

Colyer, S. L., Stokes, K. A., Bilzon, J. L. J., Cardinale, M., & Salo, A. I. T. (2017). Physical Predictors of Elite Skeleton Start Performance. *International Journal of Sports Physiology and Performance*, *12*(1), 81–89. https://doi.org/10.1123/ijspp.2015-0631

Costello, A. B., & Osborne, J. (2005). *Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis*. https://doi.org/10.7275/JYJ1-4868

*G*Power*. (n.d.). Retrieved February 21, 2025, from https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower

Hughes, M. D., & Bartlett, R. M. (2002). The use of performance indicators in performance analysis. *Journal of Sports Sciences*, *20*(10), 739–754. https://doi.org/10.1080/026404102320675602

James Cook University. (2023). *Factor Analysis*. https://www.jcu.edu.sg/__data/assets/pdf_file/0020/2067320/Factor-Analysis.pdf

Jensen, M., Stellingwerff, T., Pollock, C., Wakeling, J., & Klimstra, M. (2023). *Can principal component analysis be used to explore the relationship of rowing kinematics and force production in elite rowers during a step test? A pilot study*. https://doi.org/10.3390/make5010015

Jolliffe, I. (1990). Principal component analysis: A beginner's guide - I. Introduction and application. *Weather*, *45*, 375–382. https://doi.org/10.1002/j.1477-8696.1990.tb05558.x

Kleinbaum, D. G., Kupper, L. L., Nizam, A., & Rosenberg, E. S. (2013). *Applied Regression Analysis and Other Multivariable Methods*. Cengage Learning.

Laerd, S. (2018). *How to perform a principal components analysis (PCA) in SPSS Statistics | Laerd Statistics*. Laerd Statistics. https://statistics.laerd.com/spss-tutorials/principal-components-analysis-pca-using-spss-statistics.php

Nkansah, B. K. (2018). *On the Kaiser-Meier-Olkin's Measure of Sampling Adequacy*.

Qiu, X., & Calvo, A. L. (2020). *COMPARISON OF THE SWIMMING START PERFORMANCE BETWEEN INDIVIDUAL AND RELAY FREESTYLE RACES*.

Rencher, A. C. (2002). *Methods of multivariate analysis* (2nd ed). J. Wiley.

Selya, A. S., Rose, J. S., Dierker, L. C., Hedeker, D., & Mermelstein, R. J. (2012). A Practical Guide to Calculating Cohen's f2, a Measure of Local Effect Size, from PROC MIXED. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00111

Sewell, M. (2007). *Principal Component Analysis*.

Shrestha, N. (2021). Factor Analysis as a Tool for Survey Analysis. *American Journal of Applied Mathematics and Statistics*, *9*(1), Article 1. https://doi.org/10.12691/ajams-9-1-2

Staunton, C. A., Romann, M., Björklund, G., & Born, D.-P. (2024a). Diving into a pool of data: Using principal component analysis to optimize performance prediction in women's short-course swimming. *Journal of Sports Sciences*, *42*(6), 519–526. https://doi.org/10.1080/02640414.2024.2346670

Staunton, C. A., Romann, M., Björklund, G., & Born, D.-P. (2024b). Streamlining performance prediction: Data-driven KPIs in all swimming strokes. *BMC Research Notes*, *17*, 52. https://doi.org/10.1186/s13104-024-06714-x

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

Wu, P. P.-Y., Babaei, T., O'Shea, M., Mengersen, K., Drovandi, C., McGibbon, K. E., Pyne, D. B., Mitchell, L. J. G., & Osborne, M. A. (2021). Predicting performance in 4 x 200-m freestyle swimming relay events. *PloS One*, *16*(7), e0254538. https://doi.org/10.1371/journal.pone.0254538

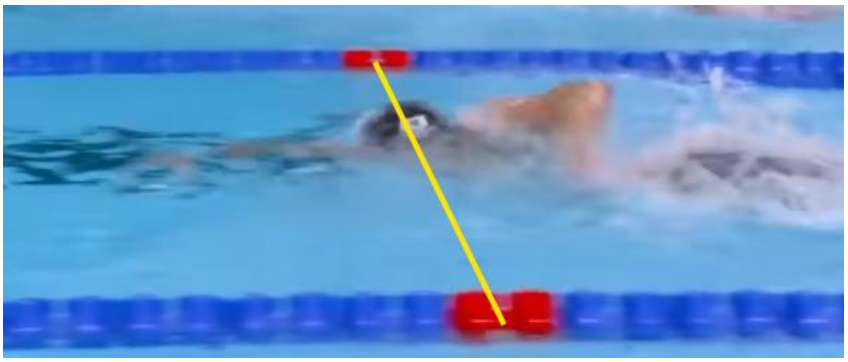
Appendices

Video Analysis Definitions

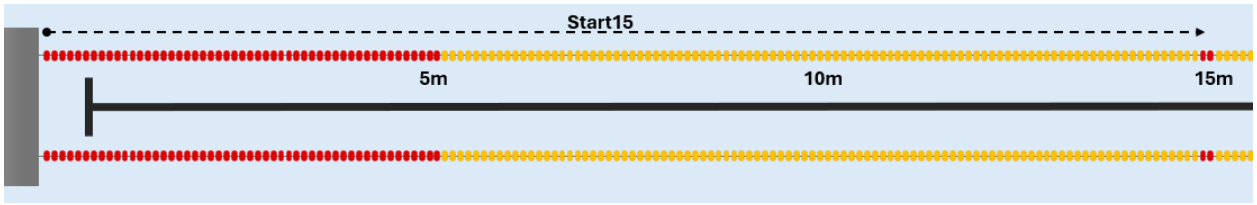
Note: All distance measurements (#4-9) use the centre of the athletes' head as the reference point.

1. Stroke Count:
    a. Number of stokes per length (50m)
2. Stroke Rate:
    a. $\frac{Stroke\ count}{60\ sec}$
3. Distance Per Stroke:
    a. $\frac{Stroke\ count}{50m}$
4. Free Swimming Speed:
    a. $\frac{45m - 15m}{Split\ time - (Out15 + In5)}$ = x m/sec

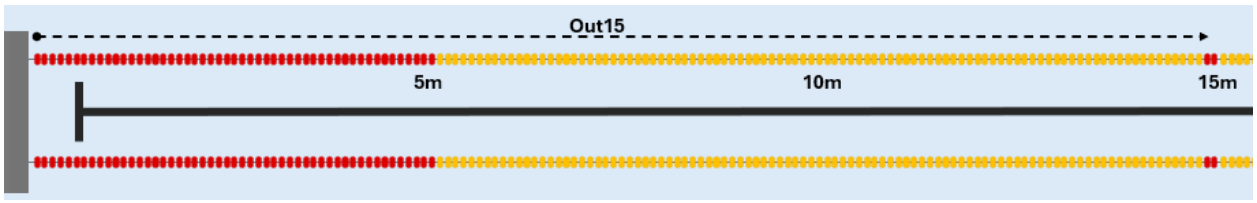
The following variables are determined when the centre of the athletes head crosses the respective distances. Example below



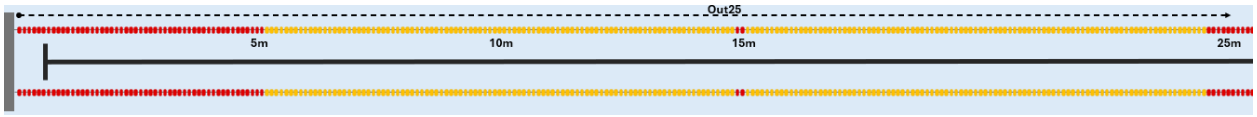
5. Start15:
    a. The time from the start gun to the 15m mark (15m is recognised by the change in lane rope colour)



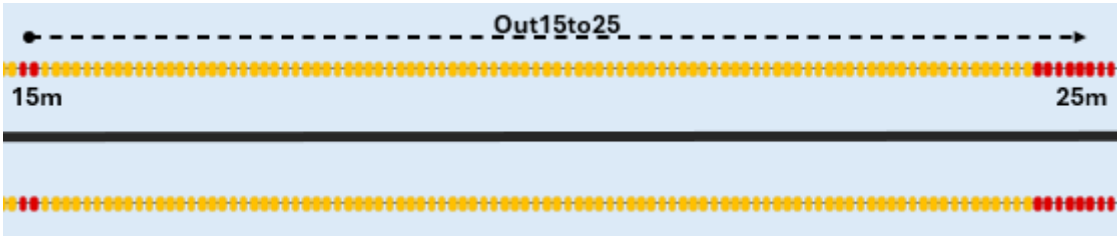
6. Out15:
    a. The duration of time from the wall to the 15m mark (15m is recognised by the change in lane rope colour)



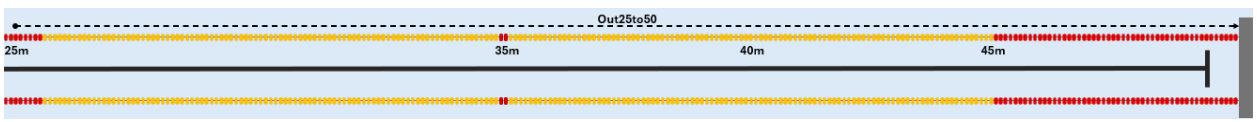
7. Out25:
    a. The duration of time from the wall to the 25m mark (25m is recognised by the change in lane rope colour)



8. Out15to25:
    a. The time difference between Out25 and Out15



9. Out25to50:
    a. The time difference between Out50 and Out25



10. In5:

a. The duration of time 5m prior to the wall (5m is recognised by the flags as well as the change in lane rope colour)

11. Rotate:

a. Measured from the last stroke to the split time

12. Breakout:

Measured when the swimmer's hand from the first stroke exits the water

1. Time:
    i. The time from the wall to breakout
2. Distance:
    i. Distance at which the breakout occurs

Statistical Procedures

1. Correlation Matrix
    a. Values outside of 2.5Std will be removed from the dataset prior to calculation
    b. Only keep variables with high correlations (>0.6)
        i. variables must have strong linear relationships for the PCA
2. Principal Component Analysis
    a. Kaiser-Meier-Olkin Measure of Sampling Adequacy
        i. Ensures sampling adequacy has been met
        ii. Similar research has used >0.5 (Staunton et al., 2024b, 2024a)
        iii. Others have suggested that a level >0.6 is acceptable with >0.8 being satisfactory (Nkansah, 2018; Rencher, 2002)
    b. Bartlett's test of Sphericity ($p<0.05$)
        i. Determines suitability for data reduction
        ii. A significance of $p < 0.001$ will be used (James Cook University, 2023; Shrestha, 2021).
    c. Rotated component Matrix (Varimax)
        i. Varimax has been previously used in similar study designs within exercise science (Colyer et al., 2017; Jensen et al., 2023, 2023; Staunton et al., 2024b, 2024a)
    d. PCs with eigenvalues > 1 are extracted
        i. The most heavily loaded variable to each component will then be retained along with the original variables which did not display a high degree of covariance
        ii. All variables will then be used in a stepwise multiple linear regression analysis with swim time as the criterion
3. Stepwise Linear regression analysis
    a. 4 Main assumptions must be met
        i. Linearity
            1. Refer to Durbin-Watson test ($p< 0.05$)
        ii. Independent Residuals
            1. Test via Residual vs Fitted plot
            2. Residual error mean should be around 0
        iii. Equal Variance
            1. Test via Scale-Location plot
            2. Residual points should be equally spread around the line of best fit
        iv. Residual errors have constant variance
            1. Test via Residuals vs Fitted values plot
            2. Values should show equal variance throughout the plot
    b. Unstandardized *B* coefficients from the linear regressions will then be used to form prediction equation(s)
    c. SEE and CI will be calculated for the prediction equation
4. *k*-fold cross-validation

a. select *k*-value based on parameter tuning
    i. Parameter tuning iterates through *k*-values, the model with the highest reported accuracy will be used
b. Provides a rigorous assessment of model stability (Colyer et al., 2017)
c. Prediction errors will be calculated for k iterations and combined to form an overall standard error of the estimate (SEE)
d. SEE of the original multiple regression analysis will be compared with the prediction model SEE
e. The correlation between predicted and actual time will be computed and compared with the R^2 value of the multiple regression (R^2 differences should not exceed 0.10)(Kleinbaum et al., 2013)

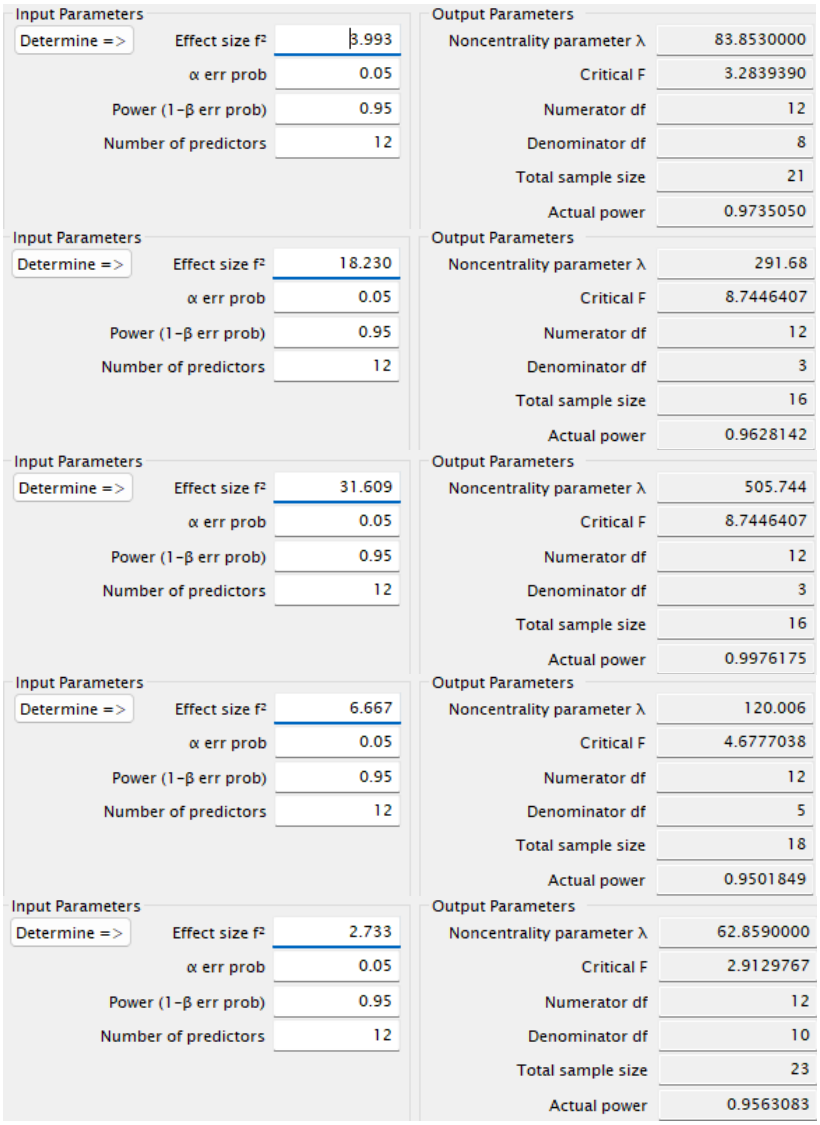
5. LOA and CI
    a. The predicted and actual swimming performances, along with 95% limits of agreement and 95% confidence intervals of the LOAs will be analysed via methods described by (Bland & Altman, 1999)

G*Power Calculations

The R^2 values are from similar research (Staunton et al., 2024a).

	50m	100m	200m	400m	800m
R^2	0.941	0.997	0.999	0.978	0.882
f^2	3.993	18.230	31.609	6.667	2.733
Gpower Sample size	21	16	16	18	23

Cohen's f (f^2) $= \sqrt{\frac{R^2}{(1-R^2)}}$



Informed consent

Title of Project: Performance Prediction in Men's long-course 4x100m Freestyle Relay

Name of Researcher: Student #2403329

Please initial box

1. I confirm that I have read and understand the information sheet for the above study and have had the opportunity to ask questions. ☐
2. I understand that my participation is voluntary and that I am free to withdraw at any time, without giving any reason and without my legal rights being affected. ☐
3. I consent to being tape recorded / video recorded (if relevant) ☐
4. I agree to take part in the above study. ☐

_____	_____	_____
Name of Participant	Date	Signature

_____	_____	_____
Researcher	Date	Signature

1 for participant; 1 for researcher

Participation information sheet

Participant Information Sheet

Performance Prediction in Men's long-course 4x100m Freestyle Relay

Thank you for reading this.

You are being invited to take part in a research study. Before you decide, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part.

What is the purpose of the study?

This study is being conducted on male athletes who performed in the 4x100m freestyle finals in the 2023 World Aquatics Championships and/or the 2024 Olympic Games. The purpose of this study is to analyse the impact of individual variables within the roles of "starters" and "non-starters".

The athletes are separated into two groups as there are key differences in the roles. Starting athletes have a standard block start where the hands are in contact with the block prior to the starting gun. Non-starting athletes perform a relay takeover start whereby they are not required to have their hands in contact with the block. Additionally, there is no starting gun, as they start when their teammate contacts the wall. The main goal of this study is to identify the most relevant variables to each of the roles so that they may be highlighted for a coach or athlete's consideration.

Why have I been chosen?

You have been chosen because you have competed in the men's 4x100m freestyle relay in the 2023 World Aquatics Championships and/or the 2024 Olympic Games.

Do I have to take part?

It is up to you to decide whether or not to take part. If you decide to take part, you will be given this information sheet to keep and be asked to sign a consent form. If you decide to take part, you are still free to withdraw and without giving a reason. A decision to withdraw, or a decision not to take part, will not affect you in any way.

What will happen to me if I take part?

As you have already participated in the preciously listed event(s), you have completed everything that needs to be done.

What are the possible disadvantages and risks of taking part?

There are no disadvantages or risks foreseen in taking part in the study.

What are the possible benefits of taking part?

Relevant variables gleaned from the analysis will be shared with you and your coaching staff. This may help support philosophies regarding your training, which in turn, may help you improve your performance.

What if something goes wrong?

If you wish to complain or have any concerns about any aspect of the way you have been approached or treated during the course of this study, please contact the Dean of the Faculty of Health, Medicine and Society, University of Chester, Parkgate Road, Chester, CH1 4BJ, 01244 511000.

The University does not accept liability for harm which does not result from its negligence. In the event that something does go wrong and a participant is harmed during the research and the harm sustained is due to the negligent acts of those undertaking the research, then the participant may have grounds to bring legal action. Anyone bringing such legal action may incur legal costs.

Will my taking part in the study be kept confidential?

All information which is collected about you during the course of the research will be kept strictly confidential so that only the researcher carrying out the research will have access to such information.

Participants should note that data collected from this project may be retained and published in an anonymised form. By agreeing to participate in this project, you are consenting to the retention and publication of data.

What will happen to the results of the research study?

The results will be written up into a dissertation for my final project of my MSc. Individuals who participate will not be identified in any subsequent report or publication.

Who is organising the research?

The research is conducted as part of a MSc in Sprots Performance Analysis within the Department of Sports Performance Analysis at the University of Chester. The study is organised with supervision from the department, by *Student #2403329*, an MSc student.

Who may I contact for further information?

If you would like more information about the research before you decide whether or not you would be willing to take part, please contact:

Student #2403329. 2403329@chester.ac.uk.

Thank you for your interest in this research.

Risk assessment form

Ref No:	**Date**: 20/02/2025	**Review Date**:	**Assessor/s:**	**Assessors Signature**:

Description of task to be assessed:	**Area or Dept**:	
	Persons Exposed (e.g. employee, contractor, public etc)	

Ref	**Hazard & Potential Harm**	**Existing Risk Control Measures**	**Level of Risk**			**Additional control measures**	**Completion date**
			Prob-ability	Seve-rity	Risk Score		
1	Using computer or workstation incorrectly. Repetitive strain injury and back injury	1. Carry out full DSE workstation assessment. 2. Ensure corrective actions implemented.	2	2	4	Researcher has been provided with an ergonomic chair and appropriate monitor arms to ensure proper posture	20/02/2025