Issues in the application of statistical techniques in sport and exercise science

David R Mullineaux

PhD

2002

Issues in the application of statistical techniques in sport and exercise science

David Ross Mullineaux

Published works submitted in partial fulfilment of the requirements of Sheffield Hallam University for the degree of Doctor of Philosophy on the basis of published work

May 2002

Acknowledgements

Oh, where to begin? Initially I worked on a full time PhD under the supervision of Roger Bartlett and Carl Payton and I must thank them for the great deal that I learnt. Liz Barnes, probably unknowingly, started me on the PhD via published work so I must thank her in part for this. My thanks to my friend and former colleague Mark Scott for the thought-provoking discussions and for the inspiration of my alternative title 'Statistics are not as bad as you think'. Thanks are belatedly extended to Suddy, Mr B, Scoot and Carlos. So as not to unintentionally miss anyone this time I extend my thanks to all my co-authors, colleagues, participants, supporting organisations, friends, family and others for the explicit and implicit support of my research. Particular thanks to Chris Barnes and Peter Walder who as office friends were invaluable for solving occasional dilemmas. In my new office I thank Tony Oldham, John Pates and Ian Renshaw for their empathy in sharing the exhilaration of nearing the end – who will be first? Their research in motor control and sport psychology topics may interest someone, says the pot to the kettle! As I write this draft I wish to thank the proof-readers Helen, Stuart and John, even though they don't know it yet, and more importantly for the walks, bike rides, DIY adventures, food and beer outings and all the other things they have still to do! Penultimately, again thanks to my family for the important things: our gatherings, sharing presents, amusing gossip, hard working dish washers and vino. All of these things keep me sane despite any transient denial. Lastly, thanks to John Alderson for supporting me in my work and for commenting on this critical synopsis of my published works.

Abstract

The aim of this research is to demonstrate the benefits and limitations of selected techniques used to analyse data derived from sport and exercise science research. Although statistical techniques are easy to access through software packages, supporting literature about their appropriate application is less common. Many researchers are unaware of the full benefits or potential pitfalls when using these techniques. An understanding of the appropriate use of statistics will benefit the researcher by maximising the potential for analysing, interpreting and applying data correctly. Furthermore, it will minimise wasted effort or dissemination of inaccurate information through incorrect analyses. In this thesis examples are derived from fifteen published articles based on five topics that illustrate the appropriate use of particular statistical techniques. Firstly, the use of 'agreement' and 'least-products-regression' as appropriate techniques for comparing repeated measures are demonstrated (e.g. Mullineaux *et al.*, 1999). Both techniques revealed that over two separate days the peak-torque-extension of the knee of healthy females is unreliable. Secondly, the use of 'allometric' scaling of body size differences that should allow for meaningful comparisons between participants' measurements is explored (e.g. Batterham, George and Mullineaux, 1997). Results showed that left ventricular mass is related to fat free mass to the power of 1.07 (0.92 to 1.22; 95% CI). Thirdly, mathematical modelling is used to explore a theory that would be difficult to test empirically (e.g. Payton, Hay and Mullineaux, 1997). Results revealed that body roll contributes substantially to the propulsive force in front crawl swimming. Fourthly, logistic regression is used to predict group membership from the combined effect of several independent variables (e.g. Mullineaux *et al.*, 2001a). It was found that the likelihood of participation in adequate physical activity to promote health can be strongly predicted from six variables. Lastly, in an invited review paper, key features in the application of research methods and statistics in biomechanics and motor control are highlighted (e.g. Mullineaux *et al.*, 2001b). These published papers form a body of work that will facilitate a greater and more appropriate use of selected statistical techniques in sport and exercise science.

Contents

Preliminary pages

Critical synopsis

Appendices of published papers

2. Mullineaux, D.R., Barnes, C.A. and Batterham, A.M. (1999). Assessment of bias in comparing measurements: a reliability example. *Measurement in Physical Education and Exercise Science*, **4**, 195-205.

3. Mullineaux, D.R., Scott, M.A. and Batterham, A.M. (1994). The use of "agreement" for assessing digitiser reliability/objectivity. In *Proceedings of the Biomechanics Section of the British Association of Sport and Exercise Sciences* (edited by J. Watkins), pp. 17-20. Leeds: BASES.

Allometric scaling

4. Batterham, A.M., Barnes, C.A. and Mullineaux, D.R. (1999a). Modelling the influence of thigh muscle cross-sectional area on vertical jump performance in young professional soccer players. *Journal of Sports Sciences*, **17**, 807-8.

5. Batterham, A.M., George, K.P. and Mullineaux, D.R. (1995). Relationship between heart size and body dimensions: an allometric scaling approach. *Medicine and Science in Sports and Exercise*, **27**, S158.

6. Batterham, A.M., George, K.P. and Mullineaux, D.R. (1997). Allometric scaling of left ventricular mass and body dimensions in males and females. *Medicine and Science in Sports and Exercise*, **29**, 181-6.

7. Mullineaux, D.R., Barnes, C.A. and Batterham, A.M. (1998a). Allometric scaling of anaerobic performance and the use of 'bootstrapping' for statistical inference from a small sample. In *Proceedings of the 2nd Australasian Biomechanics Conference* (edited by P. Hume, R. Marshall, P. Hunter, S. Stanley, I. Anderson and P. McNair), pp. 72-3. Auckland: The University of Auckland.

Mathematical modelling

8. Payton, C.J., Hay, J.G. and Mullineaux, D.R. (1995). The effect of body roll on hand velocity in the front crawl. In *Proceedings of the XVth Congress of the International Society of Biomechanics* (edited by K. Häkkinen, K.L. Keskinen, P.V. Komi and A. Mero), pp. 720-1. Jyväskylä: University of Jyväskylä.

9. Payton, C.J., Hay, J.G. and Mullineaux, D.R. (1997). The effect of body roll on hand speed and hand path in the front crawl - a simulation study. *Journal of Applied Biomechanics*, **13**, 300-15.

10. Payton, C.J. and Mullineaux, D.R. (1996). Effect of body roll on hand velocity in freestyle swimming. In *Biomechanics and Medicine in Swimming VII* (edited by J.P. Troup, A.P. Hollander, D. Strass, S.W.

Trappe, J.M. Cappaert and T.A. Trappe), pp. 59-63. London: E & FN Spon.

Logistic regression

11. Mullineaux, D.R. and Barnes, C.A. (1998). Determinants of propensity to partake in adequate physical activity to promote health. *Journal of Sports Sciences*, **16**, 394-5.

12. Mullineaux, D.R., Barnes, C.A. and Barnes, E.F. (2001a). Factors affecting the likelihood to engage in adequate physical activity to promote health. *Journal of Sports Sciences*, **19**, 279-88.

Review papers

13. Mullineaux, D.R. and Bartlett, R.M. (1997). Research methods and statistics. In *Biomechanical Analysis of Movement in Sport and Exercise* (edited by R.M. Bartlett), pp. 81-104. Leeds: BASES.

14. Mullineaux, D.R., Bartlett, R.M. and Bennett, S. (2001b). Research methods and statistics in biomechanics and motor control. *Journal of Sports Sciences*, **19**, 739-60.

15. Nevill, A.M., Atkinson, G. and Mullineaux, D.R. (2001). Editorial: new horizons in research methods. *Journal of Sports Sciences*, **19**, 737-8.

List of figures and tables

Figure

Table

Nomenclature

95% CI	95% confidence intervals
a	Constant (e.g. intercept in a linear regression; constant in non-linear scaling)
ANOVA	Analysis of variance
b	Constant (e.g. gradient in linear regression; power in non-linear scaling)
BM	Body mass
c	Constant (e.g. constant for a second variable in non-linear multiple scaling)
d	Constant (e.g. power for a second variable in non-linear multiple scaling)
δ	Mean of the differences between two repeat measures
e	The natural logarithm base number
FFM	Fat free mass
k	Number of groups
ln	Conversion of the base number from a decimal to a natural logarithm
LOA	Limits of agreement
LPR	Least-products regression
LVM	Left ventricular mass
MS_B	Between-subjects variance
MS_R	Within-subjects error/residual variance
MS_W	Within-subjects variance
R	Intraclass correlation
SE_E	Standard error of the estimate
sex	A dummy variable for sex where 0 is for males and 1 is for females
σ	Standard deviation of the differences between two repeat measures
$\dot{V}O_{2max}$	Maximal oxygen uptake
$\bar{x}$	Mean value of x
x	Variable. Typically the independent variable
x_1	Variable. Typically an independent variable in multiple regression
x_2	Variable. Typically an independent variable in multiple regression
$\bar{y}$	Mean value of y
y	Variable. Typically the dependent variable

Introduction

Sport and exercise science contributes to many areas of life including education, leisure, sport and health. At the end of the last millennium this area experienced rapid growth in the UK. In education, between 1997 and 2000, undergraduate students enrolled on Sports Science categorised courses increased by 46% to 6848 (UCAS, 2001). During this same period, in leisure, consumer spending on active sport was forecast to increase by 11.2% to nearly £6 billion (Gratton *et al.*, 1999, p. 54). In sport, National Lottery funding enables Sport England (1999) to promote sport for everyone via the Community Projects funding and to increase international recognition via the World Class funding. In health, the Government's white paper 'Saving Lives: Our Healthier Nation' details aims to reduce substantially death rates from cancer, coronary heart disease, stroke, accidents and mental ill-health (Health Development Agency, 1999). Promoting physical activity has been recognised within the white paper as an important means to achieve such aims. This growth in education, leisure, sport and health related activity demands investigation and evaluation through applied research, a significant proportion of which will lie within the scope of sport and exercise science and its associated fields.

Research in sport and exercise science will only be useful if it is well planned, conducted, analysed and reported (Mullineaux and Bartlett, 1997). Technological advances have enhanced the potential of each of these facets of the research process. There is greater access to information to plan research (e.g. via the World Wide Web), improved technology to collect data (e.g. automated on-line motion capture systems), increased computational power to analyse data (e.g. statistics software) and readily available software to report research (e.g. word processing packages). However, it is the improved access to comprehensive and advanced 'black box' statistics software packages that offers the greatest potential for misuse. This access provides a means to overcome the difficulty of performing the mathematical calculations, but it does not routinely provide information on the assumptions required for valid analyses. The 'black box' software also offers the user a wealth of statistical techniques that they would not necessarily have previously encountered.

There are many similar views on the definition of statistics. In general, *statistics* are "a mathematical technique by which data are organised, treated, and presented for interpretation and evaluation" (Vincent, 1999, p. 282). A specific application of predicting population parameters is referred to as *inferential statistics* which "consist of a set of statistical techniques that provide predictions about population characteristics based on information in a sample from that population" (Field, 2000, p. 4). The overall application of statistics provides benefits that include the ability to explore, synthesise and summarise data efficiently, simply and objectively. Statistics can also provide the scientific rigour and objectivity (Matthews, 1998) that 'gatekeepers' such as journal editors and funding bodies often seek (Mullineaux *et al.*, 2001b). Despite such potential benefits, data should only be analysed via an appropriate statistic but the appropriateness of a statistic is not always obvious. The original rationale for the development of individual tests, the subsequent mathematics incorporated and the effects of different data properties must be considered before researchers apply statistical techniques within the context of their research.

The growth in sport and exercise science, the increasing availability of powerful 'black box' statistics packages, the desire of journal editors for 'statistical significance' or 'objectivity' and researchers' frequent lack of statistical expertise have led to instances where statistics have been inappropriately applied. This can produce inaccurate and misleading findings. For example, the *t*-test, originating from the work by Student (1908), can be used to assess whether chance can be discounted as an explanation for any differences between two mean scores. The *t*-test does not provide any indication of random variation between tests. However, Atkinson and Nevill (1998) identified that sixteen out of seventy studies at the forty-third American College of Sports Medicine conference had inappropriately used the *t*-test, or equivalent test, to assess reliability or validity.

A potential way to reduce the incidence of inappropriate statistical analyses is to publish more supporting literature on appropriate uses. Hence, the aim of this research is to provide supporting literature on the appropriate use of less well known, or developing, statistical tests in sport and exercise science. In this thesis five topics incorporating statistical techniques have been selected. Firstly, as measurement tools should be reliable as a prerequisite for validity, techniques of testing repeated measures reliability are explored (e.g. Mullineaux *et al.*, 1999). The next three topics are concerned with selecting statistical analyses that are appropriate for the data based on a suitable underpinning theory. Allometric scaling is explored for testing the relationship between body dimensions and performance measures (e.g. Batterham, George and Mullineaux, 1997). Mathematical modelling is illustrated as a statistical technique for investigating theoretical propositions that would be difficult to provide support from empirical data (e.g. Payton, Hay and Mullineaux, 1997). Logistic regression as a technique for analysing relationships in data of a nominal or mixed level of measurement is illustrated (e.g. Mullineaux *et al.*, 2001a). Lastly, to promote further appropriate applications of statistical techniques in specific disciplines consideration is made of the benefits of review articles (e.g. Mullineaux *et al.*, 2001b). A brief review of each of these topics includes an explanation of the contribution of the example published papers contained in Appendices 1 to 15 for demonstrating the application of the statistical techniques or highlighting their contribution to the body of knowledge. In addition, in these topics some clarification for future directions in research concerned with the appropriate use of statistics in sport and exercise science is offered.

Repeated measures reliability

Two important characteristics of a test or method of measurement are validity and reliability. Validity relates to the degree to which a test or instrument measures what it purports to measure. Reliability relates to the consistency and dependability of the measures. It can be inherently difficult to assess validity, but a sound research design evolved from theory can minimise uncertainty and ensuring reliability can further support validity.

Reliability needs to be carefully considered at the outset of research so that it is defined and tested appropriately within the study. For instance, 'alternative forms' reliability involves using different tests to measure the same construct. Several other methods of reliability also exist, such as, 'repeated measures' reliability. This method is used to assess whether repeat measurements of the same response are reproducible (Sale, 1991). This can involve testing whether the measures obtained on two separate days, or by two testers, or with two instruments, are reproducible and thus reliable. Note that the research design used to assess repeated measures reliability should be considered with care. For example, if the number of days between repeat measures is too few, then an interaction of the *pre*-test with the *post*-test may influence the results.

Reliability is essential for validity. The tests available for appropriately assessing reliability are dependent on such aspects as the working definition of reliability, the number of repeat measures and level of measurement of the data. In this thesis the focus is on reproducible results where repeated measures reliability is most appropriate. Several tests are used for assessing repeated measures reliability (see Table 1). There may be some alternative names for repeated measures reliability, such as, 'test-retest reliability' or 'stability'. Often more than one test is available, where the choice of which to use is dependent on such things as the preferences of researchers, supervisors and journal editors on the theory underpinning the test, the statistical assumptions and the type of data output. Although different authors critique the appropriateness of these tests it is generally accepted that the majority of alternative tests not listed in Table 1 are inappropriate (e.g. Pearson product moment correlation). In addition, Table 1 refers to reliability when assessing groups of participants, but when the data relate to a single participant then alternative tests are available. For example, when several measures at the interval or ratio level of measurement are obtained for a single person, then the percentage coefficient of variation is suitable to test repeated measures reliability (Sale, 1991; Atkinson and Nevill, 1998).

The reliability coefficient is the ratio of the true measurement variance to the observed measurement variance. True measurement variance is obtained by subtracting the error variance from the observed measurement variance. This can be calculated using interclass (e.g. Pearson product moment correlation) or intraclass correlations. Interclass correlation is inappropriate for repeated measures reliability as it requires independent variables, is limited to two sets of measurements and does not detect systematic errors (Thomas and Nelson, 2001). These limitations are overcome by using intraclass correlations (*R*), and these are commonly used to assess reliability.

Table 1. Tests for the assessment of repeated measures reliability appropriate for different levels of measurement.

Level of measurement	**2 repeat measures**	**2+ repeat measures**
Nominal	Proportion of agreement[a]	
	Kappa coefficient	
Ordinal	Cronbach's alpha	Cronbach's alpha
Interval/Ratio	Limits of agreement[b,c,d]	Agreement boundary[e]
	Intraclass correlation	Intraclass correlation
	Least products regression	

Note, for tests of a nominal level of measurement see, for example, Robson (1993, p. 222) and for all other tests see text for references. Modifications to these tests include:

[a] Extend the test and use bootstrapping to obtain confidence intervals for the coefficient (Wilson and Batterham, 1999).

[b] Use with the natural logs of the data to account for violations of assumptions (Nevill and Atkinson, 1997).

[c] Account for the degrees of freedom using the *t*-distribution instead of 1.96 (Hopkins, 2000).

[d] More informative than the similar Technical Error of Measurement (Nevill and Atkinson, 2001: also known as the Method error, or if presented as a ratio of the mean the Test-retest coefficient of variation, Sale, 1991).

[e] Extension of the Standard Error of Measurement to provide 95% confidence (BSI, 1979).

When the measurement level is interval or ratio then reliability can be calculated in several ways. The 'reliability coefficient', 'limits of agreement' (LOA) and 'least-products regression' (LPR; Ludbrook, 1997) are appropriate techniques. LPR is also known as geometric mean regression (Ricker, 1973).

Analysis of variance (ANOVA) is often helpful for calculating the reliability coefficient, or specifically, intraclass correlation. A simple version of the intraclass correlation is the Average Measure R (Equation 1; McGraw and

Wong, 1996), where, the variances are separated into components for within-subjects (MS_W) and between-subjects (MS_B) effects.

$$\text{Average Measure } R = (MS_B - MS_W) / MS_B \qquad \textbf{[Equation 1]}$$

A similar equation is sometimes known as *Cronbach's Alpha* or *Alpha Coefficient* (Equation 2; Vincent, 1999), where MS_R is the within-subjects error/residual variance. This can be used to assess reliability for two or more measures at the ordinal level of measurement, or when the size of the means is to be ignored (Vincent, 1999). This equation ignores within-subject variance due to differences in between-subject means, and considers only the variance due to the within-subject order with respect to the repeated trials.

$$\text{Cronbach's Alpha} = (MS_B - MS_R) / MS_B \qquad \textbf{[Equation 2]}$$

Two of the most common versions of the intraclass correlation are described in Equations 1 and 2. Other intraclass correlations have been described by McGraw and Wong (1996) and Shrout and Fleiss (1979), many of which feature in software packages such as SPSS (2001). These authors explain where these equations may be more appropriate. For example, when no systematic difference exists between the repeated measures then the Single Measure *R* (Equation 3; McGraw and Wong, 1996; Shrout and Fleiss, 1979; where k is the number of repeat measures) is appropriate if each measurement is a single case. Alternatively, the Average Measure *R* is appropriate if each case is an average of several trials.

$$\text{Single Measure } R = (MS_B - MS_W)/(MS_B + (k-1)\cdot MS_W) \quad \textbf{[Equation 3]}$$

Interpretation of the different versions of the intraclass correlation varies between authors. Fleiss (1986) describes reliability as poor (R<0.40), fair to good (0.40<R<0.75) and excellent (R>0.75). A more conservative interpretation was provided by Vincent (1999) with acceptable but questionable (0.70<R<0.80), moderate (0.80<R<0.90) and high (R>0.90).

The variety of intraclass correlations does, however, cause problems as it is often unclear which version of the intraclass correlation has been used thus making comparisons between studies difficult. Problems also arise from the reliability coefficient being severely affected by variability of the between-subjects effects. If this variance (MS_B) is large, a larger reliability coefficient is given than for a smaller variance. This is supported by Looney (2000) who considers that correct interpretation of intraclass correlation requires a consideration of heterogeneity and that measurement error and differences between the means of the repeated measures should be taken into account. Atkinson and Nevill (1998) highlighted other potential limitations of intraclass correlation in their review of reliability testing of variables relevant to sports medicine. For example, no evidence could be found for analytical goals forming the basis for the interpretation of the ranges of R proposed in the literature (e.g. Vincent, 1999). Perhaps more important than the reliability coefficient is the estimation of the error in the repeated measurements. This can be established from ANOVA (see Mullineaux and Bartlett, 1997) or from LOA or LPR.

As intraclass correlation is a statistical relationship test, it has been proposed that it should not be used to assess reliability as it does not identify the degree of compatibility or agreement between data sets (Bland and Altman, 1986; Ottenbacher and Stull, 1993; Mullineaux *et al.*, 1994). These authors propose that the LOA test is used as it overcomes these problems and provides more meaningful information in the form of descriptive statistics rather than as statistical values. LOA for two sets of measurements are calculated by firstly subtracting the values in one set from those of the other set and then calculating the mean (δ) and standard deviation (σ) of the differences. The LOA are calculated with 95% confidence using Equation 4 (Bland and Altman, 1986). For more than two groups the Agreement Boundary, providing a single value is calculated with 95% confidence using Equation 5 (Bland, 2000; BSI, 1979; Mullineaux *et al.*, 1994 – MS_W also used instead of MS_R to produce a similar result).

$$\mathrm{LOA} = \delta \pm 1.96 \cdot \sigma \qquad \textbf{[Equation 4]}$$

$$\text{Agreement Boundary} = \pm 1.96 \cdot \sqrt{2 \cdot MS_R} \qquad \textbf{[Equation 5]}$$

In sport and exercise science, LOA is often favoured (e.g. Ottenbacher and Stull, 1993) as the calculation is simple and interpreting the data is straightforward because results are reported in the original units of measurement. Also, there is some measure of the fixed and proportional biases in the measurement that traditional tests do not assess. Fixed bias is where one method consistently measures differently to the other and proportional bias is where one method measures proportionally differently to the magnitude of the other measurement. However, fixed and proportional biases often interact and agreement does not provide a means to account for the independent effect of each. It has therefore been proposed that to account for these related entities that a LPR is used (Ludbrook, 1997). Assumptions that must be met for LPR include random error in x and y, normal distribution of the errors, homoscedasticity (homogeneity of variance between the errors and the predicted scores, i.e. additive error) and a linear relationship between the measurements. LPR of the form $\bar{y} = a + b \cdot \bar{x}$ can be calculated in several ways. The simplest are: b is a ratio of the standard deviations of the y and x values and $a = \bar{y} - b \cdot \bar{x}$ (where $\bar{y}$ and $\bar{x}$ are the mean values of y and x). In interpreting LPR, the gradient (b) should ideally be 1 and the constant (a) should ideally be 0. Whether any deviation from these ideal values is acceptable should be considered with respect to, for example, the research question where larger deviations may be acceptable for theory generating work. Confidence intervals for the gradient and constant can be obtained using bootstrapping (Zhu, 1997), the application of which has been illustrated by Mullineaux *et al.* (1998b; 1999). If criterion validity is being assessed, where a new technique is being compared to a 'gold standard', if no error is assumed to exist in the 'gold standard' an ordinary least-squares regression could be used instead of LPR.

The four techniques proposed in the literature to assess reliability are currently all acceptable (i.e. *R*, LOA, Agreement Boundary and LPR). Which to use is a matter of personal preference by the researcher or journal editor, and dependent on which might provide the most meaningful statistical evaluation of the data. The important point is that an assessment of the data is made that supports reliability and, subsequently, validity of the chosen research design. Mullineaux *et al.* (1999) compared the application of LOA and LPR and demonstrated how the assumptions of each can be checked and the results interpreted. If the proposal that these methods are more appropriate is widely accepted, it is possible that LOA, Agreement boundary and LPR will supersede the statistical relationship measurement (i.e. R). Further research applying these techniques would be useful to confirm the benefits of these tests in assessing reliability.

Allometric scaling

To express the relationship between physiological variables and body dimensions a form of scaling known as allometric scaling (Schmidt-Nielsen, 1984) can be used to primarily identify non-linear relationships, although, the result may be linear. There has been a renewed interest in this area applied to humans, but Winter and Nevill (2001) cite scientific literature on allometry dating back to 1838. Generally, allometry is appropriate for identifying the extent to which performance differences are attributable to differences in size or to differences in qualitative characteristics of the body's tissues and structures (Winter and Nevill, 2001, p. 275). In addition, as allometry can cater for non-linear relationships, it is suitable for addressing the non-isometric and isometric changes in dimensions with growth (e.g. Tanner, 1989). It is important that such analyses are underpinned by a theory. Dimensionality theory offers one possible theoretical underpinnings for an allometric scaling analysis.

Dimensionality theory is underpinned by the Système International d'Unités that comprises seven base units (mass, length, time, electric current, temperature, amount of substance and luminous intensity). From these seven units all other units can be derived (e.g. area, volume, density, force, pressure, energy, power, frequency). Often for convenience, these units are renamed, such as, the units of force of $kg.m.s^{-2}$ are denoted as Newtons (N). Dimensionality theory can be used for two main purposes (see, for example, Duncan, 1987): dimensional homogeneity and dimensional analysis. Dimensional homogeneity can be used to check that an equation is correct by partitioning both sides of an equation into their base units. If the units on both sides are partitioned into the same base units then the equation is correct. Dimensional analysis can be used to predict the relationship between different dimensions as a means to provide a theoretical foundation for a research study, such as, between several continuous variables.

When there are two continuous variables (y and x), they can be scaled with each other in many forms, three of which are common: ratio standard ($y = \mathrm{b} \cdot x$); linear regression ($y = \mathrm{a} + \mathrm{b} \cdot x$) and non-linear form ($y = \mathrm{a} \cdot x^{\mathrm{b}}$), where a and b are some constants. The use of each of these scaling techniques should be dictated by theory. This theory is not always obvious. In addition, the mathematics of, and the assumptions for, a scaling technique delimit their use theoretically and statistically. For example, non-linear scaling may be appropriate for data that are not necessarily linear, theoretically requires a zero intercept and contain multiplicative error about the regression. The opposite of these is assumed in ratio standard or linear regression scaling analyses in that the data are linear, the intercept is not fixed at zero and the error is additive.

One form of dimensionality analysis that has been used to predict the relationship between body dimensions and physiological variables is the surface law relationship. This states that the surface area is proportional to its volume to the power of 0.67 (Schmidt-Nielsen, 1984). On this basis, Winter *et al.* (1991) showed that maximal oxygen uptake ($\dot{V}O_{2max}$) is proportional to body mass (BM) to the power of 0.6 (i.e. $\dot{V}O_{2max} \propto BM^{0.6}$). Allometric scaling has also been used to scale a variety of body dimensions with performance measures or other body dimensions. These include body mass with cardiac dimensions (e.g. Batterham, George and Mullineaux, 1995; George, Batterham, Gates and Mullineaux, 1995), thigh cross sectional area with vertical jump performance (Batterham, Barnes and Mullineaux, 1999a; 1999b) and fat free mass with peak power output (Mullineaux *et al.*, 1998a).

A further study by Batterham, George and Mullineaux (1997) showed that left ventricular mass (LVM) was proportional to body mass (BM) to the power of 0.78 (0.65 to 0.91; 95% CI). This is in agreement with the surface law relationship as the power exponent is not statistically significant ($P>0.05$) from 0.67 as the 95% confidence intervals encompass this value. However, as participants vary in body composition (e.g. percentage of fat free mass, FFM), the predicted relationship derived from dimensionality analysis may be inaccurate. As a link between skeletal and cardiac muscles has been proposed (George *et al.*, 1991), FFM may be more appropriate than BM as a predictor variable of LVM. As such, LVM was found to equal FFM to the power of 1.07 (0.92 to 1.22, 95% CI; Batterham, George and Mullineaux, 1997). Hence, ratio standard scaling using BM may overestimate participants' LVM owing to greater percentages of body fat.

An additional benefit of allometric scaling is that more than one independent variable, including one dummy variable (i.e. a dichotomous independent variable coded as 0 and 1), can be included in a multiple non-linear scaling analysis. This will increase the explained variance and reduce the effect of extraneous variables, but a larger number of statistical assumptions (e.g. multicollinearity) need to be checked. Although simple non-linear scaling is easy to perform (i.e. $y = a \cdot x^b$), when there is more than one independent variable (e.g. $y = a \cdot x_1^b + c \cdot x_2^d$, where x_1 and x_2 are independent variables, and c and d are constants) it is easier to use a log-log transformation method in combination with multiple linear regression analyses. As the data are non-linear, where the error is often multiplicative, then the first log transformation generally linearises the relationship, alters the error to being additive and improves the normality distribution of the data. These are all necessary assumptions underpinning a linear regression scaling analysis. At this stage the standard error of the estimate (SE_E) is accurate as the additive error assumption is met. Subsequently, after the second log transformation is used to obtain the non-linear equation, the SE_E needs to be used in multiplicative form of $\pm e^{SE_E}$. Alternatively, the error can be considered by reporting the 95% CI for the power exponents obtained, for example, using bootstrapping (Zhu, 1997). Subsequently, Batterham, George and Mullineaux (1997) extended their analyses to include sex as an additional independent variable (i.e. as a dummy variable coded 0 for males and 1 for females) and provide 95% CI for the power exponents. The data after the first log transformation are represented by Equation 6 (excluding the 95% CI for simplicity of presentation). The second log transformation to obtain the non-linear equation, including the 95% CI for the power exponents, are presented in Equation 7 (Batterham, George and Mullineaux, 1997, p. 184).

$$\ln(LVM) = 0.96 - 0.18 \cdot sex + 1.07 \cdot \ln(FFM) \qquad \textbf{[Equation 6]}$$

$$LVM = 2.6 \cdot (e^{SEX})^{-0.18 \pm 0.10} \cdot FFM^{1.07 \pm 0.15} \qquad \textbf{[Equation 7]}$$

The confidence intervals allow for the power exponent to be assessed for differences from theoretical values. In this instance, the confidence intervals for the power exponent for FFM encompass one and is therefore not significantly different from linearity. In addition, the constant for males (i.e. $2.6\ (e^0)^{-0.18} = 2.6$) and for females (i.e. $2.6\ (e^1)^{-0.18} = 2.2$) from Equation 7 can be used to quantify differences between the groups represented by the dummy variable. Hence, independent of FFM, males possessed a LVM approximately 18% greater than for females (i.e. male to female ratio minus one then multiplied by 100, that is, $(2.6 / 2.2 - 1) \times 100$).

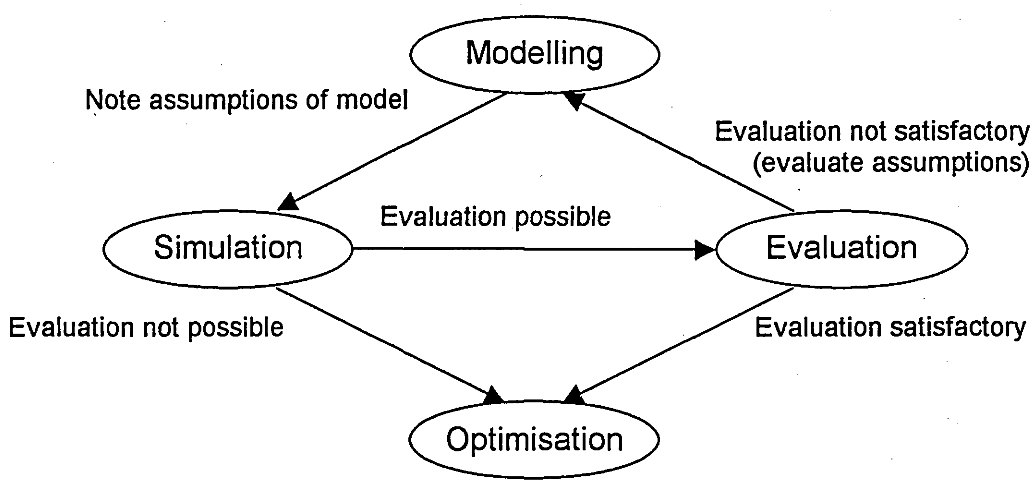
In scaling analyses, providing a theory that supports the relationship identified can be difficult. In particular, this is difficult in data analyses that routinely use ratio standard or linear regression scaling analyses and yet their use is still common. Papers on allometric scaling (e.g. Batterham, George and Mullineaux, 1997) address the need for a theoretical foundation for data analyses. A further limitation of ratio standard and linear regression scaling has been proposed by Batterham, George and Mullineaux (1997) in that extrapolation beyond the actual data range should be avoided as these tests would not generally meet the zero intercept assumption. However, this caution may also need to be applied to allometrically scaled relationships as the assumption of a zero intercept is beyond the data range that could be tested empirically. Future research on allometrically scaled relationships requires further empirical data to support the theoretical relationships. For instance, this support could be provided by research providing evidence confirming the reliability of the data analyses.

Mathematical modelling

The applications of many statistical techniques are delimited by several assumptions including the distribution of the sample. A statistical technique that is not restricted by the sampling distribution is mathematical modelling. This comprises one obligatory component (i.e. modelling) and three potential subsequent components (i.e. simulation, optimisation and evaluation). Figure 1 illustrates the links between these four concepts:

- Modelling is "an attempt to represent reality" (Nigg, 1999, p. 427).
- Simulation is "experimentation using a model" (Nigg, 1999, p. 429).
- Optimisation is an iterative process to identify the optimal simulation on a performance objective (Marshall and Elliott, 1998).
- Evaluation or validation is "providing evidence that the model is strong and powerful" (Nigg, 1999, p. 429).

Figure 1. The relationship between the components of mathematical modelling: modelling, simulation, optimisation and evaluation (from Bartlett, 1999, p. 189).



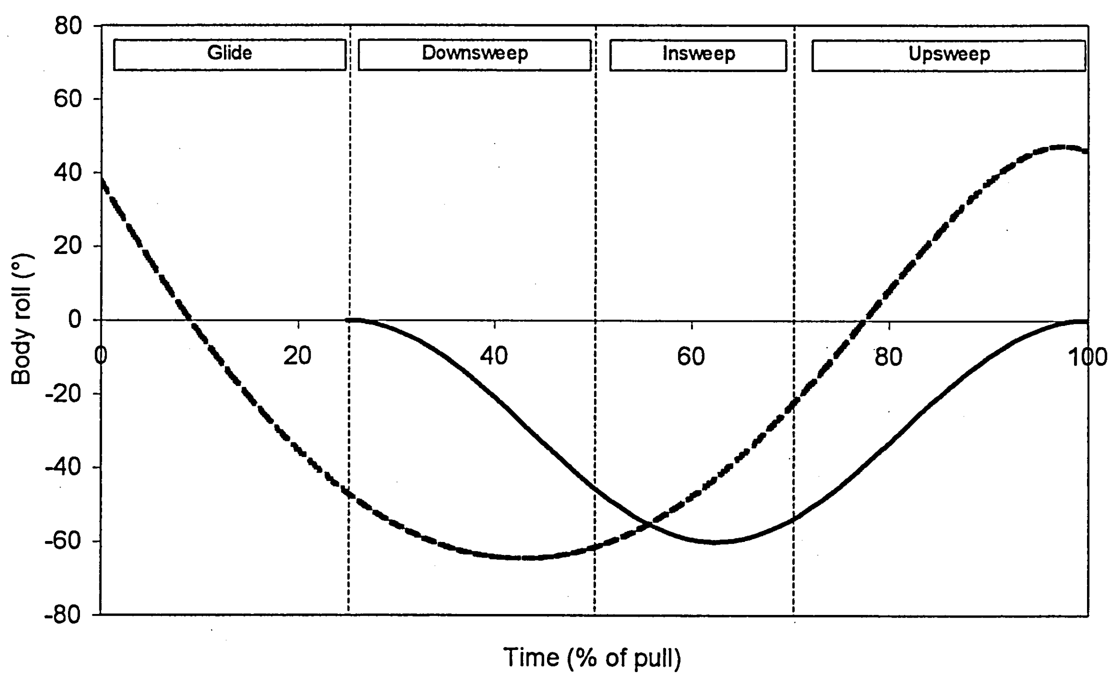
Modelling is used throughout sport and exercise science, including, dimensionality analysis in physiology (e.g. Batterham, George and Mullineaux, 1997 – see allometric scaling earlier), geometrical analysis in motor control (e.g. Lee, 1976) and the equations of motion in biomechanics (e.g. Payton, Hay and Mullineaux, 1997). Evaluating sport techniques on the basis of physical laws through mathematical modelling provides a potentially more appropriate method than using inferential statistical models. This is partly because of limitations of experimental methodologies and the corresponding benefits of mathematical modelling which include (Vaughan, 1984):

- It is safer than participants attempting potentially hazardous techniques.
- It saves time and money in running many simulations that have the potential to predict optimal performance.
- Appropriate variables are clearly selected on theoretical grounds.
- It demonstrates how the movement could be rather than how it is performed.
- Inferential statistical assumptions are not required and thus not violated.
- There is high experimental control enabling technique to be precisely manipulated to measure effects, or attribute changes to certain aspects of techniques that occur without additional alterations resulting.

Research in the biomechanics of swimming is an example of where mathematical modelling can be put to good use. This is because difficulties in experimental methodologies, such as filming underwater, has led to few studies and a poor understanding of the area (Hay, 1988). It has subsequently been proposed that mathematical modelling would probably contribute the most to a better understanding of swimming biomechanics (Martin, 1989). However, few mathematical models exist for swimming biomechanics (Gallenstein and Huston, 1973; Martin *et al.*, 1981; Hay *et al.*, 1993; Payton, Hay and Mullineaux, 1995; Payton and Mullineaux, 1996; Payton, Hay and Mullineaux, 1997). Although different types of models exist (i.e. analytical, deductive, black-box and conceptual; Nigg, 1999), a black-box approach, where a set of mathematical functions determine the input to output relationship, has typically been used in swimming biomechanics.

Mathematical modelling has its limitations (Vaughan, 1984). However, the principal concern for any model is whether or not it effectively represents the real world. Nigg (1999) suggested that the assumptions in the model required to simplify the real world can be evaluated through either direct, indirect or trend measurement methods (Nigg, 1999). In 1997, Payton, Hay and Mullineaux (1997) using a black-box model investigated the effect of body roll on hand path and hand speed in front crawl swimming. The simulations of the model demonstrated that body roll alone can account for the medio-lateral motions of the hand that are considered important in generating propulsive lift forces required for better performance (e.g. Schleihauf *et al.*, 1983). Such models provide the basis for future research (e.g. Yanai, 2001). In addition, the model by Payton, Hay and Mullineaux (1997) was the subject of a 'direct evaluation' by Payton *et al.* (1999) where the results from the model were compared to empirical data. It was concluded that the model "fails to accurately represent the movements of the trunk and upper extremity used by the swimmers" as "the assumption that the trunk rolls away from the neutral position for the duration of the insweep ... was not the case ..." (p. 694). To clarify this comparison, the body roll angles for the model and for the swimmers are illustrated in Figure 2.

Figure 2. Body roll angles during the front crawl pull. Dashed line: breath-holding trials of swimmers throughout the four phases (Payton *et al.*, 1999). Solid line: mathematical model throughout three equivalent phases (Payton, Hay and Mullineaux, 1997).



It can be seen from Figure 2 that the shape of the model is similar to that of the empirical data. However, the timing of the maximal body roll in the model at approximately 62% of the pull time needs modifying so that it occurs at approximately 42% of the pull time as observed for the empirical data. This can be achieved by including the glide phase within the model. In addition, the modelling of the body roll, as a cosine function of the shoulder extension angle, needs modifiying so that positive values aswell as negative values can be obtained. Nevertheless, care should be taken that theory is used, and not the existing action by swimmers, to develop the mathematical model further because the action might not be optimal.

This mathematical model by Payton, Hay and Mullineaux (1997), although provided the basis for future research (see Yanai, 2001 and Payton *et al.*, 1999), is currently at the stage of 'evaluation not satisfactory' (see Figure 1). However, the modifications to the black-box model to include the glide phase and to allow both positive and negative values will likely place the model at the stage of 'evaluation satisfactory'. This is because the pattern, timing and magnitude of body roll angles in the model will more accurately simulate those observed in swimmers (see Figure 2). This model would subsequently provide the basis for future research on the 'optimisation' of technique and thus promote further understanding of the influence of body roll angles in front crawl swimming.

Logistic regression

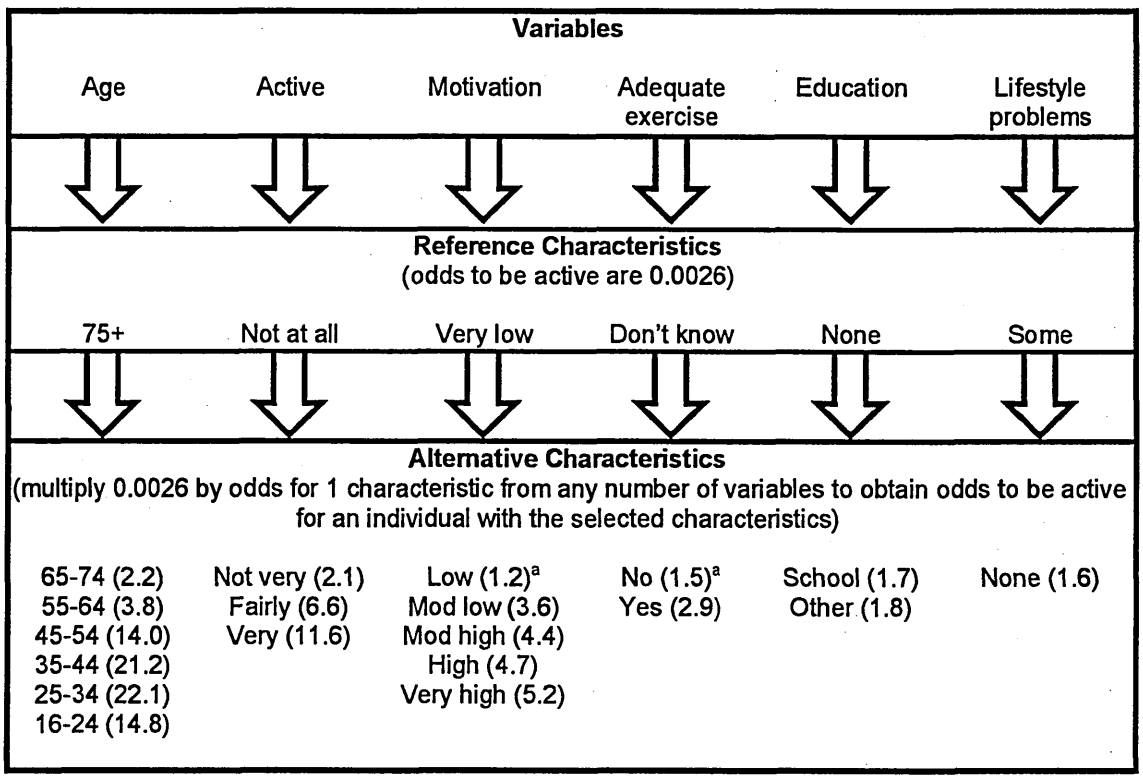
In sport and exercise science research, gathering data of a nominal level of measurement has benefits that include the ease of its presentation and interpretation of findings. As such, nominal data can be useful for publicising meaning for practical, professional and political aims. This is particularly the case for the often large data sets derived from surveys where the results are often presented in a simple uni-factorial form such as frequency counts. An example of such a survey is the Allied Dunbar National Fitness Survey of English adults (Activity and Health Research, 1992). In presenting the results from this survey only frequency data were reported in the original report. Such surveys on sport, health and exercise provide a wealth of information that allows for future secondary data analyses.

Secondary data analyses, owing to the more focused nature of the research and reduced data set, often extend the uni-factorial analyses of the original surveys and use advanced multi-factorial analyses such as discriminant analysis. Discriminant analysis has been used to predict factors that determine group membership. Although rarely used in sport and exercise psychology, Biddle *et al.* (2001) considers its wider use could be beneficial. Conversely, Munro (2001) considers logistic regression as a preferred test owing to a number of advantages over discriminant analysis. These advantages include that the independent variables can be of any level of measurement, assumptions of the data are fewer (e.g. the normal distribution is not required) and a quantification of the phenomena is calculated. This quantification is in the form of odds values that can be used to predict membership to each level of a dichotomous dependent variable.

As with all regression analyses, there must be a theoretical relationship between the dependent variable and independent variables. The complexity of the relationship between variables can be accounted for in logistic regression as the independent variables can be entered in three forms: categorical (e.g. sex), continuous (e.g. age) or interaction (e.g. sex by age as a two-way interaction). Examples in the literature use variables in combinations of the different forms. Mullineaux and Barnes (1997; 1998) and Mullineaux *et al.* (2001a) used the variables in categorical form. Brill *et al.* (2000) and Pollard and Reep (1997) used a mixture of categorical and continuous forms of the variables. Iwao *et al.* (2000) included variables in an interaction form. However, the variety of variable combinations and the traditional tabular presentation of the logistic regression model typically used in the literature can make interpretation of the results difficult. Additional limitations of logistic regression are that: the dependent variable is limited to two groups; the regression results are obtained by an 'iterative process' or 'best guess' and do not provide predictions based on an underlying mathematical theory and dissemination of findings is difficult as many practitioners' understanding of 'odds' values is limited.

The aim of the study by Mullineaux *et al.* (2001a) was to demonstrate a use of logistic regression via a secondary data analysis of the Allied Dunbar National Fitness Survey (Activity and Health Research, 1992) and to emphasise processes to overcome some of the limitations of this statistical technique. For example, for ease of analysis: all independent variables were coded as categorical variables; variables with the lowest odds were coded appropriately so that only multiplication would be required in using the model and standard error estimates were omitted. In addition, as in the literature the logistic regression model is typically presented in tabular form, which is difficult to interpret, in the study by Mullineaux *et al.* (2001a) the data were presented in an alternative graphical form (Figure 3).

Figure 3. New alternative method for presenting logistic regression results (from Mullineaux *et al.*, 2001a).



The same results from Figure 3 can be compared to the typical tabular form of the same data presented in Table 2. This graphical method provides a simpler presentation format for the logistic regression model by firstly better highlighting the characteristics included in the constant term and secondly by omitting the frequency data and confidence intervals for the odds that are not required when using the model. Subsequently, selecting characteristics from each variable are clearer to enable the odds of that individual to participate in adequate physical activity to be calculated more easily. It is suggested that this new graphical method will facilitate an easier application of logistic regression models.

There were, however, several limitations in the study by Mullineaux *et al.* (2001a). For instance, difficulty occurred in identifying any valid ethnic influence for participation in physical activity. This was due to a small and uneven sample proportions that led to an increase in the error and therefore uncertainty in the results (Menard, 1995). Nevertheless, such restrictions apply to many other statistical techniques and are not unique to logistic regression. Furthermore, the analysis compared only the 'sedentary' to 'sufficiently active' groups, and did not analyse the 'some activity' group that participated at a level lower than that deemed beneficial for health. Extending the study by Mullineaux *et al.* (2001a) to use a multinomial or polychotomous logistic regression (SPSS, 2001) may overcome this latter limitation. This may provide greater information on using such statistical tests for analysing data of a nominal or mixed level of measurement to enable the dependent variable in the research to be defined with either two or more than two groups.

Table 2. Typical tabular method used to present logistic regression results (data from Mullineaux *et al.*, 2001a).

Variable	Characteristic	Frequency	Odds	95%CI
Constant			0.0026	-3.50 to 3.51
Age	75+	246		
	65 to 74	270	2.18	-0.57 to 4.94
	55 to 64	276	3.80	0.96 to 6.63
	45 to 54	295	14.02	11.19 to 16.86
	35 to 44	313	21.23	18.38 to 24.10
	25 to 34	336	22.13	19.26 to 25.00
	16 to 24	269	14.78	11.88 to 17.67
Active	Not at all	167		
	Not very	354	2.12	-0.51 to 4.76
	Fairly	1038	6.57	3.96 to 9.18
	Very	446	11.57	8.87 to 14.26
Motivation	Very low	475		
	Low	145	1.20[a]	-1.44 to 3.84
	Mod low	270	3.63	1.08 to 6.17
	Mod high	452	4.35	1.83 to 6.87
	High	418	4.75	2.22 to 7.27
	Very high	245	5.19	2.59 to 7.78
Adequate exercise	Don't know	60		
	No	760	1.48[a]	-1.56 to 4.52
	Yes	1185	2.95	-0.09 to 5.98
Education	None	937		
	School	604	1.70	-0.59 to 3.99
	Other	464	1.84	-0.46 to 4.14
Lifestyle problems	Some	607		
	None	1398	1.63	-0.63 to 3.90

[a] No statistically significant difference from the first characteristic from each variable ($P>0.05$). These first characteristics from each variable together constitute the constant term.

Review papers

In addition to analysing specific techniques, review papers addressing a breadth of issues for different disciplines of sport and exercise science provide additional benefits for researchers. Part of the title for a review paper on sport psychology by Schutz and Gessaroli (1993), 'use, misuse and disuse', highlights potential benefits of review papers. More specifically, benefits highlighted include the breadths of traditional and novel statistical applications that are relevant for different disciplines. In addition, an indication of the diversity of theoretical topics in the area is provided, papers can be more up to date than information provided in textbooks and existing publications may be critiqued for appropriate and inappropriate applications of statistics. Owing to the potential benefits of overview papers, several have been published for each of the main disciplines of sport and exercise science. These include papers by James and Bates (1997) and Mullineaux and Bartlett (1997) for biomechanics, Shultz and Sands (1995) for physiology and Schutz and Gessaroli (1993) for psychology.

More recently, I proposed that a special issue on Research Methods and Statistics be commissioned by the Journal of Sports Sciences. This has been published and is summarised through the editorial by Nevill, Atkinson and Mullineaux (2001). This issue included invited review papers for the four areas of sport performance research (Atkinson and Nevill, 2001), psychology (Biddle *et al.*, 2001), physiology of exercise and kinanthropometry (Winter *et al.*, 2001) and biomechanics and motor control (Mullineaux *et al.*, 2001b). These papers further highlight current practice and recent developments in applications of statistics that will further facilitate more appropriate use of statistics in sport and exercise science.

Summary

The aim of this thesis is to highlight some appropriate uses of statistical techniques through applied examples in sport and exercise science research. The preceding sections provide an overview of selected techniques on five topics. Firstly, as measurement tools should be reliable for validity, LOA and LPR were proposed as appropriate techniques of testing repeated measures reliability (e.g. Mullineaux *et al.*, 1999). The next three topics were concerned with the necessity to analyse data with respect to an underpinning theory. The use of allometric scaling was shown to be appropriate for testing the relationship between body dimensions and performance measures (e.g. Batterham, George and Mullineaux, 1997). The use of mathematical modelling was demonstrated as beneficial for investigating a research question that would be difficult to test empirically (e.g. Payton, Hay and Mullineaux, 1997). The use of logistic regression to explore data of a nominal or mixed level of measurement was also highlighted (e.g. Mullineaux *et al.*, 2001a). Lastly, the benefits of review articles for promoting the understanding and appropriateness of applications of statistical techniques in specific disciplines were outlined (e.g. Mullineaux *et al.*, 2001b). In addition to the brief review of each of these topics, the contribution of the appended published papers to highlighting an appropriate use of statistical analyses and directions for future research were provided. Statistical analysis techniques will continue to become more easily available and will allow increasingly complex questions to be investigated. As a result, the need for publications that emphasise the appropriate use of statistical techniques in sport and exercise science will continue.

References

Activity and Health Research (1992). *Allied Dunbar National Fitness Survey: Main Findings*. London: Sports Council and Health Education Authority.

Atkinson, G. and Nevill, A.M. (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Medicine*, **26**, 217-38.

Atkinson, G. and Nevill, A.M. (2001). Selected issues in the design and analysis of sport performance research. *Journal of Sports Sciences*, **19**, 811-27.

Bartlett, R.M. (1999). *Sports Biomechanics: Reducing Injury and Improving Performance*. London: E & FN Spon.

Batterham, A.M., Barnes, C.A. and Mullineaux, D.R. (1999a). Modelling the influence of thigh muscle cross-sectional area on vertical jump performance in young professional soccer players. *Journal of Sports Sciences*, **17**, 807-8.

Batterham, A.M., Barnes, C.A. and Mullineaux, D.R. (1999b). Modelling the influence of thigh muscle cross-sectional area on vertical jump performance in young professional soccer players. In *Proceedings of the World Congress of Science and Football IV*, p. 11. Sydney: University of Technology.

Batterham, A.M., George, K.P. and Mullineaux, D.R. (1995). Relationship between heart size and body dimensions: an allometric scaling approach. *Medicine and Science in Sports and Exercise*, **27**, S158.

Batterham, A.M., George, K.P. and Mullineaux, D.R. (1997). Allometric scaling of left ventricular mass and body dimensions in males and females. *Medicine and Science in Sports and Exercise*, **29**, 181-6.

Biddle, S.J.H., Markland, D., Gilbourne, D., Chatzisarantis, N.L.D. and Sparkes, A.C. (2001). Research methods in sport and exercise psychology: quantitative and qualitative issues. *Journal of Sports Science*, **19**, 777-809.

Bland, J.M. (2000). *An Introduction to Medical Statistics*. Oxford: Oxford University Press.

Bland, J.M. and Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, **1**, 307-10.

Brill, P.A., Macera, C.A., Davis, D.R., Blair, S.N. and Gordon, N. (2000). Muscular strength and physical function. *Medicine and Science in Sports and Exercise*, **32**, 412-6.

BSI (1979). *Precision of Test Methods 1: Guide for the Determination and Reproducibility of a Standard Test Method (BS 5497, Part 1)*. London: British Standards Institute.

Duncan, T. (1987). *Physics: A Textbook for Advanced Level Students*. London: John Murray.

Field, A. (2000). *Discovering Statistics using SPSS for Windows*. London: Sage.

Fleiss, J.L. (1986). *The Design and Analysis of Clinical Experiments*. New York: Wiley.

Gallenstein, J. and Huston, R.L. (1973). Analysis of swimming motions. *Human Factors*, **15**, 91-8.

George, K.P., Batterham, A.M., Gates, P.E. and Mullineaux, D.R. (1995). Ratio-standard versus allometric scaling in cardiac dimensions of athletes and controls. *Medicine and Science in Sports and Exercise*, **27**, S158.

George, K.P., Wolfe, L.A. and Burggraf, G.W. (1991). The 'athletic heart syndrome': a critical review. *Sports Medicine*, **11**, 300-31.

Gratton, C., Shibli, S. and Kokolakakis, T. (1999). *Leisure Forecasts 1999-2003*. Sheffield: Leisure Industries Research Centre.

Hay, J.G. (1988). The status of research on the biomechanics of swimming. In *Swimming Science V* (edited by B.E. Ungerechts, K. Wilke and K. Reischle), pp. 3-14. Champaign, IL: Human Kinetics.

Hay, J.G., Liu, Q. and Andrews, J.G. (1993). Body roll and handpath in freestyle swimming: a computer simulation study. *Journal of Applied Biomechanics*, **9**, 227-37.

Health Development Agency (1999). *Saving Lives: Our Healthier Nation*. London: The Stationary Office.

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

Iwao, N., Iwao, S., Muller, D.C., Elahi, D., Shimokata, H. and Andres, R. (2000). A test of recently proposed BMI standards with respect to old age. *Aging-Clinical and Experimental Research*, **12**, 461-9.

James, C.R. and Bates, B.T. (1997). Experimental and statistical design issues in human movement research. *Measurement in Physical Education and Exercise Science*, **1**, 55-69.

Lee, D.N. (1976). A theory of visual control of braking based on information about time-to-collision. *Perception*, **5**, 437-59.

Looney, M.A. (2000). When is the intraclass correlation coefficient misleading? *Measurement in Physical Education and Exercise Science*, **4**, 73-8.

Ludbrook, J. (1997). Comparing methods of measurement. *Clinical and Experimental Pharmacology Physiology*, **24**, 193-203.

Marshall, R.N. and Elliott, B.C. (1998). The analysis and development of technique in sport. In *Training in Sport: Applying Sport Science* (edited by B.C. Elliott), pp. 117-44. Chichester: Wiley.

Martin, R.B. (1989). Swimming: forces on aquatic animals and humans. In *Biomechanics of Sport* (edited by C.L. Vaughan), pp. 35-51. Boca Raton, FL: CRC Press.

Martin, R.B., Yeater, R.A. and White, M.K. (1981). A simple analytical model for the crawl stroke. *Journal of Biomechanics*, **14**, 539-48.

Matthews, R. (1998). Flukes and flaws. *Prospect*, **November**, 20-4.

McGraw, K.O. and Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, **1**, 30-46.

Menard, S. (1995). *Applied Logistic Regression Analysis*. London: Sage.

Mullineaux, D.R. and Barnes, C.A. (1997). Determinants of propensity to partake in adequate physical activity to promote health. In *Proceedings of the 2nd Annual Congress of the European College of Sport Sciences* (edited by J. Bangsbo, B. Saltin, H. Bonde, Y. Hellsten, B. Iben, M. Kjær and G. Sjøgaard), pp. 651-2. Copenhagen: ECSS.

Mullineaux, D.R. and Barnes, C.A. (1998). Determinants of propensity to partake in adequate physical activity to promote health. *Journal of Sports Sciences*, **16**, 394-5.

Mullineaux, D.R., Barnes, C.A. and Barnes, E.F. (2001a). Factors affecting the likelihood to engage in adequate physical activity to promote health. *Journal of Sports Sciences*, **19**, 279-88.

Mullineaux, D.R., Barnes, C.A. and Batterham, A.M. (1998a). Allometric scaling of anaerobic performance and the use of 'bootstrapping' for statistical inference from a small sample. In *Proceedings of the 2nd Australasian Biomechanics Conference* (edited by P. Hume, R. Marshall, P. Hunter, S. Stanley, I. Anderson and P. McNair), pp. 72-3. Auckland: The University of Auckland.

Mullineaux, D.R., Barnes, C.A. and Batterham, A.M. (1998b). Assessment of fixed and proportional bias between two measurements. In *Proceedings of the*

3rd World Congress of Biomechanics (edited by Y. Matsuzaki, T. Nakamura and E. Tanaka), pp. 393. Sapporo: Hokkaido University.

Mullineaux, D.R., Barnes, C.A. and Batterham, A.M. (1999). Assessment of bias in comparing measurements: a reliability example. *Measurement in Physical Education and Exercise Science*, **3**, 195-205.

Mullineaux, D.R. and Bartlett, R.M. (1997). Research methods and statistics. In *Biomechanical Analysis of Movement in Sport and Exercise* (edited by R.M. Bartlett), pp. 81-104. Leeds: BASES.

Mullineaux, D.R., Bartlett, R.M. and Bennett, S. (2001b). Research methods and statistics in biomechanics and motor control. *Journal of Sports Sciences*, **19**, 739-60.

Mullineaux, D.R., Scott, M.A. and Batterham, A.M. (1994). The use of "agreement" for assessing digitiser reliability/objectivity. In *Proceedings of the Biomechanics Section of the British Association of Sport and Exercise Sciences* (edited by J. Watkins), pp. 17-20. Leeds: BASES.

Munro, B.H. (2001). *Statistical Methods for Health Care Research*. Philadelphia, PA: Lippincott.

Nevill, A.M. and Atkinson, G. (1997). Assessing agreement between measurements recorded on a ratio scale in sports medicine and sports science. *British Journal of Sports Medicine*, **31**, 314-8.

Nevill, A.M. and Atkinson, G. (2001). Statistical methods in kinanthropometry and exercise physiology. In *Kinanthropometry and Exercise Physiology*

Laboratory Manual: Tests, Procedures and Data. Volume 1: Anthropometry (edited by R. Eston and T. Reilly), pp. 237-73. London: Routledge.

Nevill, A.M., Atkinson, G. and Mullineaux, D.R. (2001). Editorial: new horizons in research methods. *Journal of Sports Sciences*, **19**, 737-8.

Nigg, B.N. (1999). General comments about modelling. In *Biomechanics of the Musculo-skeletal System* (edited by B.N. Nigg and W. Herzog), pp. 435-45. Chichester: Wiley.

Ottenbacher, K.J. and Stull, G.A. (1993). The analysis and interpretation of method comparison studies in rehabilitation research. *American Journal of Physical Medicine and Rehabilitation*, **72**, 266-71.

Payton, C.J., Bartlett, R.M., Baltzopoulos, V. and Coombs, R. (1999). Upper extremity kinematics and body roll during preferred-side breathing and breath-holding front crawl swimming. *Journal of Sports Sciences*, **17**, 689-96.

Payton, C.J., Hay, J.G. and Mullineaux, D.R. (1995). The effect of body roll on hand velocity in the front crawl. In *Proceedings of the XVth Congress of the International Society of Biomechanics* (edited by K. Häkkinen, K.L. Keskinen, P.V. Komi and A. Mero), pp. 720-1. Jyväskylä: University of Jyväskylä.

Payton, C.J., Hay, J.G. and Mullineaux, D.R. (1997). The effect of body roll on hand speed and hand path in the front crawl - a simulation study. *Journal of Applied Biomechanics*, **13**, 300-15.

Payton, C.J. and Mullineaux, D.R. (1996). Effect of body roll on hand velocity in freestyle swimming. In *Biomechanics and Medicine in Swimming VII* (edited by

J.P. Troup, A.P. Hollander, D. Strass, S.W. Trappe, J.M. Cappaert and T.A. Trappe), pp. 59-63. London: E & FN Spon.

Pollard, R. and Reep, C. (1997). Measuring effectiveness of playing strategies at soccer. *The Statistician*, **46**, 541-50.

Ricker, W.E. (1973). Linear regression in fishery research. *Journal of Fisheries Research Board Canada*, **30**, 409-34.

Robson, C. (1993). *Real World Research*. Oxford: Blackwell.

Sale, D.G. (1991). Testing strength and power. In *Physiological Testing of the High-Performance Athlete* (edited by J.D. MacDougall, H.A. Wenger and H.J. Green), pp. 21-106. Champaign, IL: Human Kinetics.

Schleihauf, R.E., Gray, L. and DeRose, J. (1983). Three-dimensional analysis of hand propulsion in the sprint front crawl stroke. In *Biomechanics and Medicine in Swimming* (edited by A.P. Hollander, P.A. Huijing and G. deGroot), pp. 173-83. Champaign, IL: Human Kinetics.

Schmidt-Nielsen, K. (1984). *Scaling: Why is Animal Size so Important?* Cambridge: Cambridge University Press.

Schutz, R.W. and Gessaroli, M.E. (1993). Use, misuse and disuse of psychometrics in sport psychology research. In *Handbook of Research on Sport Psychology* (edited by R.N. Singer, M. Murphey and L.K. Tennant), pp. 901-17. New York: Macmillan.

Shrout, P.E. and Fleiss, J.L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, **86**, 420-28.

Shultz, B.B. and Sands, W.A. (1995). Understanding measurement concepts and statistical procedures. In *Physiology Assessment of Human Fitness* (edited by P.J. Maud and C. Foster), pp. 257-87. Champaign, IL: Human Kinetics.

Sport England (1999). *Investing for our Sporting Future: Sport England Lottery Fund Strategy 1999-2009*. London: Sport England.

SPSS (2001). *SPSS for Windows Release 11.0.0*. Chicago, IL: SPSS.

Student (1908). The probable error of a mean. *Biometrika*, **6**, 1-25.

Tanner, J.M. (1989). *Foetus into Man*. Ware: Castlemead.

Thomas, J.R. and Nelson, J.K. (2001). *Research Methods in Physical Activity.* Champaign, IL: Human Kinetics.

UCAS (2001). Annual statistical tables: subject of study. *http://www.ucas.ac.uk/figures/archive/download/index.html*, visited 9 October.

Vaughan, C.L. (1984). Computer simulation of human motion in sports biomechanics. In *Exercise and Sport Science Reviews: Volume 12* (edited by R.L. Terjung), pp. 373-416. Lexington, KY: Collamore Press.

Vincent, W.V. (1999). *Statistics in Kinesiology.* Champaign, IL: Human Kinetics.

Wilson, K. and Batterham, A.M. (1999). Stability of questionnaire items in sport and exercise psychology: bootstrap limits of agreement. *Journal of Sports Sciences*, **17**, 725-34.

Winter, E.M., Brookes, F.B.C. and Hamley, E.J. (1991). Maximal exercise performance and lean leg volume in men and women. *Journal of Sports Sciences*, **9**, 3-13.

Winter, E.M., Eston, R.G. and Lamb, K.L. (2001). Statistical analyses in the physiology of exercise and kinanthropometry. *Journal of Sports Sciences*, **19**, 761-75.

Winter, E.M. and Nevill, A.M. (2001). Scaling: adjusting for differences in body size. In *Kinanthropometry and Exercise Physiology Laboratory Manual: Tests, Procedures and Data. Volume 1: Anthropometry* (edited by R. Eston and T. Reilly), pp. 275-93. London: Routledge.

Yanai, T. (2001). What causes the body roll in front-crawl swimming? *Journal of Applied Biomechanics*, **17**, 28-42.

Zhu, W. (1997). Making bootstrap statistical inferences: a tutorial. *Research Quarterly for Exercise and Sport*, **68**, 44-55.

Appendix 1

Mullineaux, D.R., Barnes, C.A. and Batterham, A.M. (1998b). Assessment of fixed and proportional bias between two measurements. In *Proceedings of the 3rd World Congress of Biomechanics* (edited by Y. Matsuzaki, T. Nakamura and E. Tanaka), pp. 393. Sapporo: Hokkaido University.

ASSESSMENT OF FIXED AND PROPORTIONAL BIAS BETWEEN TWO MEASUREMENTS

D.R. Mullineaux, C.A. Barnes and A.M. Batterham, School of Social Sciences, University of Teesside, Middlesbrough, TS1 3BA, UK.

Many studies in the applied sciences have compared two methods of assessment for the same variable. The rationale for such comparisons may include assessment of reliability, objectivity or criterion validity. Comparative analysis of data has traditionally been undertaken using Pearson rho or Intraclass correlation co-efficients. These tests are inappropriate, however, because they do not identify fixed bias (one method consistently measuring differently to the other) or proportional bias (one method measuring proportionally differently to the magnitude of the measurement), as examples. More recently the 'agreement' technique (Bland and Altman, 1986, *The Lancet*, **1**, 307-10) has been favoured, as the calculation is simple, and the interpretation of results is easy as they are reported in the original units of measurement. However, fixed and proportional bias often interact, and agreement does not provide the means to account for the independent effect of each. It has therefore been proposed that to account for these related entities that a 'least products regression' (LPR) should be used (Ludbrook, 1997, *Clinical and Experimental Pharmacology Physiology*, **24**, 193-203). Assumptions that must be met for LPR include random error in x and y, normal distribution of the errors in LPR, homoscedasticity (homogeneity of variance between the errors in the LPR and the predicted scores, i.e. additive error) and a linear relationship between the measurements. To highlight the use of LPR, the release speed of the ball in 30 hockey flicks was determined through digitising by two operators. The release speed was found to be 10.99 ± 2.55 m.s^{-1} and 10.95 ± 2.57 m.s^{-1} by operator 1 and 2, respectively. The necessary assumptions were all met: Z scores for the skewness and kurtosis of the errors were all within $\pm$ 2 indicating a normal distribution; Levene's test was not significant ($p = 0.88$) indicating homoscedasticity; $r = 0.99$ indicating a linear relationship. As the error is additive a LPR was performed in SPSS v 6.1 using the Loss function (NB: if the error was multiplicative the LPR would be weighted). Results provided are the parameter estimates (with 95% confidence intervals obtained by bootstrapping). For the regression equation the parameter estimate for the intercept was -0.12 (-0.43 to 0.19) and for the gradient was 1.01 (0.98 to 1.03). As the confidence intervals encompass 0 for the intercept and 1 for the gradient these indicate respectively that there is no fixed or proportional bias between the operators, thus confirming objectivity of the measurement. Least products regression provides the appropriate statistical analysis to compare two measurements, provided, the assumptions underpinning the analysis are met or controlled.

Appendix 2

Mullineaux, D.R., Barnes, C.A. and Batterham, A.M. (1999). Assessment of bias in comparing measurements: a reliability example. *Measurement in Physical Education and Exercise Science*, **4**, 195-205.

MEASUREMENT IN PHYSICAL EDUCATION AND EXERCISE SCIENCE, 3(4), 195–205

Assessment of Bias in Comparing Measurements: A Reliability Example

David R. Mullineaux
Sport Science Research Institute
Sheffield Hallam University

Christopher A. Barnes
Middlesbrough Football Club Training Headquarters

Alan M. Batterham
School of Social Sciences
University of Teesside

Comparative analyses of a variable measured twice or against a "gold standard" technique should explore the existence of any fixed and proportional biases between the 2 measurements. Levels of agreement (LOA) consider these biases together and least products regression (LPR) consider their effect independently. To compare the use of LOA and LPR, the peak torque extension (PTE) of the knee at 2.1 rad/sec during isokinetic dynamometry was obtained on 2 separate days ($N = 17$). The mean PTE (with standard deviations in parentheses) was found to be 93.6 (13.9) $N \cdot m$ on Day 1 and 92.5 (11.5) $N \cdot m$ on Day 2. The LOA were 1.06 ± 10.80 $N \cdot m$ (95% confidence), and the LPR's (with 95% confidence intervals in parentheses) intercept was –17.7 $N \cdot m$ (–37.4 to 2.03) and slope was 1.20 (1.01 to 1.40). LOA and LPR are suitable techniques to compare 2 measurements and, because the levels are large and the slope does not encompass 1, suggest that the knee's PTE at 2.1 rad/sec is unreliable.

Key words: agreement, bias, measurement, regression, reliability, validity

---

Requests for reprints should be sent to David R. Mullineaux, Sport Science Research Institute, Sheffield Hallam University, Collegiate Crescent Campus, Sheffield, S10 2BP, United Kingdom. E-mail: D.Mullineaux@shu.ac.uk

A multitude of internal and external factors have the potential to impact upon obtaining measurements for assessing human performance even when great lengths are taken to minimize such influences. To ensure confidence in the results, an assessment of the consistency or validity of the result is advisable. For example, the test–retest reliability of equipment needs to be consistent so that a single measurement can be justified. Also, for example, the results obtained simultaneously from measurements with an established technique and a new technique should be in agreement for criterion validity to be assured and to enable the new technique to be considered an acceptable alternative to the established technique. To support such comparisons, data should be analyzed using appropriate statistical tests.

The appropriateness of the statistical test used to make comparisons between more than one measurement is dependent on many factors. Due consideration should be made of the test's principles (e.g., bivariate data), underpinning assumptions (e.g., normal distribution of errors), and output (e.g., values in the original units of measurement) so that the results obtained can be used to accurately assess the aim of the comparison. Authors may argue that the test they use is the most appropriate one, and often the case for their argument can be convincing. Generally, however, no test is singularly ideal as the test may not be able to accommodate the number of repeated measurements, provide accurate results if the statistical assumptions by the data are violated, or provide results that are simple to interpret. This article therefore examines the appropriateness of statistical tests that are used for comparing two measurements on the same parameter at the interval or ratio level of measurement.

COMPARING TWO MEASUREMENTS

The notion of two repeated measurements at the interval or ratio level of measurement agreeing can be illustrated by producing a scatter plot of the data. A visual inspection is important to qualitatively assess the two measurements as to (a) whether the relation is linear, because quantitative techniques can mask nonlinear trends with high linear correlation values, and (b) how close the agreement is that will aid in interpreting the quantitative results obtained. Data that do agree would lie along the linear line of identity that passes through zero and possesses a slope of 1 (Figure 1, solid line). Plotting a least products regression (LPR) line (Figure 1, solid line)—appropriate because it minimizes errors in both x and y measurements—and comparing this with the line of identity facilitates comparison between the two measurements. If the LPR does not pass through zero, then this indicates that one measurement is consistently different from the other (Figure 1, dashed line). Also, if the regression line does not possess a slope equal to 1, then the measurements are proportionally different to each other (Figure 1, dotted line). These "none zero intercept" and "a slope not equal to 1" have been defined as *fixed* and *proportional bi-*

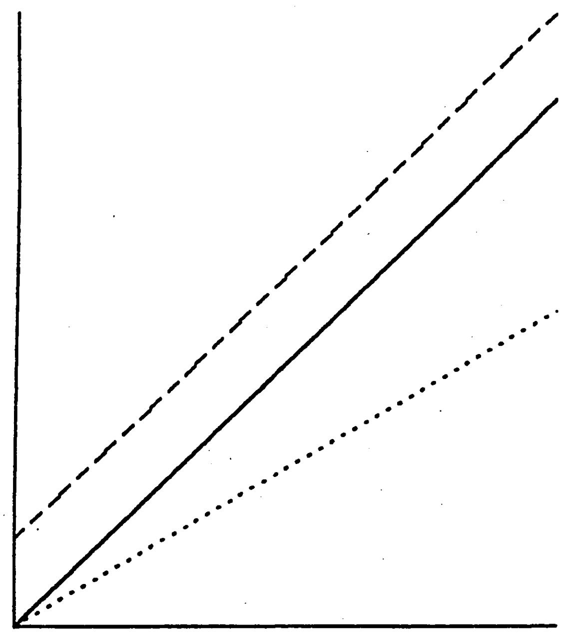


FIGURE 1 Plots illustrating (a) line of identify (solid line), (b) regression with no bias (solid line), (c) regression with fixed bias (dashed line), and (d) regression with proportional bias (dotted line).

ases, respectively (Ludbrook, 1997), and are important for interpreting the results of the comparison.

An additional important feature of the data is the distribution of points around the LPR line. If the points are evenly distributed from the line throughout the full range of the measurements then the error is additive (or homoscedastic). Alternatively, if the points are not evenly spread then the error is multiplicative (or heteroscedastic). The boundaries of the distribution of these errors can be seen in Figure 2. Many statistical tests assume that error is additive, although in the areas of sports medicine and sports science, Nevill and Atkinson (1997) suggest that multiplicative errors are the norm. Data that have multiplicative errors can be subjected to a "variance stability" transformation that will generally make the errors additive thus improving the accuracy of interpretation in any comparison. Reasons proposed for multiplicative error include the following: (a) for smaller values the score must tend toward zero as no negative value can be obtained (assuming in this instance that the experimental design constrains the measurements to being positive); (b) biologically the response is often more varied with larger values; and (c) measurement error increases with larger values.

Comparative analysis of data has been proposed as appropriate via the use of tests, including Pearson Product Moment correlation coefficients (e.g., Gravetter & Wallnau, 1996), intraclass correlation coefficients (e.g., Vincent, 1999), levels of agreement (LOA; e.g., Bland, 1995), least squares regression (LSR; e.g., Gravetter & Wallnau, 1996), and LPR (e.g., Ludbrook, 1997). Although no consensus exists as to the most appropriate test, there are a number of limitations in the

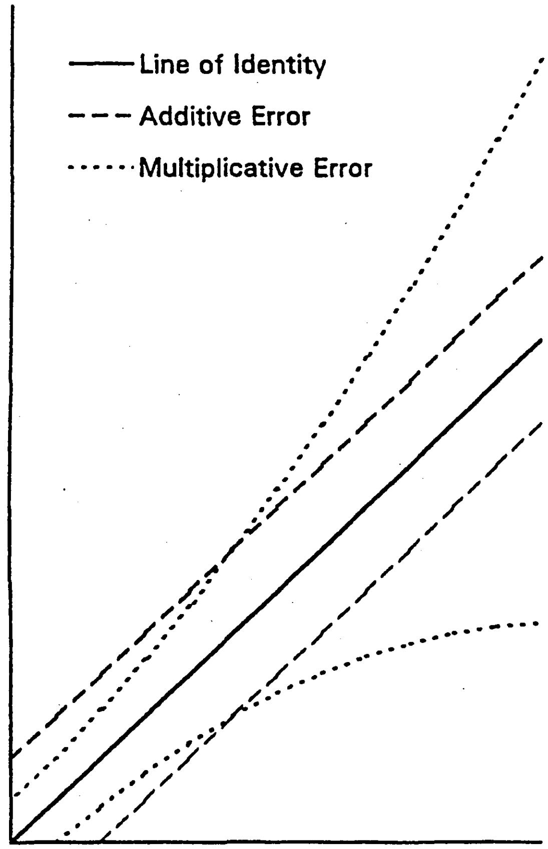


FIGURE 2 Distribution of scores between two repeated measurements indicating the line of identity (solid line) and the boundaries of the distributions of the data when the errors are additive (dashed lines) and multiplicative (dotted lines).

use of all of these tests. The intention is not to cite each of these limitations as they have been provided previously (e.g., Atkinson & Nevill, 1998; Bland & Altman, 1995). Simply, a principal limitation with the majority of these tests is that they do not provide information to assess both fixed and proportional biases. Only LOA, LSR, and LPR that test these biases are appropriate. LSR is not appropriate, however, because it only measures error in one measurement whereas error should be assumed in both measurements. Two different equations would be produced, depending on which measurement the error was minimized in for LSR. In contrast, LPR reduces the error in both measurements and would result in two different-looking equations, depending on which measurement was the dependent variable. However, these two equations are the same as they are transformations of each other.

The assessment of fixed and proportional biases in comparative studies must thus be undertaken using either LOA or LPR. Some researchers favor LOA because the calculation is simple and, because the results are reported in the original units of measurement with 95% boundaries, the interpretation of the results is straightforward. However, fixed and proportional biases often interact, and the output from LOA does not provide the means to account for the independent effect of each. For example, a positive proportional bias and negative fixed bias could interact and result in a zero mean difference. It has therefore been proposed that to account for these related entities LPR should be used because it provides separate measures of these biases. The calculation is more complex and the output is in statistical terms of reference in the form of $y = a + b \cdot x$. LPR may be considered the only "philosophically" correct technique, because it provides a separate assessment of the biases and provides a predictor equation, but it is more sensitive to the spread of the data (although not as much as the more common LSR) and does not work with more than two repeated measurements. Although LOA does not separate the biases, it does work with a small spread of data and can be extended to accommodate more than two repeated measurements.

Assumptions underpinning both LOA and LPR include that the errors should be normally distributed, have a mean value of zero, and be additive, and that there are random errors and equal variance in both sets of measurements. When the normal distribution and additive error assumptions are violated for these tests, log–log transformations can be used to correct for these violations. Nevill & Atkinson (1997) provided strong evidence that "heteroscedastic errors are the norm and, as such, advocate the use of the log transformations when assessing measurement agreement" (p. 318). However, if the normal distribution, linearity, or additive error assumptions are not violated, then these transformations should not be routinely used, because they can be misleading and are more difficult to interpret. Another solution for LPR only is to "weight" the calculation (Ludbrook, 1997), which is beneficial because it is simple and the results remain easy to interpret.

The aim of this study is to demonstrate the use of the LOA and LPR as "acceptable" tests in comparing two measurements and to provide an insight into the interpretation of their results.

METHOD

To illustrate the assessment of reliability, 17 healthy, active, female participants (means [with standard deviations in brackets] were: age, 20.9 [1.1] years; mass, 59.6 [5.5] kg; height, 1.63 [0.06] m) volunteered and provided written informed consent to participate in this study. Peak torque extension (PTE) was obtained on an isokinetic dynamometer (Cybex Orthotron, Medway, MA) for the measurement of maximal voluntary strength of the knee extensors at 2.1 rad/sec. Two measures on the dominant leg were obtained 2 days apart and at the same time of day to limit circadian variations (Table 1).

Linearity of the Data

An initial check in assessing reliability is whether the two measures are linearly related. Production of a scatter plot provides a simple technique to assess this qualita-

TABLE 1
Peak Torque Extension ($N \cdot m$) of the Dominant-Leg Knee-Extensors at 2.1 rad/sec Obtained on 2 Separate Days

	Day 1	*Day 2*	*Mean*	*Difference*	*Predicted*	*Residual*	*Absolute Residual*
	89	95	92	−6	96.56	−7.56	7.56
	113	111	112	2	115.81	−2.81	2.81
	64	76	70	−12	73.71	−9.71	9.71
	108	99	103.5	9	101.37	6.63	6.63
	108	104	106	4	107.39	0.61	0.61
	94	92	93	2	92.95	1.05	1.05
	87	88	87.5	−1	88.14	−1.14	1.14
	96	98	97	−2	100.17	−4.17	4.17
	85	89	87	−4	89.34	−4.34	4.34
	88	84	86	4	83.33	4.67	4.67
	111	104	107.5	7	107.39	3.61	3.61
	76	77	76.5	−1	74.91	1.09	1.09
	111	106	108.5	5	109.79	1.21	1.21
	87	83	85	4	82.13	4.87	4.87
	103	102	102.5	1	104.98	−1.98	1.98
	91	94	92.5	−3	95.36	−4.36	4.36
	80	71	75.5	9	67.69	12.31	12.31
M	93.6	92.5	93.1	1.06	93.6	−0.01	4.24
SD	13.9	11.5	12.5	5.51	13.9	5.47	3.28

Note. Mean = (Day 1 + Day 2)/2; Difference = Day 1 − Day 2; Predicted is Day 1 = 1.2 × Day 2 − 17.7 (from least products regression equation); Residual = Day 1 − Predicted; Absolute residual = |Residual|.

tively, which is important because quantitative techniques can indicate a linear relation for nonlinear data. Figure 3 illustrates a linear trend between the two measures, supported by the Pearson Product Moment correlation coefficient value ($r = 0.92$). Qualitatively, this illustration emphasizes that the data do not lie along the line of identity. This deviation needs to be quantified to assess whether the isokinetic chair is reliable over two separate days for the experimental conditions described.

Analysis of the Errors

The distribution of errors around the regression line is important in providing confidence in the results for accurate interpretation. Inspection of the deviation from the line of identity in Figure 3 can enable a qualitative assessment of the errors. Because the small and large errors appear to be evenly distributed throughout the range of scores, the error can be assumed to be additive. For the LPR this is quanti-

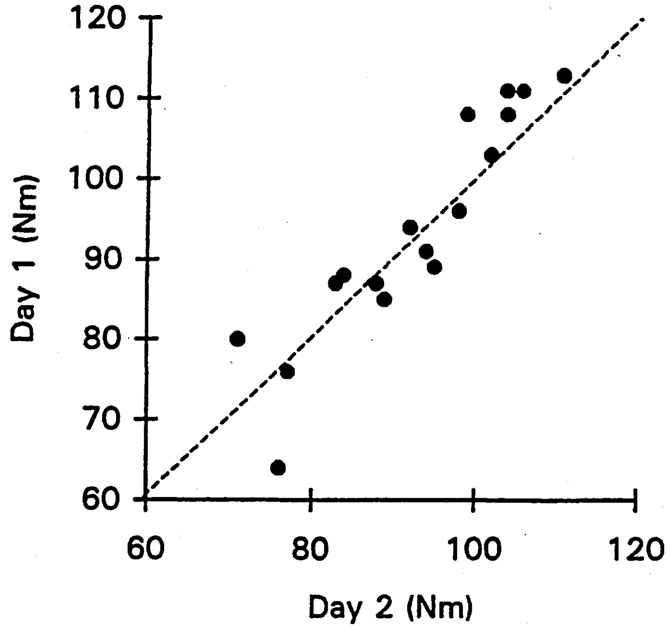


FIGURE 3 Scatter plot to illustrate the linear trend ($r = 0.92$) between the two repeated measurements of PTE (where deviation of the data from the line of identity [dotted line] can be viewed).

tatively supported, as there is no significant correlation between the y values (Day 1) and absolute residuals ($r = -0.48; p \geq .05$). For the LOA this is supported by a random distribution of the differences plotted against the mean score for each pair of measurements (Figure 4). No significant correlation between the differences and the mean score for each pair of measurements confirms that the errors in LOA are additive ($r = 0.43; p \geq .05$).[1] Finally, the error in the LPR should be zero, an assumption that it met because the mean error is $-0.01\ N \cdot m$ (see Table 1).

Normality of the Data and Errors

The use of LPR and LOA are underpinned by the normality assumption. This can be confirmed if the Z scores for the skewness (skewness ÷ SE skewness) and kurtosis (kurtosis ÷ SE kurtosis) are within ±2 (Vincent, 1999). The raw data meet this assumption for both y (Z skewness = −0.51; Z kurtosis = −0.57) and x (Z skewness = −0.38; Z kurtosis = −0.75). In addition, this assumption needs to be met for the residuals of the LPR (Z skewness = 0.64; Z kurtosis = 0.34) and the differences for the LOA (Z skewness = −1.13; Z kurtosis = 0.50), which are all acceptable.

---

[1]Although the correlations are not significant (possibly due to the small sample size), they are large, and careful consideration should be made regarding whether to perform a "variance stability" transformation (i.e., log–log transformation or Weighted LPR) or consider removing any outliers. However, the qualitative assessment suggests the errors are additive and that, because the transformations would make the interpretation more difficult, none of these techniques have been implemented.

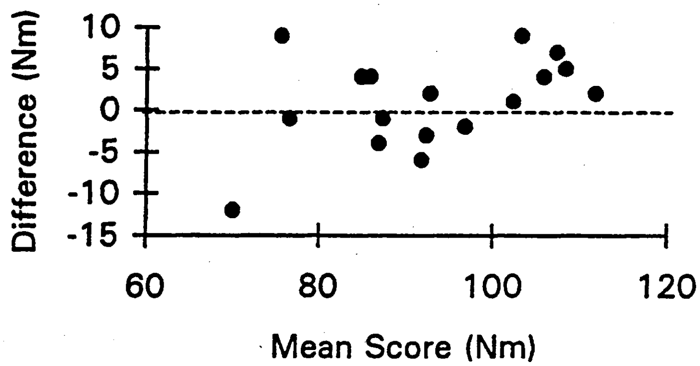


FIGURE 4 Differences between each pair of repeated measurements of PTE plotted against the mean score of each pair for assessment of additive or multiplicative error in the use of LOA.

Calculations

LOA for two sets of measurements were calculated by subtracting the values for Day 2 from those of Day 1 and then calculating the mean (δ) and standard deviation (σ) of the differences. The LOA were calculated as $\delta \pm 1.96 \cdot \sigma$ (Bland & Altman, 1986) with 95% confidence.[2]

LPR of the "Model Expression" form $y = a + b \cdot x$ was calculated using SPSS for Windows (Release 8.0.0; SPSS, 1997) using the "nonlinear regression" and inputting the "user-defined loss function" formula as $[y - (a + b \cdot x)]^2/|b|$ (where the parameters a and b were set to initial values of 0 and 1, respectively).[3] The 95% confidence intervals (CI) in a and b were obtained using the "bootstrap estimates of the standard error" bootstrapping option.[4,5]

RESULTS

The mean PTE (with standard deviations in parentheses) of the knee at 2.1 rad/sec during isokinetic dynamometry obtained for 17 healthy, active, female participants over two separate days was 93.6 (13.9) $N \cdot m$ and 92.5 (11.5) $N \cdot m$ for Days 1 and 2, respectively.

---

[2]For more than two groups, a fixed boundary value can be calculated as $\pm 1.96 \cdot \sqrt{(2 \cdot MS_R)}$ with 95% confidence (British Standards Institute, 1979; Mullineaux, Scott, & Batterham, 1994).

[3]If a weighted LPR were required then the "user-defined loss function" formula would be $[y - (a + b \cdot x)]^2/|b|(x)(y)$.

[4]Because the CI were obtained using bootstrapping, these will vary slightly each time they are calculated. See Zhu (1997) for a tutorial on bootstrapping.

[5]If the CI are not required then the simple calculations are as follows: b is the ratio of the standard deviations of the y and x values, and $a = y' - b \cdot x'$ (where y' and x' are the mean values of y and x, respectively).

The calculation of the LOA was $1.06 \pm 10.80\ N \cdot m$ (95% confidence). The mean difference suggests that there is small fixed bias of Day 1 measuring 1.06 $N \cdot m$ more than Day 2. The 95% spread of the differences of $\pm 10.80\ N \cdot m$ equates to 11.6% of the mean PTE of 93.1 $N \cdot m$. Taking the fixed bias into account, this indicates that Day 1 could be up to 9.74 $N \cdot m$ less (i.e., 1.06 – 10.80) and up to 11.86 $N \cdot m$ more (i.e., 1.06 + 10.80) than on Day 2 (95% confidence). These equate to 10.46% (i.e., 9.74/93.1 × 100) and 12.74% (i.e., 11.86/93.1 × 100) of the mean PTE, respectively. Because these values are large, it can be concluded that the data are not sufficiently reliable.

Results provided for the LPR are the parameter estimates (95% CI). For the regression equation the parameter estimate for the intercept was $a = -17.7\ N \cdot m$ (–37.4 to 2.03) and for the slope was $b = 1.20$ (1.01 to 1.40), as illustrated in Figure 5. The intercept of $-17.7\ N \cdot m$ suggests that Day 1 measures less than Day 2, although because the 95% CI encompasses zero, no fixed bias between the repeated measures can be supported. However, because the CI for the slope does not encompass 1, then proportional bias between the repeated measures exists such that Day 1 measures less than Day 2 with increasing magnitude. It can be concluded that LPR supports that the data are not sufficiently reliable.

CONCLUSIONS

The results of both LOA and LPR support that the measurement of PTE of the knee extensors at 2.1 rad/sec during isokinetic dynamometry for 17 healthy, active, fe-

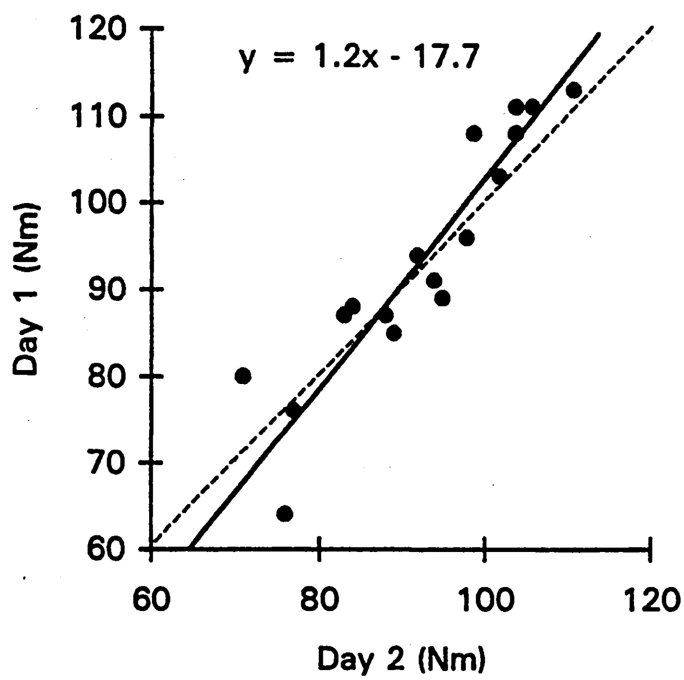


FIGURE 5 Comparison of the line of identity (dotted line) and LPR (solid line) between two repeated measurements of PTE (with the LPR equation displayed).

male participants over two separate days is not sufficiently reliable. LOA indicates an unacceptable difference between the two sets of scores, and although LPR found no fixed bias, it supports the existence of a proportional bias. A direct comparison of the two statistical techniques is not possible because LOA integrates fixed and proportional biases whereas LPR separates these two entities.

LPR possesses some benefits over LOA. First, LPR is "philosophically" correct in that it provides separate measures of fixed and proportional biases, whereas LOA provides an interactive measure that can be misleading. Second, interpretation of the intercept and slope from LPR are unambiguous in that the CI identifies whether the results deviate from the ideal, whereas identification of unacceptable differences is subjective in LOA. Conversely, LOA offers some benefits over LPR: It is easy to calculate; the results are "meaningful" because they are provided in the original units of measurement; and the technique can be extended to accommodate more than two repeated measurements.

LOA and LPR provide methods to statistically compare two measurements, provided the assumptions underpinning the analysis are met or controlled. Methods to control these have been proposed, including log–log transformations (Nevill & Atkinson, 1997) and weighted regressions (Ludbrook, 1997). The choice of test is dependent on whether the appropriate assumptions are met, the number of repeated measurements, and preference in the format of the results obtained. The important factor is that an assessment of reliability is performed so that validity can subsequently be supported (Mullineaux & Bartlett, 1997). These techniques can be extended to making comparisons between, for example, experimental and theoretical data or gold standard and new measurement methods.

ACKNOWLEDGMENT

We thank Rune Söltvedt for his contribution in collecting the data.

REFERENCES

Atkinson, G., & Nevill, A. M. (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Medicine, 26,* 217–238.

Bland, J. M. (1995). *An introduction to medical statistics.* Oxford, England: Oxford University Press.

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet, 1,* 307–310.

Bland, J. M., & Altman, D. G. (1995). Comparing two methods of clinical measurement: A personal history. *International Epidemiological Association, 24,* S7–S14.

British Standards Institute. (1979). *Precision of test methods 1: Guide for the determination and reproducibility of a standard test method (BS 5497, Part 1).* London: Author.

Gravetter, F. J., & Wallnau, L. B. (1996). *Statistics for the behavioral sciences.* Minneapolis, MN: West.

Ludbrook, J. (1997). Comparing methods of measurement. *Clinical and Experimental Pharmacology Physiology, 24,* 193–203.

Mullineaux, D. R., & Bartlett, R. M. (1997). Research methods and statistics. In R. M. Bartlett (Ed.), *Biomechanical analysis of movement in sport and exercise* (pp. 81–104). Leeds, England: British Association of Sport and Exercise Sciences.

Mullineaux, D. R., Scott, M. A., & Batterham, A. M. (1994). The use of "agreement" for assessing digitiser reliability/objectivity. In J. Watkins (Ed.), *Proceedings of the Biomechanics Section of the British Association of Sport and Exercise Sciences* (pp. 17–20). Leeds, England: British Association of Sport and Exercise Sciences.

Nevill, A. M., & Atkinson, G. (1997). Assessing agreement between measurements recorded on a ratio scale in sports medicine and sports science. *British Journal of Sports Medicine, 31,* 314–318.

SPSS. (1997). *SPSS for Windows, Release 8.0.0.* Chicago: SPSS, Inc.

Vincent, W. V. (1999). *Statistics in kinesiology.* Champaign, IL: Human Kinetics.

Zhu, W. (1997). Making bootstrap statistical inferences: A tutorial. *Research Quarterly for Exercise and Sport, 68,* 44–55.

Appendix 3

Mullineaux, D.R., Scott, M.A. and Batterham, A.M. (1994). The use of "agreement" for assessing digitiser reliability/objectivity. In *Proceedings of the Biomechanics Section of the British Association of Sport and Exercise Sciences* (edited by J. Watkins), pp. 17-20. Leeds: BASES.

THE USE OF "AGREEMENT" FOR ASSESSING DIGITISER RELIABILITY/OBJECTIVITY

D.R. Mullineaux, M.A. Scott and A.M. Batterham
the Manchester Metropolitan University

INTRODUCTION

Researchers in the field of sport and exercise sciences frequently implement improper research designs and employ inappropriate statistics. This problem is highlighted by the common practice of a large number of dependant variables being selected without *a priori* theorising and assessed for statistical significance. This "shopping basket" full of dependant variables being "checked out" for statistical significance often results in violations of alpha with a consequent incorrect and misleading statistical analysis.

The employment of the analysis of variance test for assessing the reliability of techniques or measures is a specific problem. The recommendation to estimate reliability with this method (e.g. Baumgartner, 1989; Thomas and Nelson, 1990) and the available intraclass correlation are inappropriate owing to the concept underpinning these tests. These tests aim to identify a difference and a relationship respectively between data sets and do not identify the degree of agreement between them. The use of the "agreement" technique is thus considered more appropriate.

For reliability to be assumed no statistical difference between repeated measures should be found using the analysis of variance. It is suggested by Bates *et al.* (1992) that the interpretation of non-significant findings should be performed with a statistical power analysis. A power analysis will reveal the ability of the test to identify a difference. It is hypothesised that owing to the small differences that will often exist between repeated measures that the power of this test will be small. Hence, if the ability of the test to identify a difference is small, then the criteria for reliability of no statistical difference will have a high probability of being erroneously achieved.

The objective of this study was to compare the use of the analysis of variance method with the "agreement" method for assessing reliability. The process of digitising, frequently assessed for reliability, was chosen as an example for this demonstration. However, the comparison could have been performed with any sets of data that measure the same quantity by either the same or different techniques.

METHODOLOGY

Thirty hockey players of mixed ability and sex volunteered to participate in this study. A Panasonic F-15 camera was positioned with the optical axis of the lens aligned on the penalty spot and perpendicular to the length of a hockey pitch. Each subject was filmed in two dimensions performing one hockey flick each.

The ball in each sequence was digitised at 50 Hz by two experienced digitisers on an Archimedes 440 micro computer running the Kine System (Bartlett and Bowen, 1993). For both digitisers the sequences were smoothed and analysed for the ball's release speed.

For each subject the two values obtained for the ball's release speed were used to assess the digitiser objectivity via the one-way analysis of variance (ANOVA) reliability model described by Baumgartner (1989) and the "agreement" method described by Bland and Altman (1986).

The ANOVA model required the data be input into an one-way within design. Assessment of reliability was determined by establishing whether a significant difference existed between data sets and by substituting data from the analysis of variance table into equation 1 to calculate the intraclass correlation (R) between data sets.

$$R = \frac{MS_B - MS_W}{MS_B}$$

[Equation 1]

(From Baumgartner, 1989)

The calculations for "agreement" involved, firstly subtracting the data values found by one digitiser from those found by the other digitiser and then calculating the mean (δ) and standard deviation (σ) of these differences. By substituting δ and σ into equation 2, the boundaries of agreement between the digitisers were determined.

$$\textit{Boundaries of Agreement} = \delta \pm 2.\sigma$$

[Equation 2]

(From Bland and Altman, 1986)

RESULTS

The mean release speed was found to be 10.99 ± 2.55 $m.s^{-1}$ by digitiser 1 and 10.95 ± 2.57 $m.s^{-1}$ by digitiser 2.

The analysis of variance results (table 1) demonstrated that the F value of 2.53 was less than the tabled value of 4.18 (Cohen and Holliday, 1982) at the 0.05 significance level. Thus it was concluded that there was no significant difference between the ball's release speed found by the two digitisers ($p < 0.05$). A *post hoc* power analysis (for a two group design, n = 30, effect size index =

0.014 and $\alpha = 0.05$) revealed that there was less than a 6% chance of the ANOVA finding significance (extrapolated from Cohen, 1988).

Table 1: Analysis of Variance Results.

Source of variation		DF	SS	MS	F
Between total	(B)	29	379.26	13.08	
Within total	(W)	30	0.53	0.018	
Digitisers		1	0.04	0.04	2.53
Residual	(R)	29	0.49	0.017	
Sub-total		30	0.53		
Grand total		59	379.79		

By substituting values from table 1 into equation 1 the intraclass correlation between the data sets was calculated as 0.99.

For the "agreement" method the mean of the differences was 0.05 $m.s^{-1}$ and the standard deviation of the differences was 0.18 $m.s^{-1}$. The boundaries of agreement were that the release speed of the ball could be found by digitiser 2 as 0.32 $m.s^{-1}$ lower or 0.42 $m.s^{-1}$ greater than values found by digitiser 1 (95% confidence level).

DISCUSSION

The objective of the methodology was to provide a basis for comparing techniques of assessing reliability. Limitations within the research design, such as the sampling bias, are therefore peripheral to the objectives of the study and considered irrelevant.

The non-significance ($p < 0.05$) and intraclass correlation of 0.99 between data sets calculated from the ANOVA model would suggest that there is reliability between digitisers. The boundaries of agreement, where digitiser 2 could find the ball's release speed as 0.32 $m.s^{-1}$ lower or 0.42 $m.s^{-1}$ greater than by digitiser 1, would also suggest that there is reliability between digitisers.

With this worked example a conclusion that the digitising was reliable could be made from both techniques. However, the ANOVA model is not recommended owing to two principal limitations. Firstly, it is conceptually incorrect to use difference and correlation tests for assessing reliability. Secondly, there is often a low statistical power for the ANOVA test. For this data there was a 6% power which equated to a 94% chance of a type II error occurring. Frequently this low power for ANOVA will result in no significant difference with a high type II error probability, therefore, confounding any assessment of reliability.

It is recommended that boundaries of agreement should be reported for describing reliability. This is owing to "agreement" providing more useful information than ANOVA by quantifying the potential difference between measures. Further, the use of "agreement" is considered conceptually correct and does not have the problem, as with ANOVA, of a high probability of a type II error occurring.

For situations where more than two sets of repeated measures exist, an overall agreement boundary can be calculated by substituting data from the ANOVA table into equation 3.

$$Agreement\ Boundary = \pm\ 2\ \times \sqrt{2 \times MS_R}$$ [Equation 3]

(Adapted from Fleiss, 1986)

For the current data, the agreement boundary was calculated from equation 3 as $x \pm 0.368$ m.s^{-1}, where x is any measured data value. This value is similar to two times the standard deviation of the differences ($2 \times \sigma = 0.367$ m.s^{-1}) found by the "agreement" method. The exclusive use of ANOVA combined with equation 3, however, is not recommended for cases of two data sets as the "agreement" method described provides more information which may be used in an assessment of reliability.

REFERENCES

Bartlett, R.M. and Bowen, T.M. (1993). *Kine System User Guide*. Alsager: Crewe+Alsager Faculty of *the* Manchester Metropolitan University.

Bates, B.T., Dufek, J.S. and Davis, H.P. (1992). The effect of trial size on statistical power. *Medicine and Science in Sports and Exercise*, **24**, 1059-1068.

Baumgartner, T.A. (1989). Norm-referenced measurement: reliability. In *Measurement Concepts in Physical Education and Exercise Sciences* (edited by M.J. Safrit and T.M. Wood), pp. 45-72. Champaign: Human Kinetics.

Bland, J.M. and Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, **1**, 307-310.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioural Sciences*. London: Lawrence Erlbaum.

Cohen, L. and Holliday, M. (1982). *Statistics for Social Scientists*. London: Harper and Row.

Fleiss, J.L. (1986). *The Design and Analysis of Clinical Experiments*. New York: Wiley.

Thomas, J.R. and Nelson, J.K. (1990). *Research Methods in Physical Activity*. Champaign: Human Kinetics.

Appendix 4

Batterham, A.M., Barnes, C.A. and Mullineaux, D.R. (1999a). Modelling the influence of thigh muscle cross-sectional area on vertical jump performance in young professional soccer players. *Journal of Sports Sciences*, **17**, 807-8.

Journal of Sports Sciences, 1999, 17, 807–840

CONFERENCE COMMUNICATIONS

Communications to the Fourth World Congress of Science and Football

Held at the University of Technology, Sydney, NSW, Australia, 22–26 February 1999

Modelling the influence of thigh muscle cross-sectional area on vertical jump performance in young professional soccer players

A.M. Batterham, C.A. Barnes and D.R. Mullineaux

Centre for Sport Performance, University of Teesside, Middlesbrough, UK

The ability to generate a high muscular power output underpins many actions in soccer. To make valid comparisons of performance between players, the influence of body size must be appropriately partitioned out. The aim of this study was to establish the most appropriate way to control for the influence of thigh cross-sectional area for a test of vertical jumping power in a group of elite youth soccer players.

Nine healthy male professional soccer players (mean ± s: age 18.1 ± 0.6 years, height 181 ± 6 cm, body mass 70.5 ± 5.0 kg) volunteered to participate. After a standardized warm-up, each player performed three vertical jumps with a counter-movement on a force platform (AMTI OR6-7, Watertown, MA, USA). Vertical force was sampled at 1000 Hz using a data acquisition system (AMLAB, Lane Cove, Australia). Power output was derived from the product of gross vertical force and vertical velocity of the centre of mass (the impulse of the net vertical force divided by body mass). The highest value for the three trials was recorded as peak power output. Using magnetic resonance imaging (Horizon LX 1.5 T, General Electric, Milwaukee, WI, USA), thigh muscle cross-sectional area was determined from an axial T_1-weighted scan at the mid-point of the femur.

The combined cross-sectional area of the left and right thigh muscles was 0.33 ± 0.03 m^2. The peak power output for the counter-movement jump was 4981 ± 380 W. The log-linear relationship between cross-sectional area and peak power output when expressed in the form $y = a \cdot x^b$ gave a b exponent of 0.67 ($r^2 = 0.73$, 95% confidence intervals = 0.54–1.27). The 95% confidence intervals for the b exponent in this study spanned the expected value from dimensionality theory of unity. We have shown that, when controlling for body size in inter-individual comparisons of vertical jumping performance among soccer players, appropriate scaling techniques should be adopted.

Appendix 5

Batterham, A.M., George, K.P. and Mullineaux, D.R. (1995). Relationship between heart size and body dimensions: an allometric scaling approach. *Medicine and Science in Sports and Exercise*, **27**, S158.

886 RELATIONSHIP BETWEEN HEART SIZE AND BODY DIMENSIONS: AN ALLOMETRIC SCALING APPROACH

A.M. Batterham, K.P. George, and D.R. Mullineaux,
Division of Sport Science, The Manchester Metropolitan University, Crewe and Alsager Faculty, Alsager, England.

Many physiological variables have been found to relate to body mass (BM) or fat free mass (FFM) in a log-linear fashion, according to the general allometric equation $y = a.BM^b$. Traditionally however, a ratio normalization approach has been adopted assuming that the exponent 'b' = 1. This practice is valid only if the criteria of data linearity and zero intercept are met. The purpose of this study therefore was to determine the proper relationship between left ventricular mass (LVM) and body dimensions in 142 men and women (age range 18-40 years). Echocardiography (M-mode) was used to define posterior wall thickness, septal thickness, and left ventricular internal dimensions at end diastole. LVM (g) was then calculated via a standard predictive equation. FFM was estimated via bicep, tricep, subscapular and suprailiac skinfolds. Initial data evaluation demonstrated that the criteria for ratio normalization had not been met. Allometric scaling revealed strong log-linear relationships between body size variables and LVM (lnBM: lnLVM, $r = 0.73$; lnFFM:lnLVM, $r = 0.83$). The proper value of b was found to be 1.5 for BM and 1.45 for FFM. Hence it appears that LVM does not represent a constant proportion of BM. Use of a simple ratio standard (LVM/ BM or LVM/FFM) to compare individuals or groups would therefore penalize less massive subjects, and potentially lead to spurious conclusions. In this sample, the statistically correct scaling method would be $LVM/BM^{1.5}$ or $LVM/FFM^{1.45}$. It is suggested that sample specific allometric equations should be generated in order to properly adjust echocardiographic indices of cardiac dimension.

Appendix 6

Batterham, A.M., George, K.P. and Mullineaux, D.R. (1997). Allometric scaling of left ventricular mass and body dimensions in males and females. *Medicine and Science in Sports and Exercise*, **29**, 181-6.

Allometric scaling of left ventricular mass by body dimensions in males and females

ALAN M. BATTERHAM, KEITH P. GEORGE, and DAVID R. MULLINEAUX

Department of Exercise and Sport Science, the Manchester Metropolitan University, Crewe and Alsager Faculty, Alsager, ST7 2HL, ENGLAND, and School of Social Sciences, University of Teesside, Middlesbrough, TS1 3BA, ENGLAND

ABSTRACT

BATTERHAM, A. M., K. P. GEORGE, and D. R. MULLINEAUX. Allometric scaling of left ventricular mass by body dimensions in males and females. *Med. Sci. Sports Exerc.* Vol. 29, No. 2, pp. 181–186, 1997. Physiological variables must often be scaled for body size differences to permit meaningful comparisons between subjects or groups. This study aimed to determine the proper relationship between body dimensions and left ventricular mass (LVM) *via* allometric scaling (AS) in 142 subjects (78 males, 64 females; ages 18–40). A cubic formula was used to estimate LVM from wall thicknesses and left ventricular internal dimensions derived from M-mode echocardiography. Fat free mass (FFM) was predicted from anthropometry. "Best compromise" allometric equations ($y = a{\cdot}x^b$) revealed a common body mass (BM) exponent of 0.78 (95% CI, 0.65–0.91). The widely adopted ratio scaling (RS) method assumes that the exponent $b = 1$. In this sample, use of RS would penalize heavier subjects by overcorrecting for BM. The equivalent mean FFM exponent of 1.07 was not different from unity (95% CI, 0.92–1.22). Hence, RS using BM would appear to penalize those subjects who are heavier owing to excess fat not excess FFM. Gender differences in LVM were 70, 44, and 18%, for absolute values per $BM^{0.78}$ and per $FFM^{1.07}$, respectively ($P < 0.05$). This reveals quantitative differences in heart size independent of body dimensions. We conclude that sample specific AS permits meaningful intersubject or intergroup comparisons.

POWER FUNCTION ANALYSIS, ALLOMETRY, GENDER DIFFERENCES, ECHOCARDIOGRAPHY.

Recently there has been renewed interest in the importance of body size as a potentially confounding influence in studies of physiological function (3). This paper will focus on how body dimension relates to cardiac dimension. First, the concept of scaling will be addressed, detailing various methods that have been adopted to relate body size variables to physiological variables. Second, experimental data will show the proper relationships between echocardiographically determined left ventricular mass and body dimension in a sample of males and females. Third, the general implications of these findings for scaling in sports echocardiography will be discussed.

0195-9131/97/2902-0181$3.00/0
MEDICINE AND SCIENCE IN SPORTS AND EXERCISE®

Submitted for publication June 1995.
Accepted for publication June 1996.

SCALING

A wide range of physiological variables are influenced by body dimensions, such that increments in body size result in an increase or decrease in the physiological response. To conduct meaningful interindividual or intergroup comparisons, it is often essential to partition out this confounding influence *via* the derivation of a relative index that is allegedly size independent. "Scaling" involves the normalization of a physiological variable (y) for differences in a body dimension variable (x).

Numerous approaches have been used to try to solve this problem. By far the most common approach is the use of ratio standards (RS) of the form $y = b{\cdot}x$. Scaling *via* RS involves dividing the absolute physiological (dependent) variable by a body dimension (independent) variable such as body mass (BM), fat free mass (FFM), or body surface area (BSA). The theoretical and mathematical flaws in this method were recognized nearly fifty years ago (22), and yet its use remains widespread. The RS approach assumes a linear relationship between the anthropometric and physiological variables with the line of identity passing through or close to the origin. Violation of these assumptions may lead to erroneous conclusions in research using the RS scaling method. An extension of RS involves the use of regression standards (RES) of the general form $y = a + b{\cdot}x + \in$, where b represents the gradient of the trendline, a the y-intercept, and $\in$ the additive residual error term. Unfortunately, although often providing a better fit to the data with a reduction in residual error, positive intercepts are common, indicating that someone of zero body mass would exhibit a physiological response (3). Therefore, extrapolation of the regression line beyond the actual range of data must be avoided.

ALLOMETRY

Research in comparative physiology has consistently revealed that many physiological variables relate to body size in a log-linear, rather than linear fashion (20). Plotting log-transformed data thus results in a straight-line fit.

This relationship is best described by a power function or allometric equation of the following general form:

$$y = a \cdot x^{b} \cdot \in$$

where y is the physiological variable, x is the body size variable, and a, b, and $\in$ represent the proportionality coefficient, the size exponent, and the multiplicative error term, respectively. The a and b values are derived from the log-log plot, with b representing the slope of the line and a the intercept at unity body mass (since the natural logarithm of $1 = 0$, that is, the origin on the log-log plot).

The most commonly adopted body size variable has been BM, thus enabling mass exponents (b) for several physiological variables to be identified. For example, it has been shown that maximum oxygen uptake is properly scaled by the power function ratio $\dot{V}O_{2max}$ (absolute)/$BM^{2/3}$ (16). This suggests that smaller individuals have a higher $\dot{V}O_{2max}$ per unit body mass than larger individuals because the mass exponent is less than one. Expressing $\dot{V}O_{2max}$ as a RS ($ml \cdot kg^{-1} \cdot min^{-1}$) to compare individuals or groups would thus penalize larger individuals by overcorrecting for body size.

Allometric models have provided a better fit to the data with less residual error and thus arguably represent a more valid method for producing a dimensionless physiological variable (16). An advantage of such power function approaches over regression standards models is the assumption of a multiplicative error term. Many physiological data sets indicate heteroscedasticity, a tendency for a greater spread in the data as body size increases, violating the assumption of constant error variance (homoscedasticity) in regression standards models. This further supports the statistical validity of allometric modeling (3).

RELATIONSHIP BETWEEN HEART SIZE AND BODY SIZE

There is little evidence in the literature of attempts to examine properly the relationship between left ventricular mass and body dimension indices in humans. The extant data indicate that cardiac dimensions may be closely correlated with fat free mass in particular, suggestive of a possible relationship between skeletal and cardiac muscularity (12). Animal studies (18) have revealed that the relationship between total heart mass (HM) and body mass (BM) is described best by the following equation:

$$HM = 0.006 \cdot BM^{0.98} \pm 0.02$$

The exponent of 0.98 essentially equals unity. Hence, HM in a range of mammals represents a constant proportion of BM (0.6%). It has been assumed that a similar linear, proportional relationship exists in humans and that cardiac dimensions can therefore be properly scaled by RS methods constructing a LVM/BM ratio.

Use of RS scaling has been prevalent in echocardiographic studies in sports cardiology, particularly relating to the athletic heart (12). Attempts to document training specific adaptations in cardiac dimensions are clearly dependent upon the appropriate normalization of absolute heart size data for groups of disparate body size, e.g., weight lifters and endurance athletes. It has been suggested that use of a HM/BM index may be unsatisfactory as the mass of the organ is small relative to the mass of the body. Therefore, response of the organ weight may not be proportional to that of the whole body (17). In addition, failure to meet the RS assumptions of linearity and zero intercept would result in inappropriate scaling.

The correct relationship between body size and echocardiographic indices of cardiac dimension remains to be established. The purpose of this study is to examine the proper allometric relationships between left ventricular mass and body mass and fat free mass in a human subject sample. In addition to informing scaling practice, allometry may raise interesting questions concerning gender differences in heart size-body size relationships and encourage a reevaluation of training specific changes in echocardiographic studies.

METHODS

Subjects

One hundred and forty two subjects volunteered for the study (78 male, 64 female, mean age 22 ± 1.5 yr, range 18–40). Subjects were screened medically with a standard laboratory questionnaire and all were found to be "apparently healthy," asymptomatic and free from cardiovascular disease and major risk factors for coronary heart disease. Additional exclusion criteria included obesity and the chronic use of medications that may influence resting echocardiographic dimensions. Previous testing in the same laboratory had revealed no evidence of resting or exertional hypertension or electrocardiogram abnormalities.

The sample was drawn from a population of undergraduate sport and exercise science students who appeared relatively homogeneous with respect to habitual physical activity. A simple, "global" physical activity self-assessment tool was administered in a personal interview. The instrument was modified from that used in the Allied Dunbar National (England) Fitness Survey (1), with the frequency and intensity of 20-min plus sessions in the previous 4 wk documented. An indication of lifetime physical activity participation was obtained from the proportion of years since age 14 that the subject had participated regularly in physical activity. All subjects were moderately to highly recreationally active, with 55% of males and 53% of females reporting an average of three or more 20-min sessions per week at a "vigorous" (7.5 $kcal \cdot min^{-1}$) intensity. The remainder reported an equivalent frequency of "moderate" (5 kcal·min^{-1}) intensity activity. Group equivalence for physical activity indices was confirmed via the Kolmogorov-Smirnov two-sample test ($P > 0.05$) suggesting that males and females represent the same population with respect to the distribution of physical activity indices.

Procedures were in accordance with the American College of Sports Medicine policy statement regarding the use of human subjects and written informed consent (2). A cross-sectional "snapshot" design was used, with subjects visiting the laboratory on one occasion.

Procedures

Echocardiography. A Hewlett Packard (Andover, MA) Sonos 100 ultrasound imaging system (2.5 Mhz transducer) in sector 2-D mode was used to image a longitudinal axis view of the left ventricle from the parasternal window. M-mode recordings were derived from a cursor line crossing the left ventricle at the tips of the mitral valve leaflets. All echocardiograms were conducted and analyzed by a single experienced technician. The following measurements were made in centimeters according to the Penn convention (10): septal and posterior wall thicknesses at end diastole (ST and PWT) and left ventricular internal dimensions at end diastole (LVIDd). All readings were obtained at the peak of a simultaneous EKG R-wave, with subjects in the supine or left lateral decubitus position. Measurements represented an average of 3–5 heart cycles.

Left ventricular mass (LVM, g) was estimated by means of the regression corrected cubic formula of Devereux and Reichek (10). This method involves the subtraction of internal LV volume from total LV volume, multiplied by an assumed constant for cardiac muscle density of 1.04. The major limitation of this approach lies in the calculation of LV volume by means of cubing obtained LV dimensions. Clearly, any measurement errors in ST, PWT, or LVIDd will inflate exponentially when estimating LVM. The 2.5 Mhz transducer represents a compromise between resolution and penetration, with optimal resolution of approximately 0.7–1.4 mm. A 1-mm error in PWT measurement, for example, could result in a 15% error in LVM estimation (17). Notwithstanding these limitations, Reichek and Devereux (19) reported a strong correlation between echocardiographic estimation of LVM and LVM determined at autopsy (r = 0.96). Moreover, the aim of the current study was to examine relationships between LVM and body dimensions using procedures commonly adopted in the extant literature (12) rather than criterion gold standard methods.

Body composition. Percent body fat was estimated by calculating the sum of bicep, tricep, subscapular, and suprailiac skinfolds (11). The mean skinfold of three rotations that agreed within 10% was used for subsequent analyses. Total body mass (BM ± 0.1 kg) and fat percent were used to partition body mass into its fat mass and fat free mass components.

Data analysis. Initially, it is essential to verify the inappropriateness of the RS approach before progressing to the allometry. Linearity checks were performed on BM against LVM and FFM against LVM. Scaling *via* RS can only be adopted if Tanner's "special circumstance" is satisfied (22). The coefficient of variation for the body dimension variable (x) divided by the coefficient of variation for LVM (y) must equal the Pearson product moment correlation between the two variables: Vx/Vy = r x,y (22). If this assumption is not met, RS scaling may lead to spurious conclusions.

Allometry. Prior to identifying a scaling index common to both genders, similarity of slopes of the relationship between body size and LVM must be confirmed. Significant gender differences found in the b exponents (LVM = a·BM^b (or FFM^b)) would preclude intergroup comparison of scaled LVM, as it would indicate that the groups were qualitatively different. Commonality of b exponents was tested by including a gender × ln BM (or ln FFM) interaction term in a multiple log-linear regression model:

$$ln\,LVM = ln\,a + d\cdot(gender \times ln\,BM) + c\cdot gender + b\cdot ln\,BM + ln \in$$

The interaction term for both ln BM and ln FFM was not significant ($P > 0.05$) indicating that the b exponents were similar between groups. A common "best compromise" b exponent was then fitted by including "gender" as a predictor variable alongside BM or FFM in a multiple allometric regression model (16):

$$LVM = a\cdot(GENDER^c)\cdot BM^b\,(or\,FFM^b)\cdot \in$$

To derive the power function exponents, gender was entered as a dummy variable (males coded 0, females coded 1) in a log-transformed model:

$$ln\,LVM = ln\,a + c\cdot gender + b\cdot \ln \mathrm{BM}\ (\text{or } \ln \mathrm{FFM}) + \ln \in$$

The model provides a solution for a single b exponent, isolating the "gender independent" influence of body size on LVM. In addition, the equation allows for the derivation of adjusted values of the proportionality coefficient a, isolating the effects on LVM owing only to gender. With a common b exponent, a values can thus be compared to test for size-independent gender differences in LVM (alternatively, using the best compromise" b exponent, the power function ratios LVM/BM^b or LVM/FFM^b can be compared with exactly the same result).

All analyses were carried out using the SPSS 6.0 for Windows (SPSS Inc., Chicago, IL) statistical package, providing the power function equations and 95% confidence intervals for the b exponents. Gender comparisons of LVM were conducted with independent t-tests for absolute values and the body size corrected values (using

TABLE 1. Descriptive data for males ($N = 78$) and females ($N = 64$), means ± SD.

	MALES	FEMALES
BM (kg)	75.8 ± 9.4*	61.0 ± 9.4
FFM (kg)	66.0 ± 6.6*	46.9 ± 5.7
LVIDd (cm)	5.36 ± 0.4*	4.9 ± 0.3
ST (cm)	0.97 ± 0.13*	0.77 ± 0.1
PWT (cm)	0.97 ± 0.14*	0.73 ± 0.12
LVM (g)	231 ± 52*	135 ± 33

BM = body mass; FFM = fat free mass; LVIDd = left ventricular internal dimension at end diastole; ST = septal thickness; PWT = posterior wall thickness; LVM = left ventricular mass.
* Different from females ($P < 0.05$).

power function ratios constructed from the common "best compromise" b exponents). The alpha level adopted for significance was $P < 0.05$.

RESULTS

Table 1 shows the descriptive data for BM, FFM, and LVM. Males were on average 14.8 kg heavier than females, with 19.1 kg greater FFM. Absolute LVM was approximately 70% greater in males than in females.

Linearity checks. All checks revealed that the criteria for RS normalization of linearity and zero intercept had not been met. In all cases it was found that Vx/Vy did not equal r x,y. Use of RS scaling in the current study could thus distort the data.

Allometry. Kolmogorov-Smirnov one-sample tests revealed that the log-transformed dependent and independent variables, together with the allometric model residuals, were normally distributed ($P > 0.1$). In addition, no correlation was found ($P > 0.05$) between the absolute residual and the predictor variable (ln BM or ln FFM), indicating that the assumption of homoscedasticity for the log-linear allometric model was satisfied. The allometric power function equations are reported for BM against LVM and FFM against LVM. Kolmogorov-Smirnov two-sample tests revealed that males and females represented different populations with respect to the frequency distribution of left ventricular mass ($P < 0.05$). This further confirmed that multivariate allometry was warranted to identify common exponents. One test of the ability of the allometric model to correctly partition out the influence of body size is to correlate LVM/BM^b (or FFM^b) with BM (or FFM). The correlation should be close to zero if the power function ratio has properly scaled the data. That is, there should be no relationship between relative LVM and body size variables. Correlations between the ratio standards LVM/BM and BM, and LVM/FFM and FFM are presented for comparison. If the power function correlations are closer to zero than the RS, this represents superior scaling of the data.

Body mass against left ventricular mass

$$LVM = 7.7 \cdot (GENDER^{-0.38 \pm 0.04}) \cdot BM^{0.78 \pm 0.13}$$

$$(R^2 = 0.69, P < 0.05)$$

The negative coefficient isolated for gender indicates the anticipated relationship–as gender tends towards zero (males) LVM increases. Proportionality constants (a) can be adjusted for the new common b exponent to compare the LVM of males and females independent of body mass:

$$MALE: \quad LVM = 7.7 \cdot BM^{0.78}$$

Correlation checks

$$(r\,BM, LVM/BM^{0.78} = 0.06; r\,BM, LVM/BM = -0.06)$$

$$FEMALE: \quad LVM = 5.3 \cdot BM^{0.78}$$

Correlation checks

$$(r\,BM, LVM/BM^{0.78} = -0.05; r\,BM, LVM/BM = -0.17)$$

The a values reveal that, independent of body mass, males possess approximately 44% greater LVM ($P < 0.05$). The correlation checks reveal that the expression of LVM/BM$^{0.78}$ did not penalize male or female subjects. For males, the power function ratio and the RS correctly partition out the influence of BM. For females, however, the correlation checks reveal that the RS approach results in a weak negative correlation, indicating that as body mass increases relative LVM decreases. The RS thus overcorrects for BM in females, penalizing heavier individuals in intersubject comparisons.

Fat free mass against left ventricular mass

$$LVM = 2.6 \cdot (GENDER^{-0.18 \pm 0.1}) \cdot FFM^{1.07 \pm 0.15}$$

$$(R^2 = 0.71, P < 0.05)$$

Similar to the results for BM, a negative gender coefficient was again revealed, indicating that independent of FFM males tend to have a higher LVM than females. The common b exponent of 1.07 results in the following adjustments to the proportionality constant a:

$$MALE: \quad LVM = 2.6 \cdot FFM^{1.07}$$

Correlation checks

$$(r\,FFM, LVM/FFM^{1.07} = 0.12; r\,FFM, LVM/FFM = 0.16)$$

$$FEMALE: \quad LVM = 2.2 \cdot FFM^{1.07}$$

Correlation checks

$$(r\,FFM, LVM/FFM^{1.07} = -0.06; r\,FFM, LVM/FFM = -0.04)$$

The a values demonstrate that independent of FFM, males possess approximately 18% greater LVM than females ($P < 0.05$). The correlation checks reveal that the common power function correctly partitions out the influence of FFM for the female subjects. A weak positive correlation of 0.12 for the male sample suggests that the power function is slightly undercorrecting for FFM, thus exerting a minor penalty on smaller males in intersubject comparisons. Note, however, that the allometric scaling is still superior to ratio scaling in the same sample. For the females, both the allometric and the RS correctly account for FFM differences, with correlations not different from zero ($P > 0.05$).

DISCUSSION

For a scaling technique to correctly partition out the influence of body size on a particular physiological variable, the scaled variable should be independent of body size. The findings demonstrate the statistical validity of the allometric scaling approach, with correlations between the scaled variable and the body dimension variable close to zero in all cases. In the female sample, ratio scaling of LVM resulted in a negative correlation with BM of −0.17. This overcorrection for BM would penalize larger female subjects in within-gender comparisons. The opposite effect occurred in the male sample for FFM. Ratio scaling of LVM resulted in a positive correlation with FFM of 0.16. This undercorrection would conversely penalize smaller subjects in interindividual comparisons. The findings demonstrate the utility of allometric scaling in this sample. However, Schmidt-Nielsen (20) stated that the equations cannot be extrapolated beyond the range of data on which they are based. Therefore, it is recommended that sample specific allometry be conducted in all studies where scaling is required. In the current study, intersubject and intergroup comparisons are best conducted using power function ratios constructed from the common best compromise *b* exponents identified from the multiple regression model.

The best compromise mean *b* exponent for BM was found to be 0.78 (95% confidence interval, 0.65–0.91). This exponent is different from unity ($P < 0.05$), contrary to the findings of Prothero for a range of mammals (18). The mass exponent indicates that LVM increases with body mass at a lesser rate than that predicted from simple linear proportionality. As the exponent is less than one, use of a simple ratio standard would appear to overcorrect for BM and exert a penalty on larger subjects. The theory of geometric similarity (20) indicates that as body surface area (BSA) is proportional to the square of height and body volume (BV) is proportional to height cubed, it follows that BSA is proportional to $BV^{2/3}$. As body density is approximately equal to unity, BSA can be assumed to be proportional to $BM^{2/3}$. It can be seen that the mass exponent identified in the current study is not different from 2/3 ($P > 0.05$). It would appear that LVM is therefore proportional to BSA. This relationship has been documented in numerous studies (9,13). However, some authors (25) have urged caution in scaling LVM using BSA, as the index may be confounded by differences in body composition. Clearly, any two individuals may present with similar BSA, yet differ widely in FFM. This may be of great importance given a proposed link between skeletal and cardiac muscularity (12).

The scaling of LVM for differences in FFM offers additional insight. The "best compromise" mean *b* exponent of 1.07 (95% confidence interval, 0.92–1.22) is not different from unity, indicating that the relationship between LVM and FFM is close to constant proportion. Considering the BM and FFM exponents together, it would appear that ratio scaling of LVM per BM penalizes subjects who are heavier as a result of excess body fat, not excess FFM. Strikingly similar findings have been reported recently *via* allometric modeling of peak or maximal oxygen uptake in prepubertal children (4), adult women (23), and older adults (7). In these studies, the FFM exponent was not different from unity, whereas the BM exponent was significantly less than one. These findings, together with those of the current study, lend indirect support to the documented interrelationships between cardiac dimension, skeletal muscle mass, and functional capacity (14). It appears that in echocardiographic studies requiring scaling body composition estimates must be secured. Correction of absolute cardiac dimensions by BSA or BM may be problematic because of variance in body fat percentage.

The gender difference in absolute LVM of 70% ($P < 0.05$) exceeded that reported in cross-sectional studies of 41–52% (5,9). This may be a result of disparities in sample characteristics (age, activity history, genetic factors) and/or specific methods employed (including different conventions adopted for LVM estimation). Allometric normalization of absolute LVM values for BM and FFM reduced the gender differences to 44 and 18%, respectively. These differences remained significant ($P < 0.05$). Independent of body size and composition then, males in this sample possessed larger left ventricular masses than females. Although Schmidt-Nielsen (20) issues a *caveat* that allometry is descriptive and does not represent biological laws, it is possible to briefly postulate mechanisms for these gender differences.

The proposed link between cardiac and skeletal muscularity suggests a possible dual role for testosterone. Testosterone has been used as a marker to reflect the general anabolic status of the body (8) and may positively influence cardiac growth. Gonadectomized male rats have been reported to have lower heart weights than controls (15). This regression was reversed with testosterone replacement. Several authors have reported lower basal levels of testosterone in women compared with men. In addition, there may be a lack of post-exercise testosterone spiking in women (24) although contradictory evidence exists (6).

In addition to gender differences relating to testosterone, women have higher circulating levels of estrogens. Receptor sites for estrogen have been identified in cardiac myocytes, indicating that the heart may be a target organ for estradiol (21). Estrogen may act as a testosterone antagonist in attenuating cardiac growth. This mechanism may be linked to the suggested cardioprotective influence of estrogen (12), based on epidemiological evidence of gender differences in the incidence of coronary heart disease.

CONCLUSIONS

This study has examined the proper relationship between left ventricular mass and indices of body dimension in a human subject sample. Quantitative gender differences in cardiac dimension, independent of body size and composition, have been identified. The demonstration of the value of allometric scaling raises interesting questions regarding previous studies in sports cardiology. Echocardiographic data comparing groups of disparate gender, training status, and body size, may need to be reevaluated with more appropriate scaling techniques. It is hoped that our findings will promote discussion of this interesting problem and thus inform scaling practice.

Address for correspondence: Alan M. Batterham, Department of Exercise and Sport Science, the Manchester Metropolitan University, Crewe and Alsager Faculty, Alsager, ST7 2HL, ENGLAND. E-mail: A. BATTERHAM@mmu.ac.uk.

Current address for David R. Mullineaux: School of Social Sciences, University of Teesside, Middlesbrough, TS1 3BA, England.

REFERENCES

1. Allied Dunbar National Fitness Survey. *A Report on Activity Patterns and Fitness Levels: Main Findings* (Commissioned by the Sports Council and Health Education Authority). Ipswich, England: Ancient House Press, 1992.
2. American College of Sports Medicine. Policy statement regarding the use of human subjects and informed consent. *Med. Sci. Sports Exerc.* 10:ix-x,1978.
3. Armstrong, N. and J. R. Welsman. Assessment and interpretation of aerobic fitness in children and adolescents. *Exerc. Sports Sci. Rev.* 22:435–476, 1994.
4. Batterham, A. M. and K. Tolfrey. Intraspecific allometry of aerobic performance in prepubesecent children. *Med. Sci. Sports. Exerc.* 28(Suppl.)S103:1996.
5. Cohen, J. A., P. R. Gupta, E. Lichstein, and K. D. Chadda. The heart of a dancer: noninvasive cardiac evaluation of professional ballet dancers. *Am. J. Cardiol.* 45:959–965, 1980.
6. Cumming, D. C., S. R. Wall, M. A. Galbraith, and A. N. Belcastro. Reproductive hormone response to exercise. *Med. Sci. Sports. Exerc.* 19:234–238, 1987.
7. Davies, M. J., G. P. Dalsky, and P. M. Vanderburgh. Allometric scaling of $\dot{V}O_{2max}$. by body mass and lean body mass in older men. *J. Aging Phys. Act.* 3:324–331, 1995.
8. Deschenes, M. R., W. J. Kraemer, C. M. Maresh, and J. F. Crivello. Exercise-induced hormonal changes and their effects upon skeletal muscle tissue. *Sports Med.* 12:80–93, 1991.
9. Devereux, R. B., E. M. Lutas, P. N. Casale, et al. Standardization of M-mode echocardiographic left ventricular anatomic measurements. *J. Am. Coll. Cardiol.* 4:1222–1230, 1984.
10. Devereux, R. B. and N. Reichek. Echocardiographic determination of left ventricular mass in man: anatomic validation of the method. *Circulation* 55:613–618, 1977.
11. Durnin, J. V. G. A., and J. Womersley. Body fat assessed from total body density and its estimation from skinfold thicknesses: measurement on 481 men and women aged 16–72 years. *Br. J. Nutr.* 32:77–97, 1974.
12. George, K. P., L. A. Wolfe, and G. W. Burggraf. The 'athletic heart syndrome:' a critical review. *Sports Med.* 11:300–331, 1991.
13. Henry, W. L., J. M. Gardin, and J. H. Ware. Echocardiographic measurements in normal subjects from infancy to old-age. *Circulation* 62:1054–1060, 1980.
14. Hutchinson, P. L., K. J. Cureton, H. Outz, and G. Wilson. Relationship of cardiac size to maximal oxygen uptake and body size in men and women. *Int. J. Sports Med.* 12:369–373, 1991.
15. Koenig, H., A. Goldstone, and Y. L. Chung. Testosterone mediated sexual dimorphism of the rodent heart. *Circ. Res.* 50:782–787, 1982.
16. Nevill, A. M., R. Ramsbottom, and C. Williams. Scaling physiological measurements for individuals of different body size. *Eur. J. Appl. Physiol.* 65:110–117, 1992.
17. Perrault, H. and R. A. Turcotte. Exercise-induced cardiac hypertrophy: fact or fallacy? *Sports Med.* 17:288–308, 1994.
18. Prothero, J. Heart weight as a function of body weight in mammals. *Growth* 43:139–50, 1979.
19. Reichek, N. and R. B. Devereux. Left ventricular hypertrophy: relationship of anatomic, echocardiographic, and electrocardiographic findings. *Circulation* 63:1391–1398, 1981.
20. Schmidt-Nielsen, K. *Scaling: Why is Animal Size So Important?* Cambridge: Cambridge University Press, 1984, pp. 7–32.
21. Stumpf, W. E., M. Sar, and G. Aumuller. The heart: a target organ for estradiol. *Science* 196:319–321, 1977.
22. Tanner, J. M. Fallacy of per-weight and per-surface area standards and their relation to spurious correlation. *J. Appl. Physiol.* 2:1–15, 1949.
23. Vanderburgh, P. M. and F. I. Katch. Ratio scaling of $\dot{V}O_{2max}$. penalizes women with larger percent body fat, not lean body mass. *Med. Sci. Sports Exerc.* 28:1204–1208, 1996.
24. Westerlind, K. C., W. C. Byrnes, P. S. Freedson, and F. I. Katch. Exercise and serum androgens in women. *Phys. Sports Med.* 15:87–94, 1987.
25. Wolfe, L. A., D. A. Cunningham, and D. R. Boughner. Physical conditioning effects on cardiac dimensions: a review of echocardiographic studies. *Can. J. Appl. Sport Sci.* 11:66–79, 1986.

Appendix 7

Mullineaux, D.R., Barnes, C.A. and Batterham, A.M. (1998a). Allometric scaling of anaerobic performance and the use of 'bootstrapping' for statistical inference from a small sample. In *Proceedings of the 2nd Australasian Biomechanics Conference* (edited by P. Hume, R. Marshall, P. Hunter, S. Stanley, I. Anderson and P. McNair), pp. 72-3. Auckland: The University of Auckland.

ALLOMETRIC SCALING OF ANAEROBIC PERFORMANCE AND THE USE OF 'BOOTSTRAPPING' FOR STATISTICAL INFERENCE FROM A SMALL SAMPLE

D.R., Mullineaux, C.A. Barnes, and A.M. Batterham.
University of Teesside and Manchester Metropolitan University, UK.

INTRODUCTION

In biomechanical research time and cost restrictions often dictate that a small sample is used for data collection. Analytical formulae constrained by assumptions about the sample's distribution, and rationale for a statistical test are typically used to infer the population parameters from the results. With the Pearson rho correlation coefficient, for example, it is assumed the distribution is normal and that the relationship between the dependent and independent variables is linear. Theoretical support and checks for these assumptions should be made to ensure accurate inferences.

Where small samples ($n < 30$) violate the assumption of the normal distribution to use the analytical formulae to generate the sampling distribution, the technique of 'bootstrapping' (Zhu, 1997) can be employed as it operates independently of this assumption. This technique randomly draws, with replacement, a large number of samples from the original sample, and performs statistics on each to produce a sampling distribution. The summary statistics for the distribution provide the estimate for the population.

In assessing the relationship between variables, allometric scaling may provide both statistically and biologically conceptually more viable technique to statistics confined to a linear relationship. This technique provides a zero intercept, multiplicative error and generally a smaller standard error, all necessary or beneficial in relating human performance to body size measures. For example, the peak power output (PPO) produced during a Wingate Anaerobic Test (WAnT) is influenced by the lean body mass (LBM). Although it is still common in the literature that a theoretical relationship is not fully explained, the surface law (Schmidt-Nielsen, 1984) in conjunction with dimensionality theory predicts that power should be proportional to $LBM^{2/3}$.

To demonstrate the use of these 2 techniques, PPO during WAnT will be allometrically scaled by LBM, and 'bootstrapping' used to infer the population parameters from a small sample. It should be noted however, that like all statistical tests, inappropriate uses of these techniques can simply provide a different problem. It is therefore important that the techniques only be considered amongst the myriad of conventional tests existing.

METHODS

Eight physically active, healthy, male subjects volunteered to participate in this study and provided written informed consent. The subjects' descriptive statistics and LBM, determined using an air pressure system based on Boyle's Law (Bod Pod ®, Life Measurement Instruments, USA) and utilising the formulas of Siri (1961), are provided in table 1.

Table 1: Descriptive statistics and anthropometric data of the subjects.

Variable	Range (min - max)	Mean ± SD
Age (years)	17 - 36	27 ± 7
Height (m)	1.65 - 1.84	1.75 ± 0.07
Body mass (kg)	54.9 - 97.8	73.8 ± 14.0
LBM (kg)	43.9 - 80.8	61.6 ± 11.5

Following a cycling and stretching warm-up, subjects performed a 10 s WAnT on a friction braked cycle ergometer (Monark 814E, USA) with a load of $0.1\ kp.kg^{-1}$. The PPO obtained was corrected for the inertia of the flywheel (Smith software, Lakomy). The Z scores for the kurtosis and skewness of the PPO data were 1.9 and -2.2, respectively. As both are not within the boundaries of ± 2, the data is not considered to be normally distributed (Vincent, 1995) thus violating the use of parametric statistics. To allometrically scale the relationship between PPO and LBM via log-log transformations in linear regression analysis, 1000 bootstrap samples were taken and used (Resampling Stats Inc, USA).

RESULTS AND CONCLUSIONS

The results obtained were a PPO of 909 ± 185 W, and an allometrically scaled power = 8.5 • $LBM^{1.13 \pm 0.20}$ (parametric analysis) and 8.3 • $LBM^{1.13 \pm 0.27}$ (bootstrapping analysis). Both regression formulas of the form $y = a \cdot x^{b \pm 1SE}$ are similar, and the b exponent is not significantly different from linearity. Both analyses are however significantly different from $LBM^{2/3}$, hence not supporting predictions from dimensionality theory. With regard to the parametric and bootstrapping analyses, these are very similar. The primary difference is that the standard error of the b exponent is different. Typically it would be expected that bootstrapping would provide a smaller SE, but is larger in this instance. It is hypothesised that owing to the violation of the normal distribution required of the parametric analysis the error component has been underestimated. Although bootstrapping can work with between 8 and 30 subjects (Zhu, 1997), it is proposed that linear regression prefers a ratio 20:1 (Vincent, 1995). This paper demonstrates the use of allometric scaling and bootstrapping, and emphasises that care should be taken in the use of all statistics for accurate inferences.

REFERENCES

1. Schmidt-Nielsen, K. (1984). *Scaling: Why is Animal Size so Important?* Cambridge: Cambridge University Press.
2. Siri, W.E. (1961). Body composition from fluid space and density. In *Techniques for Measuring Body Composition* (edited by J. Brozek and A. Hanschel), pp. 223-224. Washington DC: National Academy of Science.
3. Vincent, W.J. (1995). *Statistics in Kinesiology*. Champaign: Human Kinetics.
4. Zhu, W. (1997). Making bootstrap statistical inferences: a tutorial. *Research Quarterly for Exercise and Sport*, **68**, 44-55.

SPEAKER INFORMATION

David Mullineaux, School of Social Sciences, University of Teesside, Middlesbrough, TS1 3BA, UK. Tel: +44-1642-342355 (direct); Fax: 342303; D.R.Mullineaux@tees.ac.uk

Appendix 8

Payton, C.J., Hay, J.G. and Mullineaux, D.R. (1995). The effect of body roll on hand velocity in the front crawl. In *Proceedings of the XVth Congress of the International Society of Biomechanics* (edited by K. Häkkinen, K.L. Keskinen, P.V. Komi and A. Mero), pp. 720-1. Jyväskylä: University of Jyväskylä.

THE EFFECT OF BODY ROLL ON HAND VELOCITY IN THE FRONT CRAWL

Payton, C.J[1]., Hay, J.G[2], and Mullineaux, D.R[1].

[1] Division of Sport Science, Crewe+Alsager Faculty, The Manchester Metropolitan University, Alsager, UK
[2] Department of Exercise Science, University of Iowa, Iowa City, USA

INTRODUCTION

When front crawl swimmers move their hands vertically and medio-laterally through the water, forwardly directed hydrodynamic lift forces are generated. These lift forces can therefore make a significant contribution to propulsion in the front crawl. Since lift forces are proportional to the square of the hand speed, forwardly directed (propulsive) lift forces will be a function of the vertical and medio-lateral hand velocities produced by the swimmer.

Although it seems logical that elbow flexion and shoulder adduction should contribute to medio-lateral and vertical hand velocity, a recent simulation study [1] demonstrated that medial deviations of the hand to the midline of the trunk could be achieved with 19°-34° of body roll, without the need for elbow flexion or shoulder adduction. As this was well below the mean maximum body roll value of 60.8° exhibited by ten swimmers [2], it was suggested that swimmers move their arms laterally relative to the trunk in order to avoid pulling the hand too far across the midline of the trunk.

As body roll appears to influence medio-lateral hand displacement, it must also contribute to medio-lateral hand velocity and may therefore assist in the generation of propulsive lift forces in front crawl swimming. The objective of this study was to determine the effect of body roll on medio-lateral and vertical hand velocities in front crawl swimming.

METHODS

A previously reported three-dimensional model [1] was modified for this study. The right arm was represented by two rigid segments hinged at the elbow to enable flexion and extension. The arm was linked to a rigid trunk with a shoulder joint capable of extension and abduction/adduction. The trunk was free to rotate about its longitudinal axis - the *body roll axis*.

Three simulations were run, each with a pull time (t_{PULL}) of 0.7 s. The following two body movements were common to all three simulations:

1) The trunk rotated from a neutral position (shoulders horizontal) to a maximum body roll angle of 60° and back to the neutral position.
2) The shoulder extended through 180° from a position of full flexion.

Simulation 1: The elbow remained fully extended and no shoulder abduction occurred.
Simulation 2: The elbow flexed through 90° from an extended position and then fully extended again. The shoulder abducted through 55° from the neutral position and then fully adducted again.
Simulation 3: The elbow flexed through 90° from an extended position and then fully extended again. The shoulder abducted through 90° from the neutral position and then fully adducted again.

Elbow flexion and shoulder abduction angular velocities were modelled as sine functions and body roll angular velocity (ω_{BR}) as a cosine function such that they had a value of zero at the midpoint of the pull ($t_{PULL}/2$). The shoulder extension angular velocity was held constant (π / t_{PULL} rad/s) throughout the pull.

RESULTS

The mean medio-lateral ($\bar{v}_X$) and vertical ($\bar{v}_Y$) hand velocity components from $t = 0$ to $t_{PULL}/2$ were determined for each simulation. The contribution to these components made by body roll, shoulder extension and 'arm movement' (shoulder abduction plus elbow flexion) were then computed. These data are presented in Table 1.

Table 1. The contribution of body roll, shoulder extension and arm movement to the mean medio-lateral ($\bar{v}_X$) and vertical ($\bar{v}_Y$) hand velocities.

	Simulation 1		*Simulation 2*		*Simulation 3*	
	$\bar{v}_X$ (m/s)	$\bar{v}_Y$ (m/s)	$\bar{v}_X$ (m/s)	$\bar{v}_Y$ (m/s)	$\bar{v}_X$ (m/s)	$\bar{v}_Y$ (m/s)
Body Roll.	1.28	0.16	1.21	0.07	1.47	- 0.31
Shoulder Ext.	1.14	- 2.01	1.04	- 1.92	0.98	- 1.87
Arm Movement.	-	-	- 0.39	0.33	- 1.38	- 0.03
Total.	2.42	-1.85	1.86	- 1.52	1.07	- 2.21

DISCUSSION

Body roll provided the majority of medio-lateral hand velocity in all three simulations, with its estimated contribution ranging from 53% (*simulation 1*) to 137% (*simulation 3*). Body roll was less influential on vertical hand velocity. In *simulation 2* it reduced the total vertical hand velocity by 5%, whereas in *simulation 3* it increased the total vertical hand velocity by 16%.

As the body roll angular velocity-time histories were identical for each simulation, the differences in the hand velocities, produced by body roll, can be attributed to differences in the length and direction of the radius from the hand to the body roll axis in each simulation.

Arm movement reduced the total medio-lateral hand velocity generated because it produced a hand velocity which had a lateral component, when body roll and shoulder extension were producing medial velocity components (and vice versa). The mean lateral hand velocity in *simulation 3* was much greater that in *simulation 2*. This difference can be explained by the greater range of lateral hand movement produced in *simulation 3* because of the increased range of shoulder abduction.

Shoulder extension created between 47% (*simulation 1*) and 91% (*simulation 3*) of the medio-lateral hand velocity. Without body roll, shoulder extension would make no contribution to the medio-lateral hand velocity. Body roll therefore increases the medio-lateral hand velocity indirectly by changing the plane in which shoulder extension occurs. The inverse is true for vertical hand velocity, body roll reduces the contribution of shoulder extension to this velocity component.

The results suggest that body roll makes a substantial contribution to medio-lateral hand velocity in the front crawl and may therefore play an important role in the generation of propulsive lift forces.

REFERENCES

1. Hay, J.G., Liu, Q., and Andrews, J.G. (1993) Body roll and handpath in freestyle swimming: a computer simulation study. *J. Appl. Biomech.* 9 (3), 227-237.
2. Liu, Q., Hay, J.G., and Andrews, J.G. (1993) Body roll and handpath in freestyle swimming: an experimental study. *J. Appl. Biomech.* 9 (3), 238-253.

Appendix 9

Payton, C.J., Hay, J.G. and Mullineaux, D.R. (1997). The effect of body roll on hand speed and hand path in the front crawl - a simulation study. *Journal of Applied Biomechanics*, **13**, 300-15.

JOURNAL OF APPLIED BIOMECHANICS, 1997, 13, 300-315

The Effect of Body Roll on Hand Speed and Hand Path in Front Crawl Swimming—A Simulation Study

Carl J. Payton, James G. Hay, and David R. Mullineaux

The aim of this study was to predict the effect of body roll on hand speed and hand path during the pull phase in front crawl swimming. An earlier three-segment model (Hay, Liu, & Andrews, 1993) was developed to enable the hand to move out of the plane through the shoulder parallel to the sagittal plane of the rotating trunk. Elbow flexion, shoulder abduction, and body roll angular velocities were modeled as sine or cosine functions. For a given elbow flexion, an increase in maximum body roll from 45° to 60° produced a marked increase in medial hand motion. For a given body roll, an increase in maximum elbow flexion from 60° to 90° increased medial hand motion and reduced downward hand motion. An increase in body roll increased hand speed in the plane perpendicular to the swimming direction, thus increasing the potential of the hand to develop propulsive lift forces.

For many years, front crawl swimmers were instructed by coaches to pull their hands directly backward in a straight line beneath their bodies. It was believed that propulsion was gained entirely from the forward-acting drag forces that opposed the hand motion. Since the work of Brown and Counsilman (1971), it has been generally accepted that the hands of elite front crawl swimmers follow underwater trajectories that deviate considerably from a straight line. It has been suggested that in following such trajectories the hand behaves as a foil, creating lift forces that act perpendicular to the direction of movement of the hand (Brown & Counsilman, 1971; Schleihauf, 1979).

When front crawl swimmers move their hands vertically and mediolaterally in the water, perpendicular to the swimming direction, the lift force acts in a forward direction. It has therefore been suggested that lift forces contribute significantly to propulsion in the front crawl (Berger, Hollander, & De Groot, 1995; Schleihauf, Gray, & DeRose, 1983). Since lift forces are a function of the square of the hand speed relative to the water, forward-directed (propulsive) lift forces are a function of the vertical and mediolateral hand speeds generated by a front crawl swimmer.

Although the importance of mediolateral and vertical hand speed in front crawl swimming is reasonably well understood (Counsilman, 1981), little attempt has been made to identify the body movements responsible for generating these hand speeds. Barthels (1979) suggested that during the insweep phase of the front crawl, vertical and

---

Carl J. Payton is with the Department of Exercise and Sport Science, Crewe+Alsager Faculty, The Manchester Metropolitan University, Stoke-on-Trent, ST7 2HL, UK. James G. Hay is with the Department of Exercise Science, University of Iowa, Iowa City, IA 52242. David R. Mullineaux is with the School of Social Sciences, University of Teesside, Middlesbrough, Cleveland, UK.

medial hand speeds (and the resulting lift forces) are achieved by flexion of the elbow. It would also seem logical that abduction and adduction at the shoulder could contribute to the generation of mediolateral and, to a lesser extent, vertical hand speed in the front crawl.

In a recent study, Hay, Liu, and Andrews (1993) questioned the role played by the elbow and shoulder in creating mediolateral hand movement in the front crawl. Using a mathematical model, they demonstrated that movements of the hand away from/toward the midline of the trunk could be achieved entirely by body roll, without the need for elbow flexion or shoulder adduction. They also reported that the amount of body roll required to sweep the hand to the midline was only 19–34°. Because this was well below the average maximum body roll value of 60.8° exhibited by 10 male university swimmers (Liu, Hay, & Andrews, 1993), it was concluded that swimmers must move their arms laterally relative to the trunk in order to avoid pulling the hand too far across the midline of the trunk.

In the model proposed by Hay et al. (1993), the hand was constrained to move in the plane through the right shoulder and normal to the shoulder axis—the *parasagittal plane* (Figures 1 and 2). In their experimental study, Liu et al. (1993) demonstrated that this was an unrealistic constraint, because swimmers invariably moved their hands laterally relative to this plane.

Because body roll influences the mediolateral displacement of the hand relative to the water, it must also contribute to mediolateral hand speed. It thus follows that body roll may help generate propulsive lift forces in the front crawl. Prichard (1993) supported the view that front crawl swimmers generate high hand speeds and large propulsive forces by body roll. He suggested that range, speed, strength, and timing of hip rotation are important determinants of front crawl performance.

The aim of this study was to predict the effect of body roll on hand speed and hand path during the pull phase in front crawl swimming.

Methods

The Model

The model used in this study was developed from the three-dimensional mathematical model proposed by Hay et al. (1993). The model in the present study allows the hand to move out of the parasagittal plane (Figures 1 and 2).

The right arm was modeled as two rigid segments hinged at the elbow to enable flexion and extension. The arm was linked to a rigid trunk with a shoulder joint capable of extension and abduction/adduction. The trunk was free to rotate about its longitudinal axis —the *body roll axis*.

The velocity of the distal end of the right hand was determined relative to the water by defining a right-handed inertial reference frame fixed in the pool. The x-axis of the frame was directed down the pool in the swimming direction, the y-axis was directed vertically upward, and the z-axis was directed horizontally to the right wall of the pool.

The following terms were used to describe the direction of hand motion:

- Right-hand movement along the x-axis: forward (+*ve*) or backward (–*ve*)
- Right-hand movement along the y-axis: upward (+*ve*) or downward (–*ve*)
- Right-hand movement along the z-axis: lateral (+*ve*) or medial (–*ve*).

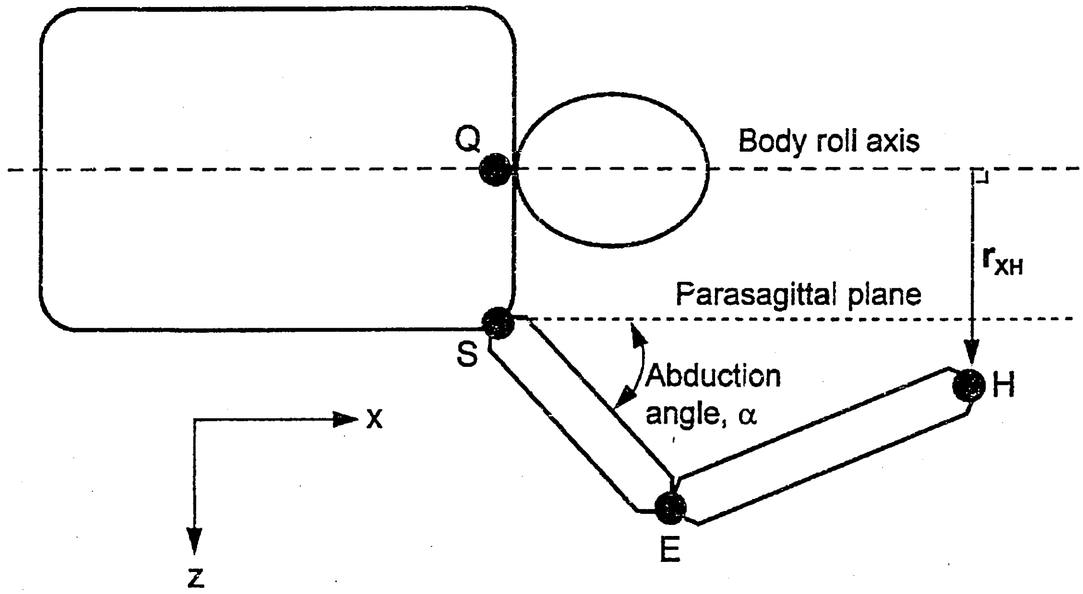


Figure 1 — Body roll model viewed from above.

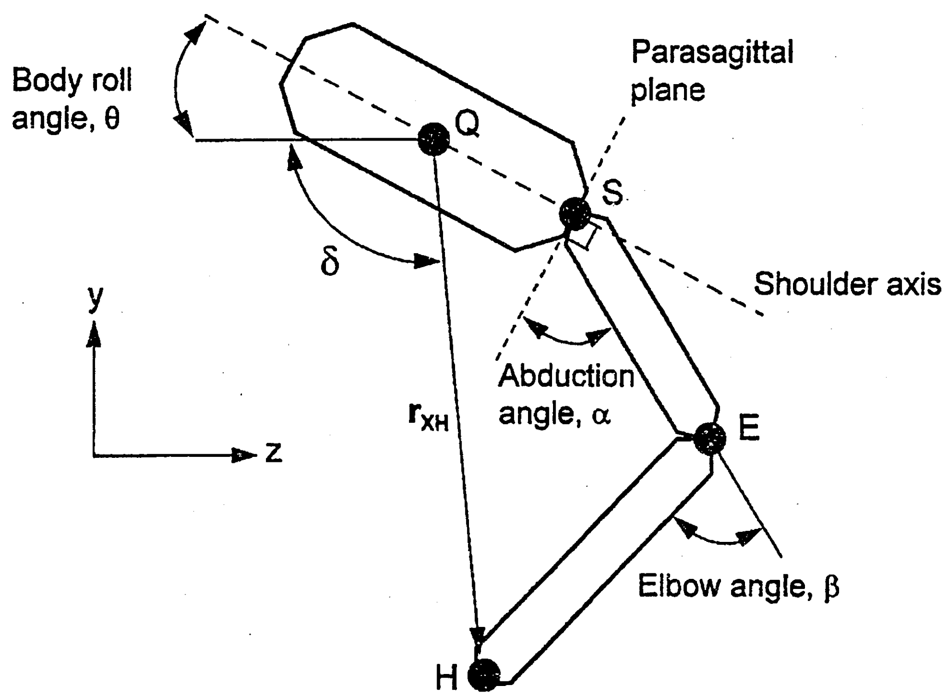


Figure 2 — Body roll model viewed from behind.

The velocity of the hand relative to the pool-fixed reference frame was modeled as a function of the following variables:

- *Body roll angle* (θ): the angle between the line connecting the two shoulder joints (the shoulder axis) and the horizontal (x–z) plane (Figure 2).
- *Body roll angular velocity* (ω_{BR}): the rate of change of θ with respect to time.
- *Shoulder extension angle* (ϕ): the angle between the shoulder-to-elbow position vector ($\mathbf{r}_{SE}$) and the x-axis, when projected onto the sagittal plane (Figure 3).

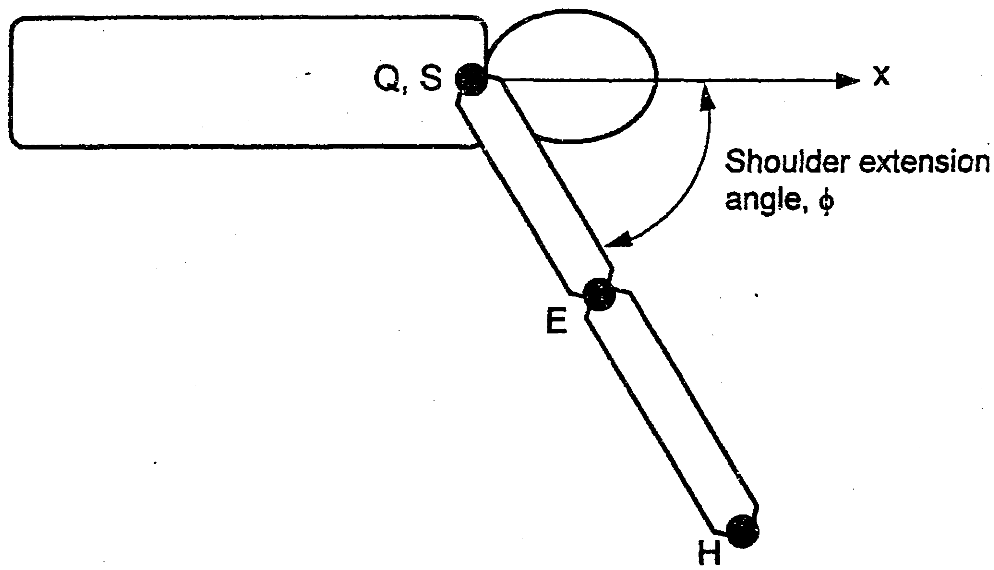


Figure 3 — Body roll model viewed in the sagittal plane.

- *Shoulder extension angular velocity* (ω_{SE}): the rate of change of ϕ with respect to time.
- *Shoulder abduction angle* (α): the angle between the shoulder-to-elbow position vector ($\mathbf{r}_{SE}$) and the parasagittal plane (Figures 1 and 2).
- *Shoulder abduction angular velocity* (ω_{SA}): the rate of change of α with respect to time.
- *Elbow flexion angle* (β): the angle between the upper arm and lower arm segments subtracted from 180° (Figure 2).
- *Elbow flexion angular velocity* (ω_{EF}): the rate of change of β with respect to time.
- *Pull time* (t_{PULL}): the time taken for the arm to rotate from a 0° to 180° shoulder extension angle.
- *Trunk velocity* ($v_{Q/P}$): the velocity of the trunk-fixed reference point Q along the x-axis.
- *Half shoulder-width* (l_{QS}): the distance from the midpoint of the shoulders Q to the right shoulder S.
- *Upper arm length* (l_{SE}): the distance from the right shoulder (S) to the right elbow (E).
- *Lower arm length* (l_{EH}): the distance from the right elbow (E) to the distal end of the right hand (H).

Model dimensions l_{QS}, l_{SE}, and l_{EH} were held constant for all simulations. Angles ϕ, θ, α, and β were varied as a function of the pull time (t_{PULL}) within a simulation trial as described in the following section.

Variables

Body Roll Variables (θ, ω_{BR}). The trunk was rotated from the neutral position ($\theta = 0°$) to a preselected angle of maximum body roll θ_{MAX} and back to the neutral position in the pull time t_{PULL}. The time to reach θ_{MAX} was equated to half the pull time. Body roll angular velocity ω_{BR} was modeled as a cosine function such that ω_{BR} was maximum as the trunk rotated through the neutral position and zero at the instant of maximum body roll.

Shoulder Extension Variables (ϕ, ω_{SE}). From a position of full flexion ($\phi = 0°$), the shoulder was extended at a constant rate through an angle of 180° in the time t_{PULL}. The shoulder extension angular velocity ω_{SE} was therefore a function of the pull time ($\omega_{SE} = 180°/ t_{PULL}$).

Elbow Flexion and Shoulder Abduction Variables (β, ω_{EF}, α, ω_{SA}). The straight–bent–straight elbow pull pattern characteristic of front crawl swimming was simulated in the model by allowing the elbow to flex from a fully extended position ($\beta = 0°$) to a specified maximum flexion angle β_{MAX} and then back to full extension in the time t_{PULL}. The shoulder joint was abducted from a neutral position ($\alpha = 0°$) to a specified maximum abduction angle α_{MAX} and then back to the neutral position in the time t_{PULL}.

The elbow flexion angular velocity (ω_{EF}) and shoulder abduction angular velocity (ω_{SA}) were both modeled as sine functions such that they were zero at the start and finish of the pull and at the instant of maximum body roll. The time to reach α_{MAX} and β_{MAX} therefore coincided with the time to reach maximum body roll ($t_{PULL}/2$).

Computation of Hand Velocity

The velocity of the hand relative to the pool-fixed reference frame ($\mathbf{v}_{H/P}$) is equal to the sum of the velocity of the hand relative to the trunk-fixed reference point Q ($\mathbf{v}_{H/Q}$) and the velocity of Q relative to the pool-fixed reference frame ($\mathbf{v}_{Q/P}$) :

$$\mathbf{V}_{H/P} = \mathbf{V}_{H/Q} + \mathbf{V}_{Q/P} \tag{1}$$

The velocity of the hand relative to the trunk-fixed point Q has a component due to shoulder extension angular velocity ($\mathbf{v}_{SE}$), body roll angular velocity ($\mathbf{v}_{BR}$), elbow flexion angular velocity ($\mathbf{v}_{EF}$), and shoulder abduction angular velocity ($\mathbf{v}_{SA}$) and was calculated using Equation 2:

$$\mathbf{V}_{H/Q} = \mathbf{V}_{SE} + \mathbf{V}_{BR} + \mathbf{V}_{EF} + \mathbf{V}_{SA} \tag{2}$$

The contribution to hand velocity made by the shoulder extension angular velocity ($\mathbf{v}_{SE}$) is given by Equation 3:

$$\mathbf{V}_{SE} = \boldsymbol{\omega}_{SE} \times \mathbf{r}_{SH} \tag{3}$$

where $\mathbf{r}_{SH}$ is the position vector from the right shoulder to the distal end of the right hand, and the symbol $\times$ denotes the vector cross product.

The forward–backward ($\mathbf{v}_{SE(X)}$), upward–downward ($\mathbf{v}_{SE(Y)}$), and mediolateral ($\mathbf{v}_{SE(Z)}$) components of vector $\mathbf{v}_{SE}$ are given by Equations 4, 5, and 6, respectively:

$$\mathbf{V}_{SE(X)} = |\boldsymbol{\omega}_{SE} \times \mathbf{r}_{SH}| \sin\phi\, \mathbf{i} \tag{4}$$

$$\mathbf{V}_{SE(Y)} = |\boldsymbol{\omega}_{SE} \times \mathbf{r}_{SH}| \cos\phi \cdot \cos\theta\, \mathbf{j} \tag{5}$$

$$\mathbf{V}_{SE(Z)} = |\boldsymbol{\omega}_{SE} \times \mathbf{r}_{SH}| \cos\phi \cdot \sin\theta\, \mathbf{k} \tag{6}$$

where **i**, **j**, and **k** are the unit vectors along the *x*-, *y*-, and *z*-axes, respectively.

The contribution to hand velocity made by the body roll angular velocity ($\mathbf{v}_{BR}$) is given by Equation 7:

$$\mathbf{V}_{BR} = \boldsymbol{\omega}_{BR} \times \mathbf{r}_{XH} \tag{7}$$

where $\mathbf{r}_{XH}$ is the perpendicular position vector from the body roll axis to the distal end of the right hand, as illustrated in Figures 1 and 2.

The forward–backward ($\mathbf{v}_{BR(X)}$), upward–downward ($\mathbf{v}_{BR(Y)}$), and mediolateral ($\mathbf{v}_{BR(Z)}$) components of vector $\mathbf{v}_{BR}$ are given by Equations 8, 9, and 10, respectively:

$$\mathbf{V}_{BR(X)} = 0 \tag{8}$$

$$\mathbf{V}_{BR(Y)} = |\omega_{BR} \times \mathbf{r}_{XH}| \cos \delta \, \mathbf{j} \tag{9}$$

$$\mathbf{V}_{BR(Z)} = |\omega_{BR} \times \mathbf{r}_{XH}| \sin \delta \, \mathbf{k} \tag{10}$$

where δ is the angle made by the vector $\mathbf{r}_{XH}$ to the horizontal (negative z-axis) as shown in Figure 2.

The contributions to hand velocity made by the elbow flexion angular velocity ($\mathbf{v}_{EF}$) and shoulder abduction angular velocity ($\mathbf{v}_{SA}$) are given by Equations 11 and 12, respectively:

$$\mathbf{V}_{EF} = \omega_{EF} \times \mathbf{r}_{EH} \tag{11}$$

$$\mathbf{V}_{SA} = \omega_{SA} \times \mathbf{r}_{SH} \tag{12}$$

where $\mathbf{r}_{EH}$ is the position vector from the right elbow to the distal end of the right hand. The three Cartesian components of vectors $\mathbf{v}_{EF}$ and $\mathbf{v}_{SA}$ were obtained directly from the vector cross products described by Equations 11 and 12.

Input Parameters for Simulations

Anthropometric Data (l_{QS}, l_{SE}, l_{EH}). The half shoulder-width (l_{QS}), upper arm length (l_{SE}), and forearm plus hand length (l_{EH}) were assigned values of 0.25 m, 0.35 m, and 0.5 m, respectively. These were the mean values previously reported for 10 male competitive swimmers (Liu et al., 1993).

Elbow and Shoulder Abduction Angles (β_{MAX}, α_{MAX}). Two arm pull conditions were simulated. In each simulation, the elbow was allowed to flex to a maximum angle (β_{MAX}) of either 60° (low flex) or 90° (high flex). The shoulder was permitted to abduct to a maximum angle (α_{MAX}) of 90° in each simulation. Shoulder abduction angles have not previously been reported in the literature, so this figure was based on photographic evidence (Maglischo, 1982 , Figure 2.6 I, page 58).

Body Roll Angle, Pull Time, and Trunk Velocity (θ_{MAX}, t_{PULL}, $v_{Q/P}$). Simulations were performed for maximum body roll angles of 45° (low roll) and 60° (high roll). These angles are similar to the mean values previously reported for front crawl swimmers by Beekman (1986) and Liu et al. (1993), respectively.

A pull time of either 0.75 s (fast) or 1.10 s (slow) was used in each simulation. These times were based on mean values reported for sprint swimming (Ringer & Adrian, 1969) and distance swimming (Maglischo et al., 1988). A constant trunk velocity of 1.6 m/s was maintained in each simulation. Although trunk velocity has been shown to influence forward–backward hand motion (Hay et al., 1993), it was not manipulated in the present study since it does not influence the effect of body roll on vertical or mediolateral hand motion.

A total of four simulations were performed in both the fast and slow pull times. Figures 4 to 7 illustrate the four simulations as viewed from behind, up to the point of maximum body roll.

Equations 3–12 were used to compute the contributions made by body roll, shoulder extension, elbow flexion, and shoulder abduction to hand velocity. As each simulation was symmetrical about its midpoint ($t_{PULL}/2$), the velocity contributions were only calculated for the first half ($t = 0$ to $t_{PULL}/2$) of each simulation (to avoid the positive and nega-

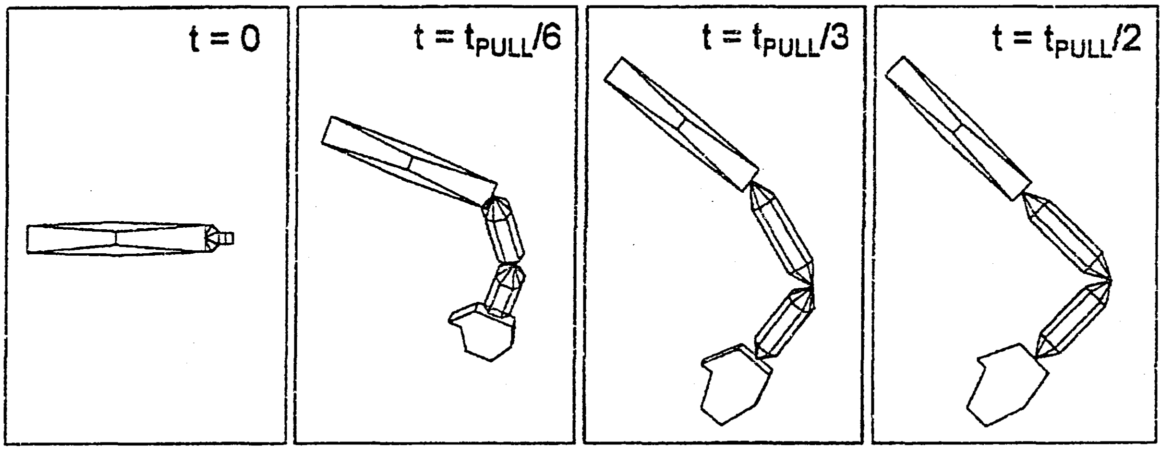


Figure 4 — Low-roll/high-flex simulation involving 45° of body roll, 90° of elbow flexion, and 90°of shoulder abduction, viewed from behind.

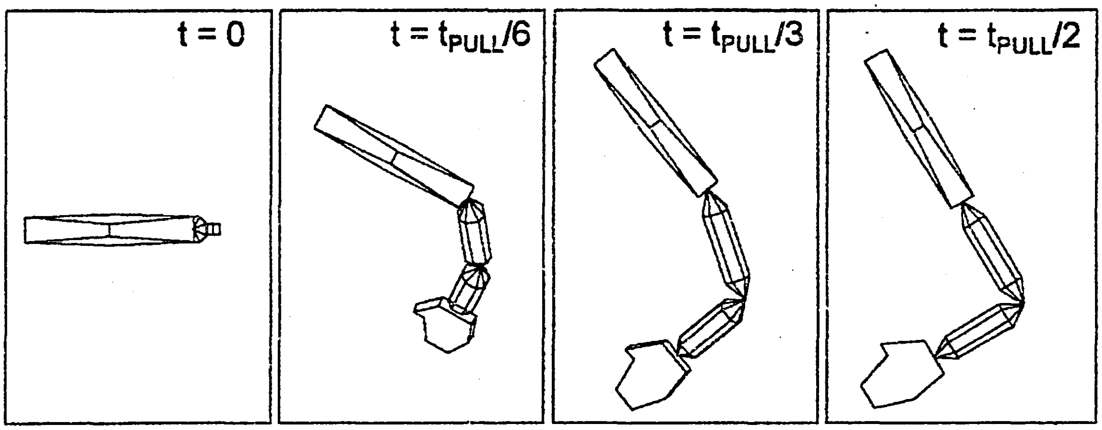


Figure 5 — High-roll/high-flex simulation involving 60° of body roll, 90° of elbow flexion, and 90° of shoulder abduction, viewed from behind.

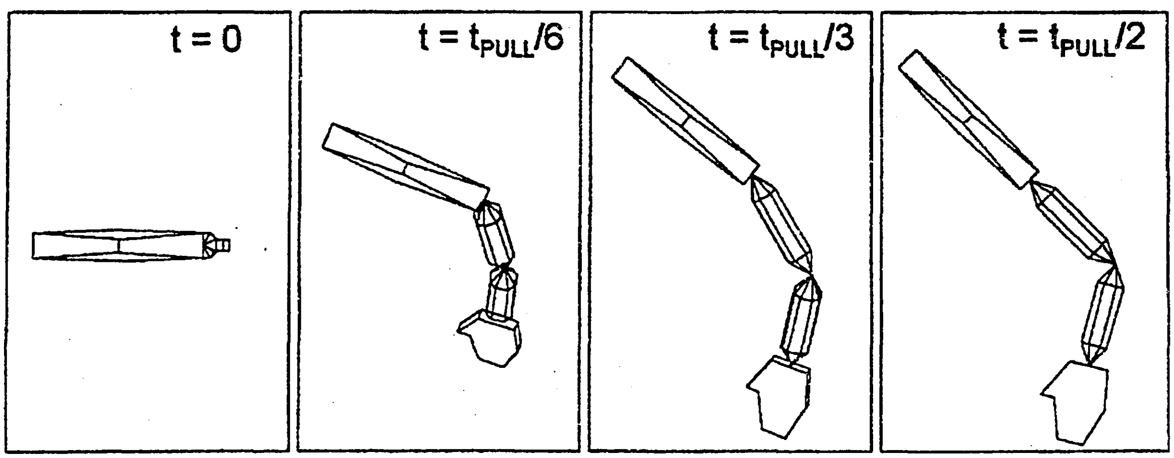


Figure 6 — Low-roll/low-flex simulation involving 45° of body roll, 60° of elbow flexion, and 90° of shoulder abduction, viewed from behind.

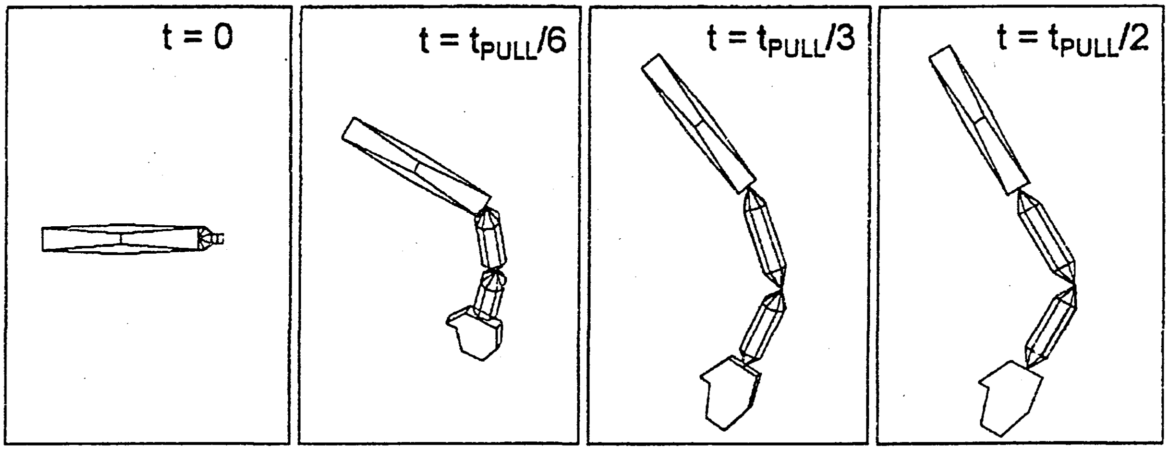


Figure 7 — High-roll/low-flex simulation involving 60° of body roll, 60° of elbow flexion, and 90° of shoulder abduction, viewed from behind.

tive velocities canceling). Therefore, the results presented in the following section apply to the simulations up to the point of maximum body roll.

Results and Discussion

The effects of body roll on the three Cartesian components of the hand velocity are described by Equations 8, 9, and 10. Equation 8 indicates that body roll makes no contribution to backward hand velocity, which suggests that body roll cannot be used to produce forward-acting (propulsive) drag forces on the hand. Shoulder extension, elbow flexion, and shoulder abduction may all help create propulsive drag forces, because they all have the potential to generate backward hand velocity, as evidenced by Equations 4, 11, and 12, respectively.

The contribution made by body roll to hand velocity is confined to the plane perpendicular to the swimming direction (the y–z plane), as indicated by Equations 9 and 10. Therefore, body roll has the potential to assist in the creation of forward-acting (propulsive) lift forces by the hand.

Hand Speed and Hand Speed Squared

Propulsive lift forces are created by moving or *sculling* the hand in a plane perpendicular to the swimming direction. The speed at which the hand moves in this plane (henceforth referred to as the *hand sculling speed*) influences the magnitude of the propulsive lift forces produced. Specifically, the propulsive lift force acting on the hand is proportional to the square of the hand sculling speed.

Simulations for the fast pull time produced hand sculling speeds that were 47% higher than the equivalent simulations run for the slow pull time (Figure 8). Although there was little variation in hand sculling speed among the four simulations, the highest hand sculling speeds were found for the high-roll/low-flex simulation, and the lowest hand sculling speeds were found for the low-roll/high-flex simulation.

The extent to which body roll contributes to hand sculling speed depends on the body roll angular velocity (ω_{BR}) and the displacement of the hand from the body roll axis ($\mathbf{r}_{XH}$) as indicated by Equation 7. Figure 8 illustrates that for the fast pull time, body roll produced a mean hand sculling speed ranging from 1.15 m/s (low-roll/high-flex) to 1.70

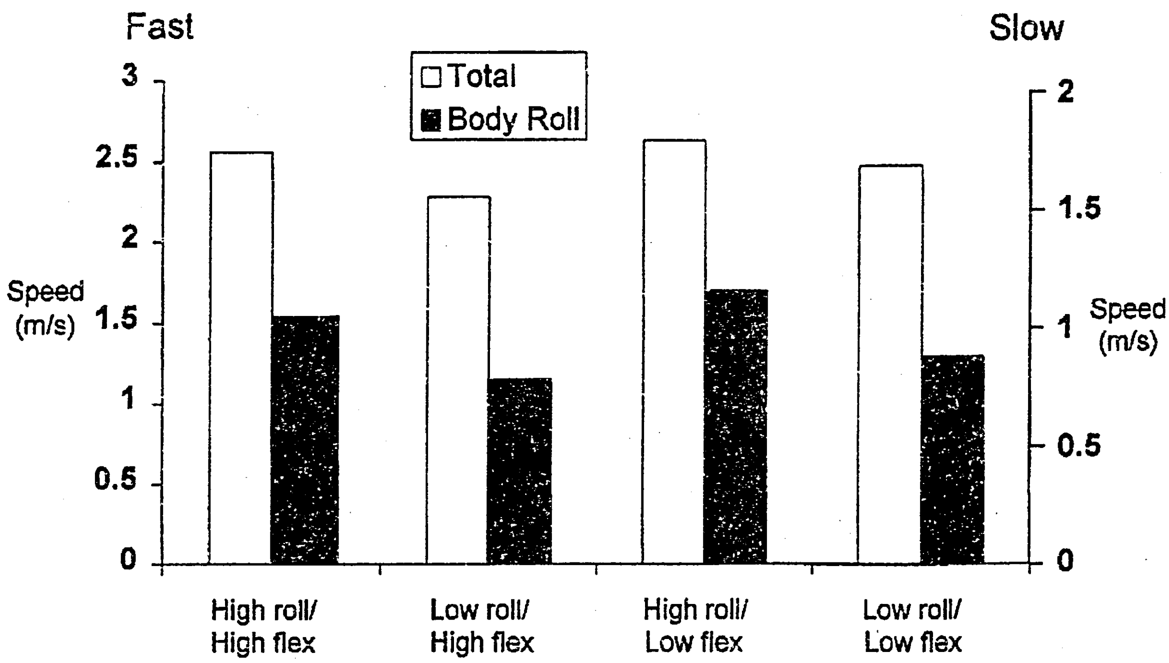


Figure 8 — The mean hand sculling speeds produced by the four simulations (Total) and the mean hand sculling speeds produced by body roll alone (Body Roll). Hand sculling speeds for the fast pull time (t_{PULL} = 0.75 s) and slow pull time (t_{PULL} = 1.10 s) are shown on the left and right vertical axes, respectively.

m/s (high-roll/low-flex). This variability was due to differences in ω_{BR} and/or $\mathbf{r}_{XH}$ between simulations.

The two simulations in which the body rolled to 60° in $t_{PULL}/2$ produced greater body roll angular velocities than did those employing 45° of body roll in the same time. The displacement of the hand from the body roll axis ($\mathbf{r}_{XH}$) varied as a function of the pull time for the high and low elbow flexion conditions (Figure 9), with the two simulations involving 60° of elbow flexion yielding higher values of $\mathbf{r}_{XH}$ than did those using 90° of flexion. Body roll thus made its greatest contribution to hand sculling speed in the high-roll/low-flex simulation, because this involved the highest values of $\mathbf{r}_{XH}$ and ω_{BR}. Conversely, body roll contributed least to hand sculling speed in the low-roll/high-flex simulation, where $\mathbf{r}_{XH}$ and ω_{BR} had their lowest values.

An increase in the hand sculling speed produced by body roll was not matched by an equivalent increase in the total hand sculling speed (Figure 8). This was because the hand velocity vector produced by body roll ($\mathbf{v}_{BR}$) did not necessarily coincide with the total velocity vector of the hand ($\mathbf{v}_{H/P}$), even when viewed in the plane perpendicular to the swimming direction (*y–z* plane). Only a component of $\mathbf{v}_{BR}$ contributed to hand sculling speed at any instant during the pull, and the magnitude of this component varied between simulations. Additionally, although decreasing the maximum elbow angle from 90° to 60° increased the hand sculling speed produced by body roll (for a given maximum roll angle), it also reduced the hand sculling speed produced by elbow flexion. Figure 10 shows that simulations for the fast pull time (left vertical axis) produced mean squared hand sculling speeds that were 115% higher than for the equivalent simulations for the slow pull time (right vertical axis).

The squared hand speed data followed the same trend as the hand speed data (the results are in the same rank order), but the differences between simulations were more pronounced. For example, although the high-roll/low-flex simulation produced a mean

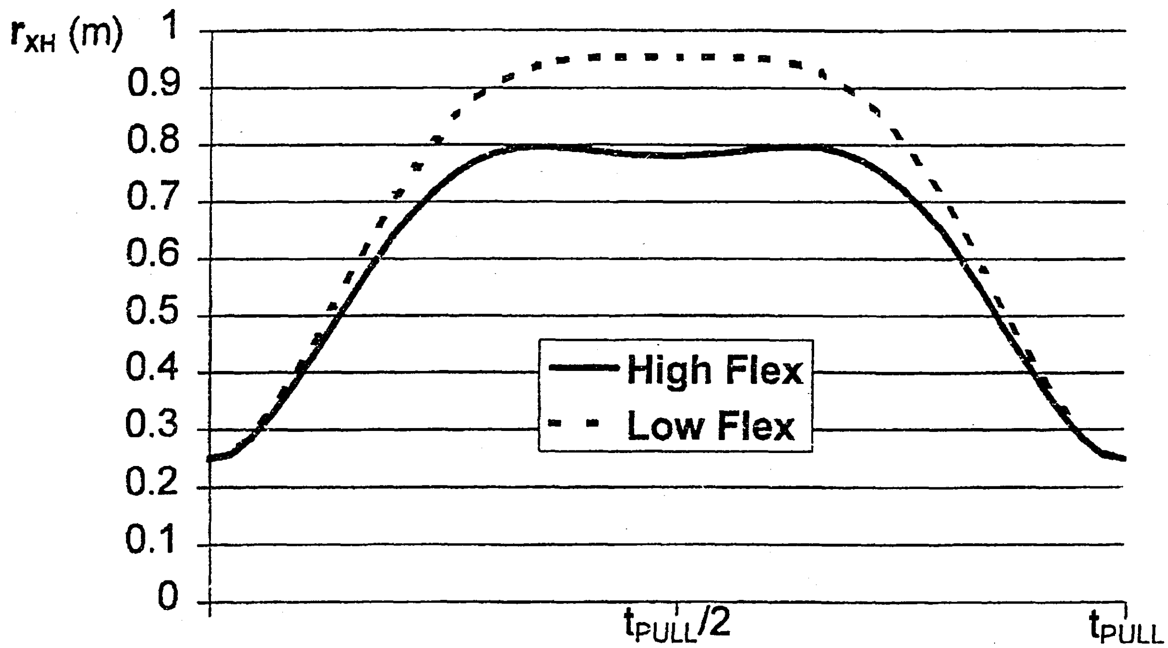


Figure 9 — Displacement of the distal end of the hand from the body roll axis (r_{XH}) as a function of the pull time. The broken line shows the displacements for the high-roll/low-flex and low-roll/low-flex simulations. The solid line shows the displacements for the high-roll/high-flex and low-roll/high-flex simulations.

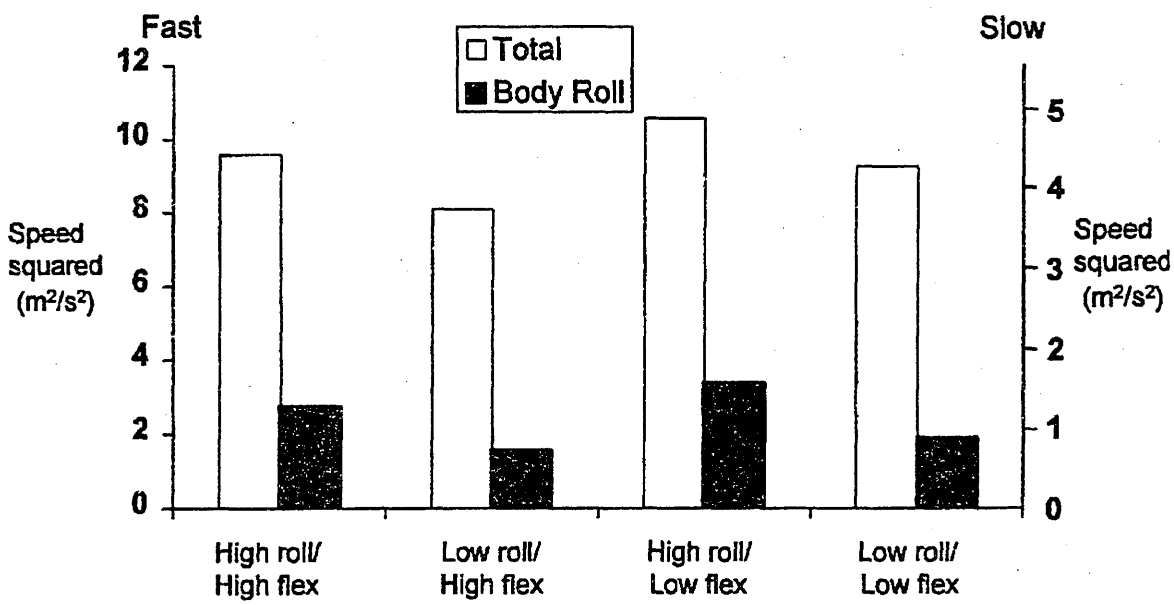


Figure 10 — The mean squared hand sculling speeds produced by the four simulations (Total) and the mean squared hand sculling speeds produced by body roll alone (Body Roll). Squared hand sculling speeds for the fast pull time (t_{PULL} = 0.75 s) and slow pull time (t_{PULL} = 1.10 s) are shown on the left and right vertical axes, respectively.

hand sculling speed that was only 15% higher than for the low-roll/high-flex simulation, it yielded a mean squared hand sculling speed that was 31% higher.

Body roll produced the greatest squared hand sculling speed for the high-roll/low-flex simulation and the least for the low-roll/high-flex simulation. This can again be explained by differences in body roll angular velocity and hand-to-roll axis distance discussed earlier.

The mean squared hand sculling speed for the high-roll/high-flex simulation was 18% higher than for the low-roll/high-flex simulation. Similarly, the high-roll/low-flex simulation created 14% more squared hand sculling speed than the low-roll/low-flex simulation. Because propulsive lift forces are proportional to the square of the hand sculling speed, it could be inferred from these results that, for a given pull time, a front crawl technique that involves 60° of body roll has the potential to generate more propulsive lift than one involving 45° of body roll. Finally, it is worth noting that, for a given maximum body roll, simulations involving 60° of elbow flexion produced higher squared hand sculling speeds than those using 90° of elbow flexion.

While the results in Figures 8 and 10 indicate which techniques produce the highest hand sculling speeds, they give no indication of the path followed by the hand during the pull. The handpath is an important feature of front crawl swimming and has previously been shown to be influenced by body roll (Hay et al., 1993). The present model provided an opportunity to revisit some issues concerning the effect of body roll on the geometry of the handpath.

Mediolateral Hand Motion

Figure 11, a–d, illustrates the top view of the handpath of the right hand relative to the water, for the four simulations conducted for the fast pull time. In the high-roll/high-flex simulation (Figure 11a), the right hand was medially displaced 38 cm at the point of maximum body roll. This took the hand beyond the midline of the trunk (the vertical plane containing the body roll axis). When the body roll angle was reduced by 15° and the same degree of elbow flexion was used, the medial deviation of the hand was reduced to 18 cm, as shown by the low-roll/high-flex simulation (Figure 11b). The high-roll/low-flex simulation (Figure 11c) medially displaced the hand 20 cm by maximum body roll. In contrast, the low-roll/low-flex simulation (Figure 11d) resulted in a slight lateral displacement of the hand by maximum body roll.

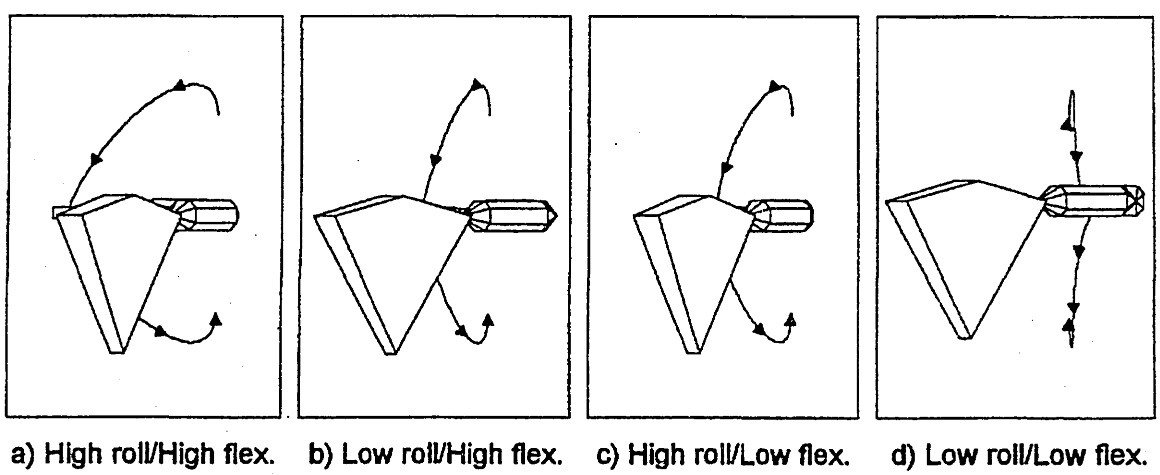


Figure 11 — Overhead view of the path followed by the right hand (relative to the water) in the four simulations. Handpaths shown are for the fast pull time (t_{PULL} = 0.75 s) and a trunk velocity of 1.6 m/s. The position of the trunk and upper extremity at the instant of maximum body roll is also shown.

Two conclusions can be drawn from these observations: An increase in body roll increased medial hand movement, and an increase in the maximum elbow flexion angle also increased medial hand movement. Although the first conclusion is in agreement with the results of the previous simulation study by Hay et al. (1993), the second conclusion is not. Hay et al. (1993) found that the medial deviation of the hand was inversely proportional to the maximum elbow flexion angle. This contradictory finding was due to a difference in the two models. In the Hay et al. (1993) model, the hand was constrained to move in the parasagittal plane through the right shoulder normal to the shoulder axis (Figures 1 and 2), and an increase in elbow flexion simply reduced the displacement of the hand from the body roll axis ($\mathbf{r}_{XH}$). This inevitably reduced the medial deviation of the hand for a given body roll. In the present model, the hand was free to move out of this parasagittal plane (Figure 1), and an increase in the maximum elbow flexion angle from 60 to 90° had two effects. First, it brought the hand closer to the body roll axis (as in the earlier model), which is illustrated in Figure 9. Second, when the elbow was allowed to flex to 90°, the hand remained closer to the parasagittal plane than when the elbow only flexed 60°. That is, 90° of elbow flexion more effectively counteracted the tendency of shoulder abduction to move the hand laterally away from the parasagittal plane than did 60° of elbow flexion. The closer the hand was to the parasagittal plane at maximum body roll, the greater was its medial displacement (for a given $\mathbf{r}_{XH}$ and maximum roll angle).

The displacement values of Table 1 were obtained by integrating the respective velocity–time functions. For example, the contribution of body roll to medial hand displacement was defined as the definite integral of Equation 10 between $t = 0$ and $t_{PULL}/2$.

Although an increase in the maximum elbow flexion angle from 60 to 90° reduced the contribution that body roll made to medial hand movement, medial hand movement was still enhanced. This was because elbow flexion and shoulder extension made increased contributions to medial hand motion. Also, the tendency of shoulder abduction to cause lateral hand motion was reduced when the maximum elbow flexion angle was increased from 60 to 90° (Table 1).

In the high-roll/high-flex simulation, the hand was medially displaced 38 cm by the point of maximum body roll, compared to only 20 cm in the high-roll/low-flex simula-

Table 1 Mediolateral Hand Displacement Up to $t_{PULL}/2$ and the Contribution to This Displacement From Body Roll, Shoulder Extension, Elbow Flexion, and Shoulder Abduction (cm)

	High roll/ high flex	Low roll/ high flex	High roll/ low flex	Low roll/ low flex
Mediolateral displacement	–38	–18	–20	+5
Contribution				
Body roll	–52	–38	–58	–41
Shoulder extension	–35	–27	–34	–26
Elbow flexion	–58	–67	–45	–49
Shoulder abduction	+107	+114	+117	+121

tion. Although the body roll contribution to medial hand displacement was 6 cm less in the high-roll/high-flex simulation than in the high-roll/low-flex simulation, this was more than compensated for by increased contributions from elbow flexion (13 cm more) and shoulder extension (1 cm more). Also, the tendency of shoulder abduction to cause lateral hand motion was 10 cm less in the high-roll/high-flex simulation than in the high-roll/low-flex simulation. The same pattern is seen in the data when the low-roll/high-flex simulation is compared to the low-roll/low-flex simulation.

The extent to which shoulder extension contributes to mediolateral hand motion is determined by the amount of body roll. Equation 6 shows that shoulder extension is only able to contribute to mediolateral hand motion once body roll has commenced ($\theta \neq 0$), and that for a given shoulder extension angle ϕ, the greater the body roll, the greater is the contribution. Table 1 indicates that shoulder extension made a considerable contribution to medial hand motion in all four simulations. When maximum body roll was increased from 45 to 60°, this contribution increased by 8 cm in both the high and low elbow flexion conditions.

Vertical Hand Movement

The downward motion of the hand in the two simulations involving 60° of elbow flexion (Figure 12, c and d) was considerably greater than in the simulations involving 90° of elbow flexion (Figure 12, a and b). The displacement values of Table 2 were obtained by integrating the respective velocity–time functions. For example, the contribution of body roll to vertical hand displacement was defined as the definite integral of Equation 9 between $t = 0$ and $t_{PULL}/2$. Table 2 shows that an increase in elbow flexion from 60 to 90° reduced downward hand motion from 95 to 77 cm in the high-roll simulations and from 91 to 78 cm in the low-roll simulations.

When the elbow was allowed to flex to 90°, a distinct upward indentation in the handpath was produced prior to maximum body roll (Figure 12, a and b). This feature was not present in the handpaths when elbow flexion was reduced to 60° (Figure 12, c and d). This finding is consistent with the observations of Schleihauf (1979), who suggested that techniques which involved limited elbow flexion during the inward scull and elbow ex-

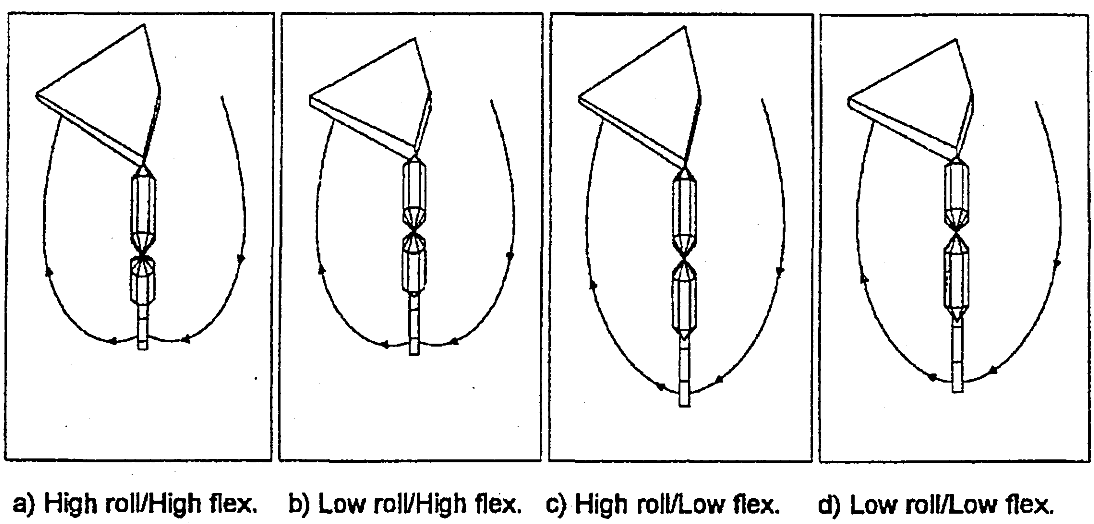


Figure 12 — Side view of the path followed by the right hand (relative to the water) in the four simulations. Handpaths shown are for the fast pull time (t_{PULL} = 0.75 s) and a trunk velocity of 1.6 m/s. The position of the trunk and upper extremity at the instant of maximum body roll is also shown.

Table 2 Vertical Hand Displacement Up to $t_{PULL}/2$ and the Contribution to This Displacement From Body Roll, Shoulder Extension, Elbow Flexion, and Shoulder Abduction (cm)

	High roll/ high flex	Low roll/ high flex	High roll/ low flex	Low roll/ low flex
Vertical displacement	–77	–78	–95	–91
Contribution				
Body roll	–11	–15	–18	–21
Shoulder extension	–65	–71	–65	–70
Elbow flexion	+49	+39	+24	+16
Shoulder abduction	–50	–31	–36	–16

tension during the outward scull produced handpaths without an upward indentation. Techniques that employed a greater range of movement at the elbow were characterized by handpaths with a distinct upward indentation.

A change in the maximum body roll angle had relatively little effect on the downward motion of the hand. When maximum body roll was increased from 45 to 60° in the low-flex simulations, an additional 4 cm of downward hand motion was produced. In contrast, the same increase in body roll in the high-flex simulations created 1 cm less downward hand motion (Table 2).

Body roll tended to assist downward hand motion in the first half of all four simulations. In the low-roll/low-flex simulation, body roll was responsible for 21 cm of the downward hand displacement compared to only 11 cm in the high-roll/high-flex simulation.

Equation 9 indicates that once the hand has moved beyond the midline of the trunk (where $\mathbf{r}_{XH}$ passes the vertical and δ is less than 90°), body roll makes a positive (upward) contribution to hand motion until maximum roll is reached. Thus, in the high-roll/high-flex simulation (Figure 12a), body roll aids downward hand motion until the hand crosses the midline of the trunk at a roll angle of approximately 44°. The remaining 16° of body roll reduces the downward hand motion. Because the hand does not reach the midline of the trunk in the other three simulations (Figure 12, b–d), body roll assists downward hand motion until maximum roll is achieved.

Backward Hand Movement

Although Equation 8 indicates that body roll has no effect on backward hand motion, a couple of observations are still worthy of comment. Figure 12, a–d, shows that the hand exited the water 67 cm behind the point at which it entered for each simulation run for the fast pull time (t_{PULL} = 0.75 s). Although this feature is characteristic of some world-class swimmers, many swimmers use the "classic technique" (Schleihauf, 1979), in which the hand leaves the water ahead of its point of entry. To achieve this, the swimmer's mean forward trunk speed must exceed the mean backward hand speed between hand entry and exit. When each simulation was conducted for the slow pull time (t_{PULL} = 1.1 s), this condition was met and the hand left the water 5 cm ahead of its entry point.

Model Evaluation

A number of assumptions are inherent in the model presented in this paper. The main assumptions are as follows:

1. The line joining the two shoulders (the shoulder axis) remains perpendicular to the swimming direction (the x-axis).
2. The right shoulder, elbow, and distal end of the hand remain in a plane perpendicular to the parasagittal plane.
3. The right forearm and hand act as a single rigid segment; that is, there is no wrist motion.
4. The shoulder always abducts to 90° regardless of the maximum elbow flexion and body roll angles.
5. Maximal values of body roll, elbow flexion, and shoulder abduction are all reached midway through the pull ($t_{PULL}/2$).
6. An increase in maximum body roll from 45° to 60° can be achieved without an increase in pull time.

Although the model appears to capture the essential features of the front crawl technique, further work is still necessary to fully evaluate each of these assumptions.

Conclusions

The following conclusions can be drawn from the results of this study:

1. Body roll influences both mediolateral and vertical hand motion in front crawl swimming and therefore contributes to hand speed in the plane perpendicular to the swimming direction.
2. For a given maximum elbow flexion, an increase in maximum body roll from 45° to 60° markedly increases medial hand motion but has relatively little effect on vertical hand motion.
3. For a given maximum body roll, an increase in maximum elbow flexion from 60 to 90° increases medial hand motion and reduces downward hand motion.
4. An increase in body roll is accompanied by an increase in the squared hand speed in the plane perpendicular to the swimming direction. Therefore, an increase in body roll increases the potential of the hand to develop propulsive lift forces.

Because the validity of the model presented in this paper has not yet been established, the results must only be considered as preliminary indications of how body roll affects hand speed and hand path. Nevertheless, the results suggest that coaches and front crawl swimmers should view body roll as a means of enhancing hand speed and consequently propulsion from lift forces, rather than as simply a reaction to other parts of the stroke.

References

Barthels, K.M. (1979). The mechanism for body propulsion in swimming. In J. Terauds & E.W. Bedingfield (Eds.), *Swimming III* (pp. 45–54). Baltimore: University Park Press.

Beekman, K.M. (1986). *Characteristics of the front crawl techniques of swimmers with shoulder impingement syndrome.* Unpublished master's thesis, University of Iowa, Iowa City.

Berger, M.A.M., Hollander, A.P., & de Groot, G. (1995). Contribution of lift force to propulsion in front crawl swimming. In *Proceedings of the XVth Congress of the International Society of Biomechanics* (pp. 102–103). Jyväskylä, Finland: International Society of Biomechanics.

Brown, R.M., & Counsilman, J.E. (1971). The role of lift in propelling the swimmer. In J. M. Cooper (Ed.), *Selected topics in biomechanics: Proceedings of the C.I.C. Symposium on Biomechanics* (pp. 179–188). Chicago: The Athletic Institute.

Counsilman, J.E. (1981). Hand speed and acceleration. *Swimming Technique*, **18**(1), 22–26.

Hay, J.G., Liu, Q., & Andrews, J.G. (1993). Body roll and handpath in freestyle swimming: A computer simulation study. *Journal of Applied Biomechanics*, **9**, 227–237.

Liu, Q., Hay, J.G., & Andrews, J.G. (1993). Body roll and handpath in freestyle swimming: An experimental study. *Journal of Applied Biomechanics*, **9**, 238–253.

Maglischo, E.W. (1982). *Swimming faster.* Palo Alto, CA: Mayfield.

Maglischo, C.W., Maglischo, E.W., Higgins, J., Hinrichs, R., Luedtke, D., Schleihauf, R.E., & Thayer, A. (1988). A biomechanical analysis of the 1984 U.S. Olympic freestyle distance swimmers. In B.E. Ungerechts, K. Wilke, & K. Reischle (Eds.), *Swimming science V* (pp. 351–360). Champaign, IL: Human Kinetics.

Prichard, R. (1993). A new swim paradigm: Swimmers generate propulsion from the hips. *Swimming Technique*, **30**(1), 17–23.

Ringer, L.B., & Adrian, M.J. (1969). An electrogoniometric study of the wrist and elbow in the crawl arm stroke. *Research Quarterly*, **40**, 353–363.

Schleihauf, R.E. (1979). A hydrodynamic analysis of swimming propulsion. In J. Terauds & E.W. Bedingfield (Eds.), *Swimming III* (pp. 70–109). Baltimore: University Park Press.

Schleihauf, R.E., Gray, L., & DeRose, J. (1983). Three-dimensional analysis of hand propulsion in the sprint front crawl stroke. In A.P. Hollander, P.A. Huijing, & G. de Groot (Eds.), *Biomechanics and medicine in swimming* (pp. 173–183). Champaign, IL: Human Kinetics.

Appendix 10

Payton, C.J. and Mullineaux, D.R. (1996). Effect of body roll on hand velocity in freestyle swimming. In *Biomechanics and Medicine in Swimming VII* (edited by J.P. Troup, A.P. Hollander, D. Strass, S.W. Trappe, J.M. Cappaert and T.A. Trappe), pp. 59-63. London: E & FN Spon.

8 EFFECT OF BODY ROLL ON HAND VELOCITY IN FREESTYLE SWIMMING

C.J. PAYTON
Division of Sport Science, The Manchester Metropolitan University, Alsager, UK
D.R. MULLINEAUX
School of Human Studies, University of Teesside, Middlesborough, UK

Abstract
A mathematical model was developed [1] to evaluate the contribution of body roll to medial-lateral (v_X) and vertical (v_Z) hand velocities in freestyle swimming. The right arm was modelled as two rigid segments joined at the elbow to enable flexion and extension. The arm was linked to a rigid trunk with a joint capable of shoulder extension and abduction/adduction.
Simulations were performed for pull times (t_{PULL}) from 0.7 s to 1.1 s and maximum body rolls (θ_{MAX}) from 50° to 70° with a straight arm, and with the elbow flexing through 90°. For a simulation involving elbow flexion (t_{PULL} = 0.7 s, θ_{MAX} = 60°), mean values for v_X and v_Z were 1.86 m/s and 2.24 m/s respectively. These values were 24% and 22% less, respectively, than those obtained from an equivalent simulation performed with a straight arm. These reduced velocities were due to the elbow flexion: 1) reducing the hand's radii of rotation, and 2) creating a hand velocity component which opposed that resulting from body roll.
It was concluded that body roll makes a substantial contribution to medial-lateral and vertical hand velocities in freestyle swimming and may therefore play an important role in the generation of propulsive lift forces.
Keywords: Body roll, freestyle swimming, hand velocity, modelling.

1 Introduction

When freestyle swimmers scull their hands vertically and medial-laterally, a forwards acting lift force is generated. It has been suggested that these lift forces can make a significant contribution to propulsion [2]. Since lift forces are proportional to the square of the hand speed, forwardly directed (propulsive) lift forces will be a function of the vertical and medial-lateral hand velocities produced by the swimmer.

Little attempt has been made to identify the body movements responsible for generating medial-lateral and vertical hand velocities in the freestyle. Although it seems logical that elbow flexion and shoulder adduction should make a contribution to hand velocity, a recent study [1] questioned the role played by elbow and shoulder movements. Using a mathematical model the authors demonstrated that medial deviations of the hand to the midline of the trunk could be achieved entirely with body roll, without the need for elbow flexion or shoulder adduction. It was also found that the amount of body roll required to sweep the hand to the midline was only 19°-34°. As this was well below the mean maximum body roll value of 60.8° exhibited by ten male university swimmers [3], it was suggested that swimmers may move their arms laterally relative to the trunk in order to avoid pulling the hand too far across the midline of the trunk.

As body roll appears to influence medial-lateral displacement of the hand relative to the water, it must also contribute to medial-lateral hand velocity. It then follows that body roll may assist in the generation of propulsive lift forces in freestyle swimming. The aim of this study was to determine the effect of body roll on medial-lateral and vertical hand velocities during the pull phase in freestyle swimming.

2 Methods

A previous three-dimensional model [1] was modified for this study. The right arm was modelled as two rigid segments hinged at the elbow (E) to enable flexion and extension. The arm was linked to a rigid trunk with a joint (S) capable of shoulder extension and shoulder abduction/adduction (Figure 1).

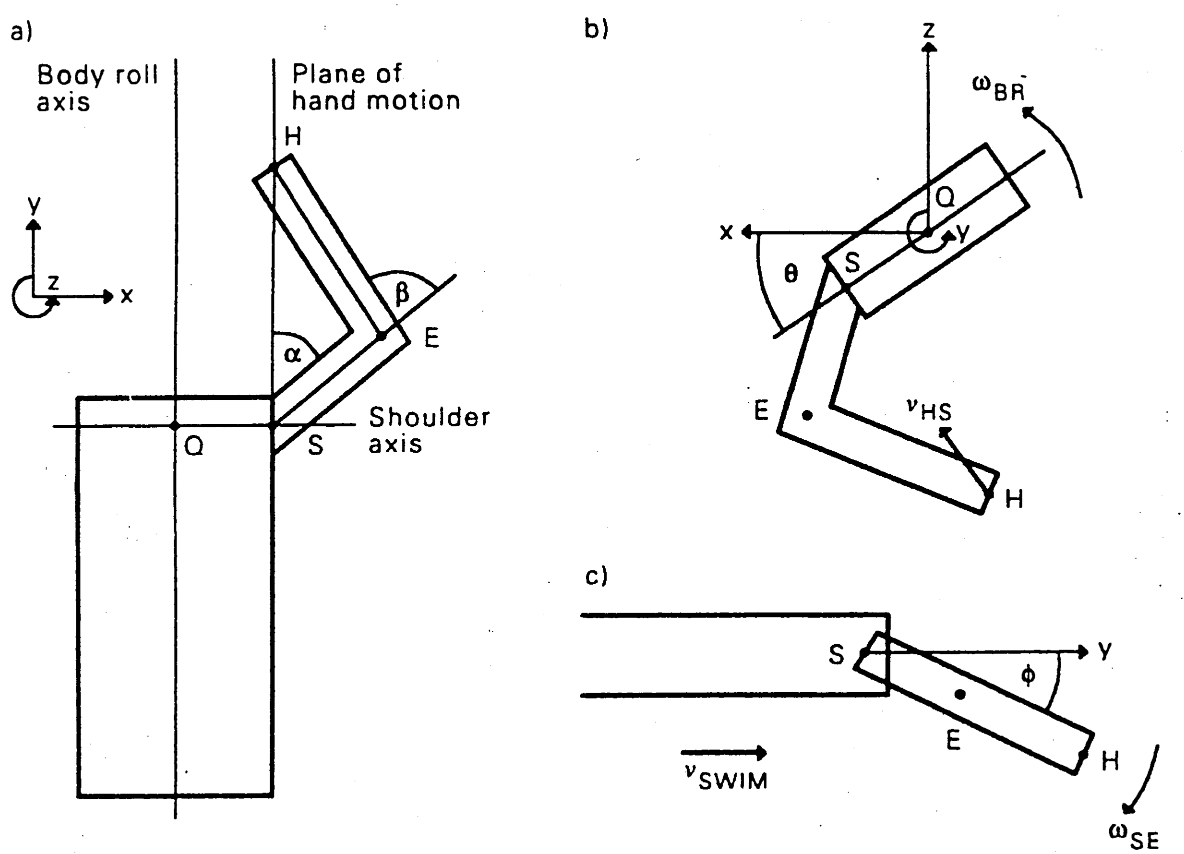


Figure 1. Model viewed in frontal plane (a), transverse plane (b) and sagittal plane (c).

The trunk was free to rotate about its long axis - the *body roll axis*. The hand (H) was constrained to move in the plane located through joint S and normal to the *shoulder axis* - the line connecting the two shoulder joints. The shoulder abduction angle (α) was therefore a function of the elbow angle (β) to ensure the hand remained in this plane - the *plane of hand motion*.

The half shoulder width (l_{QS}), upper arm length (l_{SE}) and forearm plus hand length (l_{EH}) were assigned values of 0.25 m, 0.35 m and 0.5 m respectively, based on reported anthropometric data on competitive swimmers [3].

Simulations were run for fast (0.7 s), medium (0.9 s) and slow (1.1 s) pull times. Within this time:

1. The trunk rotated from a neutral position ($\theta = 0°$) to a pre-selected maximum body roll angle (θ_{MAX}) of either 50°, 60° or 70°, and back to the neutral position.
2. The shoulder extended through 180° from a position of full flexion ($\phi = 0°$).
3. The elbow flexed through 90° from an extended position ($\beta = 0°$) and then fully extended again.

The body roll and elbow flexion angular velocities (ω_{BR} and ω_β respectively) were modelled as sine functions such that they were zero at the start, midpoint ($t_{PULL}/2$) and end of each simulation. The shoulder extension angular velocity (ω_{SE}) was held constant in each simulation. The effect of ω_{BR} on hand velocity was dependent on the perpendicular vector from the hand to the body roll axis (l_{YH}). Similarly, the effect of ω_{SE} on hand velocity was a function of the shoulder to hand vector (l_{SH}).

The elbow flexion angular velocity, coupled with the shoulder abduction angular velocity (ω_α), produces a hand velocity component (v_{HS}) directed toward the shoulder (Figure 1b). The effect of ω_α and ω_β on v_{HS} was determined by the lengths of the two arm segments (l_{SE} and l_{EH}).

The equations used to compute the medial-lateral (v_X), superior-inferior (v_Y) and vertical (v_Z) components of the hand velocity, relative to a pool-fixed Cartesian reference frame (Figure 1a), are presented in Table 1.

Table 1. Equations for computing the Cartesian components of the hand velocity.

	Body roll	Shoulder extension	Elbow flexion		
$v_X =$	$\omega_{BR} . l_{YH} . \sin\delta$	$+\ \omega_{SE} . l_{SH} . \cos\phi . \sin\theta$	$+\ v_{HS} . \sin\phi . \sin\theta$		(1)
$v_Y =$		$\omega_{SE} . l_{SH} . \sin\phi$	$+\ v_{HS} . \cos\phi . \cos\theta$	$+\ v_{SWIM}$	(2)
$v_Z =$	$\omega_{BR} . l_{YH} . \cos\delta$	$+\ \omega_{SE} . l_{SH} . \cos\phi . \cos\theta$	$+\ v_{HS} . \sin\phi . \cos\theta$		(3)

where δ is the angle vector l_{YH} makes to the vertical.

3 Results

Table 2 presents the mean hand velocity components from a representative simulation performed without elbow flexion ($\theta_{MAX} = 60°$, $t_{PULL} = 0.7$ s). Table 3 presents mean hand velocity components from an equivalent simulation performed with the elbow flexing through 90°. Mean hand velocity components were defined as the mean absolute value (magnitude) of the hand velocity component from $t = 0$ to t_{PULL}.

The contributions made by body roll, shoulder extension and elbow flexion to the mean hand velocity components are also shown in Tables 2 and 3. The contributions made by elbow flexion to $\overline{v}_X$ and $\overline{v}_Z$ have been assigned a negative value. This indicates that elbow flexion served to reduce the mean value of these two velocity components. Swimming velocity (v_{SWIM}) was assumed to remain constant within each simulation and was given an arbitrary value of 1.8 m/s. This value has been included in the calculation of the mean superior-inferior ($\overline{v}_Y$) hand velocity values.

Table 2. Mean hand velocity components from a simulation with 60° of body roll, a pull time of 0.7 s, a swim velocity of 1.8 m/s and no elbow flexion.

	Body roll contribution (m/s)	Shoulder ext. contribution (m/s)	Mean velocity (m/s)
$\overline{v}_X$	1.67	0.78	2.45
$\overline{v}_Y$	-	2.43	0.63
$\overline{v}_Z$	0.71	2.17	2.88

Table 3. Mean hand velocity components from a simulation with 60° of body roll, a pull time of 0.7 s, a swim velocity of 1.8 m/s and the elbow flexing through 90°.

	Body roll contribution (m/s)	Shoulder ext. contribution (m/s)	Elbow flexion contribution (m/s)	Mean velocity (m/s)
$\overline{v}_X$	1.52	0.68	-0.35	1.85
$\overline{v}_Y$	-	2.05	0.30	0.55
$\overline{v}_Z$	0.56	2.07	-0.39	2.24

4 Discussion

The effect of body roll on the medial-lateral (v_X), superior-inferior (v_Y) and vertical (v_Z) hand velocities are described by equations 1, 2 and 3 respectively. The equations reveal that body roll angular velocity has a direct influence on medial-lateral and vertical hand velocities but has no effect on the superior-inferior component.

When the body rolls through 60° in 0.35 s ($t_{PULL}/2$) with a fully extended arm, the mean medial-lateral and vertical hand velocities produced are 2.45 m/s and 2.88 m/s respectively (Table 2). For identical body roll conditions, but with the elbow flexing through 90°, the mean medial-lateral hand velocity is reduced by 24% to 1.85 m/s. There is also a 22% reduction in the mean vertical hand velocity to 2.24 m/s resulting from elbow flexion. Although it has previously been suggested [4] that elbow flexion makes a positive contribution to medial-lateral and vertical hand velocities, these results do not support this view.

Elbow flexion reduces the mean medial-lateral and vertical hand velocities in two ways. Firstly, elbow flexion reduces the shoulder to hand distance l_{SH}. Consequently, for any given shoulder extension angle ϕ, the perpendicular distance from the hand to the body roll axis l_{YH} is also reduced. These shortened radii reduce the body roll and shoulder extension contributions to the mean medial-lateral and vertical velocities (Table 3). Secondly, elbow flexion, coupled with shoulder abduction, produces a hand velocity component (v_{HS}) directed toward the shoulder (Figure 1b). The medial-lateral and vertical components of v_{HS} will always oppose the direction of the body roll and shoulder extension contributions to hand velocity.

Of the 24% reduction in medial-lateral hand velocity caused by elbow flexion, 10% is attributable to the shortening of l_{SH} and l_{YH}. The remaining 14% reduction is due to the direction of the medial-lateral component of v_{HS}. Similarly, of the 22% decrease in vertical hand velocity, 8% results from the shortened radii with the vertical component of v_{HS} accounting for the remaining 14%.

In addition to directly contributing to hand velocity ($\omega_{BR} \wedge l_{YH}$), body roll also influences the medial-lateral and vertical velocicities produced by shoulder extension. Equation 1 shows that once body roll has commenced ($\theta \neq 0$), shoulder extension makes a contribution to medial-lateral hand velocity. For a given shoulder extension angle ϕ, the greater the body roll, the greater is the contribution. The inverse is true for vertical hand velocity (Equation 3), as body roll increases (for a given ϕ) the magnitude of this component decreases.

The validity of the model presented in this paper has not yet been established. The results must therefore only be considered as tentative indications of how body roll affects hand velocity in the freestyle. Nevertheless, body roll may make a substantial contribution to medial-lateral and vertical hand velocities in freestyle swimming and therefore play an important role in the generation of propulsive lift forces.

5 References

1. Hay, J.G., Liu, Q., and Andrews, J.G. (1993) Body roll and handpath in freestyle swimming: a computer simulation study. *Journal of Applied Biomechanics*, Vol. 9, No. 3. pp. 227-237.
2. Schleihauf, R.E., Gray, L., and DeRose, J. (1983) Three-dimensional analysis of hand propulsion in the sprint front crawl stroke, in *Biomechanics and Medicine in Swimming*, (eds. A.P. Hollander, P.A. Huijing and G. de Groot), Human Kinetics, Champaign, pp. 173-183.
3. Liu, Q., Hay, J.G., and Andrews, J.G. (1993) Body roll and handpath in freestyle swimming: an experimental study. *Journal of Applied Biomechanics*, Vol. 9, No. 3. pp. 238-253.
4. Barthels, K.M. (1979) The mechanism for body propulsion in swimming, in *Swimming III*, (eds. J. Terauds and E.W. Bedingfield), University Park Press, Baltimore, pp. 45-54.

Appendix 11

Mullineaux, D.R. and Barnes, C.A. (1998). Determinants of propensity to partake in adequate physical activity to promote health. *Journal of Sports Sciences*, **16**, 394-5.

promoting health by reducing the risk for development or recurrence of disease, or enhancing physical fitness. These reduced risks result in improved health and feelings of well-being, a better quality of life, lower costs for individuals, government and industry, and a lower incidence of disease.

In light of the potential health-related benefits of physical activity, recommendations for appropriate and safe exercise prescriptions have been proposed. With regard to cardio-respiratory endurance, individuals are recommended to partake in rhythmic activity at an intensity of 55–90% of maximum heart rate, 20–60 min, 3–5 days per week. These are similar to the level 5 recommendations included in the Health Education Authority's report on the Allied Dunbar National Fitness Survey (ADNFS), that an intensity of 60–80% of maximum heart rate should be maintained for at least 20 min on a minimum of three occasions per week. Recently, it has been suggested that health benefits can be accrued through exercising at lower intensities (Pate *et al.*, 1995), including 30 min of 'moderate' activity on most days of the week.

On the basis of the ACSM recommendations, the Centers for Disease Control and Prevention (CDCP) reported that approximately 73% of Americans' physical activity levels are insufficient. Similarly, in the UK, results from the ADNFS revealed that approximately 70% of English adults do not engage in sufficient physical activity to gain health benefits. Existing analyses tend to be uni-factorial and descriptive and do not identify those factors which are associated with a sedentary lifestyle. The use of a logistic regression analysis considers the combined effect of variables and predicts the 'odds' for whether an individual is sedentary or is sufficiently active to reap the associated health benefits. The aim of this study was to produce a multifactorial model to indicate individual propensity to participitate in an adequate level of physical activity to promote health.

The database of the ADNFS of 1990 contained responses from 4316 subjects (16+ years old) to multiple levels of 122 questions on mental, physical, social and environmental factors associated with involvement in physical activity. Responses to questions on mode, intensity, duration and frequency of physical activity at home, work and during leisure time were combined to define subjects as sedentary (involved in no moderate or vigorous bouts of activity) or sufficiently active to accrue benefits to health (fulfilling either the level 5 criteria of the ADNFS or the recommendations of CDCP/ ACSM. Twenty potential independent variables were selected on the basis of their use in previous research. To maintain a high level of power, only variables with more than a 90% response rate were included in the analysis.

A logistic regression was run on SPSS on Windows v. 6.1 to predict the odds of membership to either the sedentary ($n = 870$) or active ($n = 1135$) groups. The analysis was a forward likelihood model. By including the 'constant' in the model, a 'reference person' was defined by the level of each independent variable that described the person as least likely to be active. The independent variables were defined as indicators and the basis for their inclusion in the model was dictated by setting statistical significance to an alpha level of 0.05.

From the population of 4316 subjects, 20.5% ($n = 883$) were categorized as sedentary and 26.5% ($n = 1144$) were deemed sufficiently active to gain health benefits. Con-

Determinants of propensity to partake in adequate physical activity to promote health

D. Mullineaux and C. Barnes

School of Social Sciences, University of Teesside, Middlesbrough, UK

Appropriate physical activity is widely accepted as being beneficial for health and fitness. The American College of Sports Medicine (ACSM) summarizes the benefits as either sequently, 53.0% ($n = 2289$) of the sample were physically active to a lesser degree than that deemed beneficial for health, and were excluded from the regression analysis. The respective percentages are <1% different to those from the ADNFS. These values are not directly comparable, however, as the criteria for group membership were not corrected for age and sex as in the ADNFS.

The accuracy of the logistic regression model was good, correctly predicting 72% of the sedentary group and 89% of the active group. More importantly, the low 'goodness-of-fit' value (1909 and 1985 df) indicates that the model has a high 'probability power' to predict correctly. The final model contained six independent variables: one 'physical' (age), one 'social' (academic qualifications) and four 'attitude-based' (motivation to exercise; physically active; exercise enough; lifestyle problems). Alternative variables that were not included were hypothesized as having a small effect, or having an effect catered for by the variables retained in the analysis. Age has the greatest effect on physical activity participation and accounts for almost half the power, but all variables contribute statistically significant ($P < 0.05$) power to the model.

The 'reference person' (defined as: 75+ years; possessing no qualifications or of CSE grades 2–5); very low motivation to be active; perceives themselves as physically inactive; does not know whether they exercise enough for benefits; and has lifestyle problems from health) has odds of 0.0026, the reciprocal of which indicates they are 385 times more likely to be sedentary than to be active. Changing the 'reference person' to the most likely person to be active (defined as: 25–34 years; possessing other qualifications; highly motivated; very active; gets enough exercise; and has no lifestyle problems) results in someone who is 31 times more likely to be active than to be sedentary. Thus, persons with a tendency to be active are generally younger, possess further qualifications, are highly motivated, perceive themselves as active and getting enough exercise, and have no lifestyle problems.

The small percentage of people involved in a sufficient amount of physical activity, as defined in recent recommendations, emphasizes the need for effective health promotion campaigns. This model could provide a tool for developing, evaluating and prescribing health promotion initiatives by identifying and quantifying target groups who face health problems due to inactive lifestyles. Efficient and economical health promotion may warrant targeting sedentary individuals identified from the model as having better odds to become active. Further research is needed to confirm why certain factors influence participation in physical activity, and evaluate the effects of associated health promotion initiatives.

Appendix 12

Mullineaux, D.R., Barnes, C.A. and Barnes, E.F. (2001a). Factors affecting the likelihood to engage in adequate physical activity to promote health. *Journal of Sports Sciences*, **19**, 279-88.

Journal of Sports Sciences, 2001, 19, 279–288

Factors affecting the likelihood to engage in adequate physical activity to promote health

DAVID R. MULLINEAUX,[1]* CHRISTOPHER A. BARNES[2] and ELIZABETH F. BARNES[3]

[1]*School of Sport and Leisure Management, Sheffield Hallam University, Collegiate Hall, Sheffield S10 2BP,* [2]*Middlesbrough Football Club Training Headquarters, Hurworth Place, Darlington DL2 2DH and* [3]*School of Health, University of Teesside, Middlesbrough TS1 3BA, UK*

Accepted 18 December 2000

The aim of this study was to assess the likelihood of individuals to participate in enough physical activity to promote fitness and, more conservatively, to accrue only health benefits. Sedentary ($n = 883$; 20.5%) and active ($n = 1144$; 26.5%) groups were identified from the 1990 Allied Dunbar National Fitness Survey of English adults ($n = 4316$). The data were analysed using logistic regression. Participants were described using 20 variables identified from previous research, six of which made a significant contribution to the model ($P < 0.05$). The odds of being sedentary increased with age, self-perception of lifestyle problems, and lower scores on education, self-perception of motivation to exercise, perception of own participation in physical activity and recognition of exercising enough for health benefits. The odds of being active were associated with the opposite characteristics to those observed for sedentary behaviour. The extreme scores varied from individuals who may be 385 times more likely to be sedentary, to those who were 29 times more likely to be active, depending on scores on the selected variables. The results of this study provide a means to determine individual propensity to participate in adequate physical activity, and to identify those who may benefit most from health promotion campaigns.

Keywords: age, education, health promotion, odds, sedentary, self-perception.

Introduction

Appropriate physical activity is widely accepted as being beneficial for health and fitness. The American College of Sports Medicine (ACSM, 1995) summarized the benefits as either promoting health by reducing the risk for development or recurrence of disease, or enhancing physical fitness. These reduced risks result in improved health and feelings of well-being, better quality of life, lower costs for individuals, government and industry, and a lower incidence of disease. Increased participation in physical activity by adults reduces the risk of coronary heart disease, stroke, hypertension, non-insulin-dependent diabetes mellitus, osteoporotic fractures, depression and some cancers (Riddoch and Boreham, 1995).

On the basis of the ACSM recommendations, the US Centers for Disease Control and Prevention (CDCP, 1995) reported that approximately 73% of Americans' physical activity is insufficient. This tendency towards a sedentary lifestyle has been estimated to be responsible for 250,000 premature deaths a year in the USA (Hahn *et al.*, 1990). Similarly, in the UK, results from the Allied Dunbar National Fitness Survey (ADNFS) revealed that approximately 70% of English adults do not engage in sufficient physical activity to accrue health benefits (Activity and Health Research, 1992).

Several large surveys of English adults have identified factors indicating those individuals who engage in sufficient physical activity to gain health benefits. In the most comprehensive survey of physical activity patterns of English adults ($n = 4316$), the ADNFS showed activity to vary greatly between the sexes, across age groups and from one socio-economic group or ethnic group to another (Activity and Health Research, 1992). Similar findings were reported in the UK Health Education Authority's (HEA) National Survey of Activity and Health ($n = 2837$; Walker and Hoinville, 1995) and in the Department of Health's Health Survey of England

* Author to whom all correspondence should be addressed. e-mail: d.mullineaux@shu.ac.uk

Journal of Sports Sciences ISSN 0264-0414 print/ISSN 1466-447X online
http://www.tandf.co.uk/journals

in 1991 and annually since 1993 (for example, in 1998, $n = 15,908$; Prior, 1999). The analysis of these more recent surveys for the purpose of identifying factors that determine individual involvement in physical activity is less appropriate than the ADNFS, as they used reduced methods of the ADNFS and analysed fewer factors. The factors identified for the UK are similar to those in other Western countries. For example, in the USA, analysis of the Third National Health and Nutrition Examination Survey from 1988 to 1994 revealed that physical inactivity is more prevalent for the less educated, those living below the poverty line, those living in households with incomes below US$20,000 and retired persons (Crespo *et al.*, 1999). More generally, the USA's National Institutes of Health (Luepker *et al.*, 1996) consensus statement on physical activity and cardiovascular health identified age, socio-economic status, cultural influences and health status as factors that influence adopting and maintaining a physically active lifestyle.

Although much information has been accumulated from such surveys on health status and behaviours, analysis of the resultant data has often tended to be unifactorial and descriptive, and thus may have been limited in its power to identify key factors associated with a sedentary lifestyle. For example, although a strong association was found between physical inactivity and multiple indicators of social class (Crespo *et al.*, 1999), the interrelation between these variables was not assessed. Additional limitations apply to the ADNFS classification of 'appropriate physical activity' for health benefits. The level 5 criteria within the ADNFS (Fentem *et al.*, 1994) are the same as the HEA's (1995) target of vigorous exercise at 60–80% of maximum heart rate for at least 20 min on a minimum of three occasions per week. The analysis does not consider recent recommendations on exercise prescription, such as those proposed by Pate *et al.* (1995), who suggested that cumulative bouts of lower intensity exercise could also potentially improve health status such as through accumulation of 30 min of moderate activity on most days of the week. This has been reflected in the ACSM (1998) modifying its exercise prescription guidelines, such that the 20–60 min of continuous exercise can also be accumulated in 10 min bouts throughout the day. Within these ACSM guidelines, moderate activity is defined as exercising between 55 and 69% of maximum heart rate.

A greater understanding of the factors that have a positive influence on patterns and amounts of participation in physical activity should facilitate the implementation of more efficient and effective health promotion strategies and, subsequently, lead to improved health status of the nation through more active lifestyles. Such factors may not appear significant in isolation and, indeed, may only be shown to exert an influence in conjunction with others. As such, before producing recommendations regarding promotional activities, a thorough data analysis must be undertaken. Many multi-factorial statistical tools exist and are routinely used for such purposes. Logistic regression is one such tool that can be used to determine the combined effect of a number of independent variables, of any level of measurement and not necessarily normally distributed, towards predicting the odds of a participant falling in one category or another on a dichotomous dependent variable. The latest Health Survey of England (Prior, 1999) in 1998 did consider more recent recommendations on exercise prescriptions (e.g. Pate *et al.*, 1995) and did use logistic regression, but the number of variables used was small and still the analysis was unifactorial in that several independent variables were not combined in the logistic regression analysis. The aim of this study was to conduct a secondary analysis on the data of the Allied Dunbar National Fitness Survey to produce a multifactorial model to predict individuals' likelihood to participate in adequate physical activity to promote health while catering for more recent exercise prescription guidelines.

Methods

The Allied Dunbar National Fitness Survey (ADNFS) of 1990 is considered to provide a good estimate of the physical activity profile of the English population (Activity and Health Research, 1992). Permission was obtained from the UK Sports Council and Health Education Authority to perform a secondary data analysis on the ADNFS database provided by The Data Archive (Activity and Health Research, 1995). Full details of the methods used in the ADNFS are described elsewhere (Fentem *et al.*, 1994), a summary of which is presented below.

The ADNFS contained responses from 4316 randomly sampled participants (16+ years old) to multiple levels of 122 questions on mental, physical, social and environmental factors associated with involvement in physical activity. The responses to questions about the mode, intensity, duration and frequency of physical activity, whether at home, in the workplace or during leisure time, were combined in this study to re-categorize participants as either 'sedentary' or 'active'. Where no moderate or vigorous bouts of activity were reported, the participant was deemed 'sedentary'. The definition within this study for 'active' (i.e. sufficient to accrue benefits to health), therefore, was defined using both the level 5 criteria of the ADNFS (Fentem *et al.*, 1994) and the recommendations of Pate *et al.* (1995), and are outlined in detail in the introduction to this paper. To identify those characteristics that determine or predict whether an individual is likely to engage in adequate physical activity, 20 independent variables (detailed in Table 1), which were not used to categorize participants as either sedentary or active, were selected from the database. Many of these have been cited in previous research (e.g. Activity and Health Research, 1992; Walker and Hoinville, 1995; Luepker *et al.*, 1996; Crespo *et al.*, 1999; Prior, 1999) together with explanations for their effects on participation in physical activity. For ease of interpretation, all independent variables were analysed by dividing the responses into characteristic groupings. These are described in Table 1 and in greater detail for the subsequent variables retained in the logistic regression model (Table 2).

A logistic regression was run on SPSS (1999) to predict the odds of membership to either the sedentary ($n = 883$) or active ($n = 1144$) groups. The analyses used the forward likelihood-ratio selection procedure and the 'constant' was included in the model so that the results defined Reference Characteristics that describe one person by the first characteristic from each of the independent variables retained in the model. The remaining characteristics within each variable are collectively termed the Alternative Characteristics. The variables were defined as indicators; the basis for their inclusion in the model was dictated by setting alpha at 0.05, a value that is considered appropriate for research in the social sciences (Cohen, 1988).

As logistic regression selects only participants that possess responses to all independent variables, the analysis was re-run with the unselected variables removed. The final model resulted in the deletion of 13 participants from the sedentary group and nine from the active group. When it is possible to define the Reference Characteristics by the level within each independent variable possessing the lowest odds, then only multiplication and no division is required in using the logistic regression model. To simplify the results, the regression was re-run to define the Reference Characteristics by the level within each independent variable possessing the lowest odds of being active. The odds of being active for a person with Alternative Characteristics can then be obtained by multiplying the odds of the Reference Characteristics with up to one Alternative Characteristic's odds within each variable. Where the Reference and Alternative Characteristics within a variable are not statistically significantly different, then no multiplication is required (see 'Using the logistic regression model' below for clarification).

Table 1. Independent variables selected for the logistic regression grouped under physical, social and self-perception types

Type	Variable[a]	Description
Physical	*Age*	All age groups from 16 years and upward
	Diet	Diet's nutritional quality
	Health status	Self-appraisal of current health
	Health threshold	Self-appraisal of resistance to poor health based on existing health problems
	Sex	Male or female
	Smoking	Smoker, ex-smoker or non-smoker
Social	*Education*	Highest educational certificate obtained
	Ethnic/racial background	Classified in accordance with the 1991 census (OPCS)
	Marital status	Married, single or other
	Social class	Based on the 'classification of occupations' (OPCS, 1980)
	Socio-economic group	Based on 'head of household' (OPCS, 1980)
Self-perception	*Active*	Self-report on level of participation in physical activity
	Adequate exercise	Perception of whether participate in enough exercise to benefit health
	Fitness	Self-report on perception of own fitness
	Importance of activity	Perception of the importance of activity for health
	Lifestyle problems	Self-report of any health problems that affect lifestyle
	Motivation	The recognition that exercise will contribute to achieving some important personal goals
	One vigorous bout	Participation in this much activity per week: yes or no
	Three vigorous bouts	Participation in this much activity per week: yes or no
	Well-being	Positive outlook on life

[a] Italics indicate the six variables that were included in the resultant regression model and which make up the Reference and Alternative Characteristics depending on the level of the characteristic within the variable.

Table 2. Independent variables and their characteristics used to determine the Reference and Alternative Characteristics as retained in the logistic regression model and their frequency of responses for all participants ($n = 4316$) and for the sample in this study ($n = 2005$)

		All participants		This study	
Variable	Characteristic [a]	Frequency	%	Frequency	%
Age [b]	75+	367	8.5	246	12.3
	65–74	556	12.9	270	13.5
	55–64	616	14.3	276	13.8
	45–54	688	15.9	295	14.7
	35–44	740	17.1	313	15.6
	25–34	773	17.9	336	16.8
	16–24	576	13.3	269	13.4
Education	None [c]	1799	41.7	937	46.7
	School [d]	1415	32.8	604	30.1
	Other [e]	1102	25.5	464	23.1
Active [f]	Not at all	242	5.6	167	8.3
	Not very	865	20.0	354	17.7
	Fairly	2383	55.2	1038	51.8
	Very	790	18.3	446	22.2
Adequate exercise	Don't know	117	2.7	60	3.0
	No	1893	43.9	760	37.9
	Yes	2306	53.4	1185	59.1
Lifestyle problems	Some	3185	73.8	1398	69.7
	None	1131	26.2	607	30.3
Motivation	Very low	812	18.8	475	23.7
	Low	336	7.8	145	7.2
	Moderately low	674	15.6	270	13.5
	Moderately high	1076	24.9	452	22.5
	High	917	21.2	418	20.8
	Very high	501	11.6	245	12.2

[a] The first characteristic from each of the six variables defines the Reference Characteristics; the remaining characteristics are the Alternative Characteristics.
[b] Age of all participants = 47 ± 19 years (mean ± s; range = 16–96 years); age of sample in the present study = 44 ± 16 years (range = 16–74 years).
[c] Includes grades 2–5 in the UK Certificate of Secondary Education (CSE) qualification taken at age 16 years.
[d] 16- or 18-year-old school leaving certificate higher than CSE grades 2–5.
[e] Vocational or higher academic qualifications than for the School characteristic.
[f] Thirty-six non-responses.

Results

Group constituents

The percentages of the sample classified as sedentary and active are shown in Table 3. The respective percentages for this study are approximately no more than 1% different from those identified in the 'ADNFS modified criteria' (Activity and Health Research, 1992). This demonstrates that the present study has classified the physical activity patterns of the English adult population in a similar fashion to the Allied Dunbar National Fitness Survey (ADNFS). These values are not directly comparable, however, as the criteria for group membership in this study were not adjusted for age and sex as in the 'ADNFS modified criteria' (see Table 3). Twenty-two participants were excluded from the final analysis as they failed to respond to one or more questions. This resulted in 870 sedentary and 1135 active participants (i.e. 20.2% and 26.3% of the original sample, respectively).

Table 3. Percentage of the sample for all ages ($n = 4316$) and 16- to 74-year-olds ($n = 3949$) defined as active, irregularly active and sedentary based on the criteria of the original and modified Allied Dunbar National Fitness Survey (ADNFS) and of the present study

	ADNFS original criteria[a]		ADNFS modified criteria[b]		Present study[c]	
Group	All ages	16–74 years	All ages	16–74 years	All ages	16–74 years
Active	8.2	9.0	27.5	29.3	26.5	28.6
Irregularly active[d]	71.2	74.7	51.9	54.4	53.0	55.0
Sedentary[e]	20.6	16.3	20.6	16.3	20.5	16.4

[a] Active is defined as fulfilling the level 5 criteria of the ADNFS (Fentem *et al.*, 1994).
[b] Active is defined as fulfilling the level 5 criteria of the ADNFS (Fentem *et al.*, 1994) and controlling for age and sex (Activity and Health Research, 1992) where less physical activity is required for health benefits with increases in age and smaller increases in age for women.
[c] Active is defined as fulfilling the level 5 criteria of the ADNFS (Fentem *et al.*, 1994) or the recommendations of Pate *et al.* (1995).
[d] Engaging in physical activity but less than either an accumulation of 30 min of moderate exercise in at least 10 min bouts (Pate *et al.*, 1995) or three 20 min bouts of vigorous exercise (Fentem *et al.*, 1994) per week.
[e] Engaging in no physical activity sufficient to gain health benefits (i.e. no bouts of at least 10 min of moderate exercise).

Independent variables

The accuracy of the logistic regression model resulted in the correct prediction of 72% of the sedentary group, 89% of the active group and 81% overall. The Hosmer and Lemeshow goodness-of-fit ($\chi^2 = 3.70$, d.f. = 8, $P = 0.88$), which compares the observed to the predicted probabilities, was non-significant, indicating that the model fits well. Although issues related to statistical power are not yet covered for logistic regression, including this goodness-of-fit value that is analogous to the standard error in linear regression (Munro, 2001), the high P-value for the goodness-of-fit suggests that the model's ability to predict correctly is high. This is confirmed by the large extremes of the odds for predicting some participants to either the sedentary or active groups.

The final model contained six independent variables: one 'physical' (age), one 'social' (education) and four 'self-perception' (active, adequate exercise, lifestyle problems and motivation). All variables have been included in the model as categorical variables, of which columns 1 and 2 of Table 2 describe their name and characteristics, respectively. The frequencies for each characteristic within each variable tended to have an equal distribution with no individually small sample proportion that is beneficial for logistic regression (Menard, 1995). The sequence of inclusion of the independent variables in the model was determined by a significant improvement in χ^2, as detailed in Table 4. This value indicates the explanatory power of the model (Munro, 2001). Age had the greatest effect on physical activity participation and accounted for almost half the explanatory power, but all variables contributed statistically significant explanatory power to the model ($P \leq 0.001$).

Logistic regression model

The final model is presented in Fig. 1, which illustrates the variables, Reference Characteristics and Alternative Characteristics for the model. The numbers are the beta coefficients (β) presented as odds (i.e. odds = e^{β}), of which the constant is referred to as the Reference Characteristics. The calculations of each odds value to be active, for a person defined by combinations of these characteristics, can take account of standard error estimates in their presentation, although for simplicity of presentation, calculation and interpretation of these have not been included or used.

Using the logistic regression model

The logistic regression model illustrated in Fig. 1 was used to identify the likelihood of a categorized person participating in adequate physical activity. The Reference Characteristics (i.e. those characteristics described in the middle box of the model) have an 'odds value' of 0.0026, and suggest this is the least likely person defined by the model to be active. An odds value of less than 1 indicates that the defined person is more likely to be sedentary; the lower this value, the greater this tendency. The reciprocal of odds less than 1 provides the odds to be sedentary; hence the reciprocal of 0.0026 indicates that a person possessing these characteristics is 385 times more likely to be sedentary than to be active. In this example, the Reference Characteristics are defined as: 75+ years old; perceives oneself as not at all active, very low motivation to be active and don't know whether I participate in adequate exercise for benefits; possessing no education; and perceives oneself has some lifestyle problems from health.

Multiplying the odds for the Reference Characteristics with one Alternative Characteristic's odds, from any number of variables, provides the odds for a newly described person to be active. For example, changing the Reference Characteristics to the following – 25–34 years old; perceives oneself as very active, very highly motivated and gets adequate exercise; possessing other education; and perceives oneself has no lifestyle problems – provides odds of 29 (i.e. $0.0026 \times 22.1 \times 11.6 \times 5.2 \times 2.9 \times 1.8 \times 1.6$). This person is 29 times more likely to be active than to be sedentary and is the most likely person to be defined by the model as active. An odds value of greater than 1 indicates that the defined person is more likely to participate in physical activity; the greater this value, the greater this tendency.

Table 4. Sequence of inclusion of variables in the logistic regression, including an indication of the model's explanatory power and accuracy to predict physical activity status

Variable	Explanatory power[a] χ^2	d.f.	Accuracy[b] (% correct)
Age	698	6	77
Active	963	9	80
Motivation	1028	14	81
Adequate exercise	1053	16	81
Education	1072	18	81
Lifestyle problems	1084	19	81

[a] The inclusion of variables provides a significant improvement in the explanatory power of the model to predict participants as either sedentary or active as indicated by χ^2 ($P \leq 0.001$).

[b] The overall accuracy of the model to correctly predict participants as either sedentary or active.

As all the Alternative Characteristics' odds are greater than zero, changing any characteristic from the Reference Characteristics will increase the odds for a person to be active. Also, the further the deviation from the first Alternative Characteristic, within a variable, the greater the likelihood the individual will be active (with the exception of the youngest age category where the odds reduce slightly). Defining a person with any combination of Alternative Characteristics makes a statistically significant difference ($P < 0.05$) from the Reference Characteristics in all except for two conditions. There are no significant differences between defining a person as perceiving they have a 'very low' or 'low' motivation to be active, and between whether they perceive that they 'don't know' or 'do not' get enough exercise. If a person possesses the latter element to either pair of these responses, then changing to these Alternative Characteristics from the Reference Characteristics is not required. Essentially, individuals with a tendency to be active are generally younger, perceive themselves to be active, getting enough exercise and highly motivated to be active, more educated and having no lifestyle problems. Example individuals identified from the logistic regression model are provided in Table 5, indicating their odds to be active at equal intervals on a linear scale of the beta coefficients (β), where $\beta = \ln(\text{odds})$.

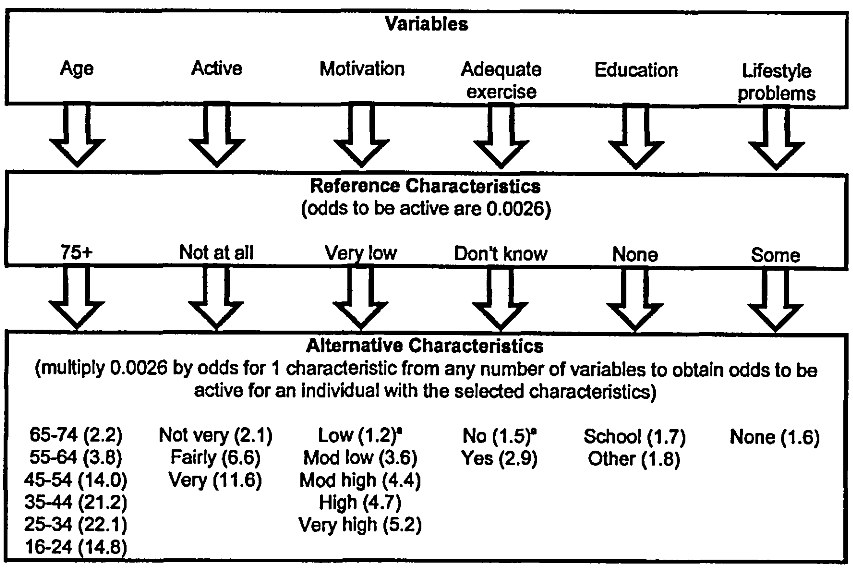


Fig. 1. Simultaneous logistic regression model for calculating individuals' odds to participate in adequate physical activity. The Reference Characteristics odds of 0.0026 should be multiplied by the odds for one Alternative Characteristic from any number of variables to obtain new odds for the defined individual to be active. [a]No statistically significant difference from the respective Reference Characteristics ($P > 0.05$).

Table 5. Exemplar person characteristics and their odds to be active ranging from the most likely to be sedentary to the most likely to participate in adequate physical activity for health benefits

Exemplar characteristics						
Age (years)	Active	Motivation	Adequate exercise	Education	Lifestyle problems	Odds to be active[a]
75+	Not at all	Very low	Don't know	None	Some	0.0026 (384.6)
55–64	Not at all	Very low	Don't know	School	Some	0.0168 (59.5)
55–64	Not very	Very high	Don't know	None	Some	0.1079 (9.3)
25–34	Not very	Moderately low	Don't know	None	None	0.6950 (1.4)
16–24	Fairly	Moderately low	Yes	School	Some	4.5
25–34	Very	Very high	Yes	Other	None	28.9

[a] The logistic regression results provide the odds to be active for a person possessing these six characteristics. Where the defined person is more likely to be sedentary than active (i.e. odds <1), then the reciprocal of the odds (i.e. odds to be sedentary) are described in parentheses. The odds range from the most likely to be sedentary at the top of the table down to the most likely to be active in equal intervals on a linear scale of the beta coefficients (β), where $\beta = \ln(\text{odds})$.

Discussion

The proportion of this sample of English adults that was deemed to participate in a sufficient amount of physical activity for health benefits was 26.5%. The remaining 73.5% participated in less activity than is deemed to be beneficial to health (Fentem *et al.*, 1994; Pate *et al.*, 1995). This includes 20.5% of the sample who lead a sedentary lifestyle. These findings are similar to those from previous studies in the UK (see Table 3) and USA (24% sedentary; Crespo *et al.*, 1999), and the small percentage of people involved in sufficient physical activity emphasizes the need for effective health promotion campaigns.

The results of the logistic regression model provided additional information on the physical activity of English adults by combining several independent variables and by using these to quantify their likelihood to participate in physical activity. The six variables included one 'physical' (age), one 'social' (education) and four 'self-perception' (active, adequate exercise, lifestyle problems and motivation).

The results of our analysis demonstrate that age is the strongest indicator of those individuals who are likely to be active. Participation is greatest between 25 and 44 years of age and decreases rapidly with increasing age. Life-changes towards a more independent lifestyle (Sallis and Hovell, 1990) have been hypothesized to explain the young (16–24 years) being less active than 25- to 44-year-olds, whose lives tend to be more structured. The Allied Dunbar National Fitness Survey (Activity and Health Research, 1992) proposed two theories for the decline in physical activity participation after the age of 44. First, the ageing process itself can be a deterrent for physical activity, although on its own this is extremely difficult to substantiate given the large differences in the rates of age-related changes across the population (McArdle *et al.*, 1996). The rate of ageing has been proposed to be causally linked to inactivity (Wood, 1992), which is supported by a deterioration of body functions in response to physical inactivity (Bortz, 1984). Secondly, the opportunities to participate in vigorous physical activity have increased dramatically with an increase in leisure facilities and sports provision. This is supported by the Leisure Industries Research Centre in the UK (Gratton *et al.*, 1999), who link an increase in consumer spending on active sport since 1993 to: an increase in the provision of health and fitness clubs, changes in self-perceptions towards a greater preference to engage in physical activity, and a rise in the number of 15- to 24-year-olds who have a tendency to be active. It is possible that the older generation has had little experience of vigorous physical activity other than through walking, physical tasks and occupations, but with future generations this may change.

With increasing age there may also be changes in self-perceptions, although the interaction of age with the self-perception variables retained in the analysis has not been explored. Nevertheless, as perceptions of competence in physical activity have been linked to high efficacy and thus to intrinsic motivation for participation in exercise (McAuley *et al.*, 1991), there may be a reduction in such competence associated with ageing that would in itself result in reduced motivation. Self-perceptions made a significant independent contribution to the model. Individual perceptions of physical activity patterns were found to be accurate where assessments were made of one's own participation, particularly whether this was adequate to promote health. Those who perceive themselves to be in good health are more likely to exercise (Sallis and Hovell, 1990).

Similarly, perceptions of possessing higher motivation and possessing no barriers to participation through lifestyle problems from health were found to increase the likelihood of participation. Participation has been related to those who recognize the benefits of exercise, and to a belief in one's control over health outcomes (Dishman, 1982). An understanding of the potential for improved health may explain why a high risk of heart disease has been proposed to be a strong determinant of participation in exercise, although Dishman *et al.* (1985) reported contradictory evidence that men at risk of coronary heart disease are less likely to participate. Education and the development of positive self-perceptions would appear, therefore, to be a useful tool to increase participation, but this may only be a means of stimulating initial participation rather than promoting better adherence. Further research is needed to explore whether intervention strategies to increase participation can also increase ongoing participation in exercise.

The results of our analysis suggest that individual perceptions of participation could be useful to identify who should be targeted in future health promotion campaigns. Identifying who to target based on a combination of several variables may be beneficial, whereas targeting based on single variables in isolation may result in key persons being omitted from health promotion activities. For example, motivation as an isolated factor has not been shown to relate to compliance with exercise programmes (Wankel *et al.*, 1985).

One 'social' variable was found to influence participation in physical activity. The greater the education of an individual, the more likely it was he or she would participate in sufficient physical activity. This was also identified in a study in the USA (Crespo *et al.*, 1999). In addition, as the characteristics of self-motivation and exercise behaviour skills have been linked with an increased likelihood of the more educated to be active (Dishman *et al.*, 1985), then educational achievement as a predictor of participation is inherently linked to self-perceptions and knowledge. The better educated might understand and value the benefits of exercise (Crespo *et al.*, 1999) and other lifestyle behaviours affecting health, such as diet, which may lead to greater participation in physical activity. Another factor associated with educational attainment is environmental barriers. Environmental features and related perceptions of the environment have been reported to be barriers to exercise (Barnes *et al.*, 1994). The factors described within this include family and social support, physical environment and availability of facilities. If less education is related to socio-economic status, then it is likely that the less educated will perceive more constraints to participation. In the USA, it was reported that only education made a significant contribution to a logistic regression model in women of different ethnic groups (Ransdell and Wells, 1998). Four possible mechanisms were described for lower participation with less education or lower social class, although the authors did emphasize that further research is required. The four mechanisms were that these individuals could be less informed, have less disposable income, work longer and more inflexible hours, and participate in more physically demanding jobs, which has previously been shown to be inversely related to physical activity participation (White *et al.*, 1987).

Other variables that were not included were hypothesized to have a small effect, or to have an effect catered for by the variables retained in the analysis. For instance, the inclusion of education may have excluded socio-economic group and social class. Education and socio-economic group have been considered as part of a multi-factorial measure of social class (Crespo *et al.*, 1999), but the probable strong relationship between these factors may clarify why education alone has been retained in the model. It is also beneficial to include education as an indicator because it is widely used, considered a stable measure for adults and is often a precursor to income and occupational status (Crespo *et al.*, 1999).

The sex of the individual also made no significant contribution to our model. Although there is evidence to suggest that participation patterns do differ between the sexes (Sallis *et al.*, 1985; Stephens *et al.*, 1985), there is little evidence that sex is a predictor of participation. Its absence from our model could suggest that the physical differences between the sexes and sociological issues of gender affecting individuals' participation in physical activity can be explained by the combined effect of the variables retained in the model. For example, as almost one-quarter of women are family or home keepers (Central Statistical Office, 1990 – taken from the same year as the ADNFS and the definition used then), who perceive their health to be significantly worse than those with other occupations (Stonks *et al.*, 1997), then a difference between the sexes may be accounted for by the self-perception of 'lifestyle problems' variable retained in the analysis. Alternatively, these differences could simply exist within the irregularly active group, which was omitted from the analysis. Although this sample is considered to provide a good estimate of the physical activity profile of the English population (Activity and Health Research, 1992), any traits of minority groups are obscured by logistic regression and many other statistical techniques simply owing to the problems of small and uneven sample proportions that increase the error and uncertainty in the results (Menard, 1995). Therefore, identifying any valid ethnic influence on physical activity participation is difficult and further confounded by the proportionally large difference between the non-white British population of 5.9% (Office of Public Censuses and Surveys, 1991) compared with 3.6% in this sample.

The resultant model from this study presents a powerful and accurate (81% correct) means to provide odds for selected individuals to participate in physical activity based on specific characteristics. The model spans a surprisingly large range in propensity to be sufficiently active, supported by the high explanatory power of the model (Hosmer and Lemeshow goodness-of-fit, $P = 0.88$). The odds to be active range from 0.0026 (i.e. the reciprocal of which provides those who are 385 times more likely to be sedentary than active) to 29 (i.e. those who are 29 times more likely to be active than sedentary). These ranges suggest that the tendency of the sedentary group to remain sedentary is much greater than that for the active group to remain active. This is supported by the ability of the model to correctly predict 89% of the sedentary group versus 72% of the active group. The results can be used to identify people that should and could be targeted in future health promotion campaigns. Efficient and economical health promotion may thus warrant targeting sedentary individuals with high odds to be active (e.g. from 0.6950 to 1), as they may more easily be encouraged to change and thus participate in adequate physical activity to promote health. In addition, targeting individuals who are more likely to be active than sedentary but with low odds (e.g. odds from 1 to 4.5) may be beneficial so that they remain active. The Transtheoretical Model of behaviour change (see, for example, Marcus and Simkin, 1994) may support targeting individuals with these odds to be active from 0.6950 to 4.5, which encompass one-fifth of the linearized range of odds as presented in Table 5. As this is the fourth of the five stages in Table 5, potentially this can be aligned with the 'action' stage of the Transtheoretical Model. Hence, changing the behaviour of this group to possess odds greater than 4.5 would put them in the top one-fifth of the odds range, potentially classifying them in the 'maintenance' stage of the Transtheoretical Model. Exploring these stages, and owing to the self-perception variables retained in the model, linking them with the self-efficacy component of the model may provide a focus for future work on changing physical activity participation behaviour. In addition, in targeting these individuals successfully in health promotion interventions, consideration should also be made of other factors, such as the appropriateness of specific intervention strategies for different points in the odds continuum and in light of other barriers to participation. A variety of barriers to participation and intervention strategies available to the practitioner have previously been highlighted (Barnes *et al.*, 1994; Dishman and Buckworth, 1996).

Future research should include only variables that have an *a priori* rationale for their effect on activity participation. Consideration should be given to: (1) confirming how the combined effect of these variables influences participation in physical activity; (2) utilizing this information in the development of subsequent large-scale analyses on physical activity; (3) investigating the irregularly active group potentially using a multinomial logistic regression (SPSS, 1999); and (4) assessing the effectiveness of health promotion strategies implemented to promote appropriate involvement in physical activity. Attention must also be given to the type of physical activity in which people engage, as there is evidence to suggest that predictors of participation differ according to the nature of activity or exercise (Sallis and Hovell, 1990). This work will test the validity of this model and approach, and provide more information on the effect of the English culture on physical activity participation.

Acknowledgements

We wish to thank the UK Sports Council and Health Education Authority for authorizing access to the Allied Dunbar National Fitness Survey and The Data Archive, Essex, England for supplying the database.

References

American College of Sports Medicine (1995). *ACSM's Guidelines for Exercise Testing and Prescription.* Baltimore, MD: Williams & Wilkins.

American College of Sports Medicine (1998). The recommended quantity and quality of exercise for developing and maintaining cardiorespiratory and muscular fitness, and flexibility in healthy adults. *Medicine and Science in Sports and Exercise*, 30, 975–991.

Activity and Health Research (1992). *Allied Dunbar National Fitness Survey: Main Findings.* London: Sports Council and Health Education Authority.

Activity and Health Research (1995). *Allied Dunbar National Fitness Survey: 1990 Computer File.* Colchester: The Data Archive.

Barnes, E., Woodhouse, D. and Bunton, R. (1994). *Sport, Activity and Health.* Newcastle: Northern Region of the Sports Council and Northern Regional Health Authority.

Bortz, W.M. (1984). The disuse syndrome. *Western Journal of Medicine*, 151, 691–694.

Centers for Disease Control and Prevention (1995). Prevalence of recommended levels of physical activity among women – behavioral risk factor surveillance system (1992). *Morbidity and Mortality Weekly Report*, 44, 105–113.

Central Statistical Office (1990). *Social Trends 20: 1990 Edition.* London: HMSO.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioural Sciences.* London: Lawrence Erlbaum.

Crespo, C.J., Ainsworth, B.E., Keteyian, S.J., Heath, G.W. and Smit, E. (1999). Prevalence of physical inactivity and its relation to social class in US adults: Results from the Third National Health and Nutrition Examination Survey, 1988–1994. *Medicine and Science in Sports and Exercise*, 31, 1821–1827.

Dishman, R.K. (1982). Compliance/adherence in health-related exercise. *Health Psychology*, 1, 237–267.

Dishman, R.K. and Buckworth, J. (1996). Increasing physical activity: A quantitative synthesis. *Medicine and Science in Sports and Exercise*, 28, 706–719.

Dishman, R.K., Sallis, J.F. and Orenstein, D.R. (1985). The determinants of physical activity and exercise. *Public Health Reports*, 100, 158–171.

Fentem, P.H., Collins, M.F., Tuxworth, W., Walker, A., Hoinville, E., Cooke, C., Harries, U.J. and Rayson, M. (1994). *Allied Dunbar National Fitness Survey: Technical Report.* London: Sports Council and Health Education Authority.

Gratton, C., Shibli, S. and Kokolakakis, T. (1999). *Leisure Forecasts 1999–2003.* Sheffield: Leisure Industries Research Centre.

Hahn, R.A., Teutsch, S.M., Rothenberg, R.B. and Marks, J.S. (1990). Excess deaths from nine chronic diseases in the United States, 1986. *Journal of the American Medical Association*, 264, 2654–2659.

Health Education Authority (1995). *Health Update 5: Physical Activity.* London: HEA.

Luepker, R.V., Johnson, S.B., Breslow, L., Chobanian, A.V., Davis, C.E., Duling, B.R., Kumanyika, S., Lauer, R.M., Lawson, P., McBride, P.E., Oparil, S., Prineas, R.J. and Washington, R.L. (1996). Physical activity and cardiovascular health. *Journal of the American Medical Association*, 276, 241–246.

Marcus, B.H. and Simkin, L.R. (1994). The transtheoretical model: Applications to exercise behavior. *Medicine and Science in Sports and Exercise*, 26, 1400–1404.

McArdle, W.D., Katch, K.I. and Katch, V.L. (1996). *Exercise Physiology: Energy, Nutrition and Human Performance.* Philadelphia, PA: Williams & Wilkins.

McAuley, E., Wraith, S. and Duncan T.E. (1991). Self-efficacy, perceptions of success and intrinsic motivation for exercise. *Journal of Applied Psychology*, 21, 139–155.

Menard, S. (1995). *Applied Logistic Regression Analysis.* London: Sage.

Munro, B.H. (2001). *Statistical Methods for Health Care Research.* New York: Lippincott.

Office of Public Censuses and Surveys (1980). *Classification of Occupations.* London: HMSO.

Office of Public Censuses and Surveys (1991). *Census 1991.* London: HMSO.

Pate, R.R., Pratt, M., Blair, S.N., Haskell, W.L., Macera, C.A., Bouchard, C., Buchner, D., Ettinger, W., Heath, G.W., King, A.C., Kriska, A., Eon, A.S., Marcus, B.H., Morris, J., Paffenbarger, R.S., Patrick, K., Pollock, M.L., Rippe, J.M., Sallis, J. and Wilmore, J.H. (1995). Physical activity and public health: A recommendation from the CDCP and the ACSM. *Journal of the American Medical Association*, 273, 402–407.

Prior, G. (1999). Physical activity. In *Health Survey for England: Cardiovascular Disease '98* (edited by B. Erens and P. Primatesta), pp. 181–220. London: The Stationery Office.

Ransdell, L.B. and Wells, C.L. (1998). Physical activity in urban white, African-American, and Mexican-American women. *Medicine and Science in Sports and Exercise*, 30, 1608–1615.

Riddoch, C.J. and Boreham, C.A.G. (1995). The health-related physical activity of children. *Sports Medicine*, 19, 86–102.

Sallis, J.F. and Hovell, M.F. (1990). Determinants of exercise behaviour. *Exercise and Sport Sciences Reviews*, 18, 307–330.

Sallis, J.F., Haskell, W.L., Wood, P.D., Fortmann, S.P., Rogers, T., Blair, S.N. and Paffenbarger, R.S. (1985). Physical activity assessment methodology in the five-city project. *American Journal of Epidemiology*, 121, 91–106.

SPSS (1999). *SPSS for Windows: Release 10.0.0.* Chicago, IL: SPSS Inc.

Stephens, T., Jacobs, D.R. and White, C.C. (1985). A descriptive epidemiology of leisure-time physical activity. *Public Health Report*, 100, 147–158.

Stonks, K., Van de Mheen, H., Van den Bos, J. and Mackenbach, J.P. (1997). The interrelationship between income, health and employment status. *International Journal of Epidemiology*, 26, 592–600.

Walker, A. and Hoinville, E. (1995). *Physical Activity Matters Across England – Variation by Region.* London: Health Education Authority.

Wankel, L.M., Yardley, J.K. and Graham, J. (1985). The effects of motivational interventions upon the exercise adherence of high and low self-motivated adults. *Canadian Journal of Applied Sport Sciences*, 10, 147–156.

White, C.C., Powell, K.E., Hogelin, G.C., Gentry, E.M. and Forman, M.R. (1987). The behavioral risk factor surveys: IV. The descriptive epidemiology of exercise. *American Journal of Preventive Medicine*, 3, 304–310.

Wood, P.D. (1992). Ageing: Don't take it sitting down. In *Medical and Health Annual* (edited by E. Bernstein), pp. 135–149. Chicago, IL: Encyclopaedia Britannica.

Appendix 13

Mullineaux, D.R. and Bartlett, R.M. (1997). Research methods and statistics. In *Biomechanical Analysis of Movement in Sport and Exercise* (edited by R.M. Bartlett), pp. 81-104. Leeds: BASES.

7 RESEARCH METHODS AND STATISTICS

7.1 Introduction

Advances in sport and exercise biomechanics depend on many factors; these are technical, personnel and research-based. For advances to occur, developments are needed not only in equipment, but also in the experience, knowledge and expertise of personnel, so that researchers propose, conduct and report research correctly and thoroughly. The aim of this chapter is to highlight research design and statistical issues that should be considered in all research. An understanding of these issues should assist in the production of quality research that will advance sport and exercise biomechanics. This chapter provides an overview of the important factors to be considered; further details on specific aspects (particularly the use of discrete tests) should be sought from more detailed texts, such as those listed in section 7.8.

7.2 Types of research

Research is often categorised as basic or applied; within each of these there are many ways to classify research. Research is generally positioned on a continuum from descriptive (where events are observed and described) to experimental (events are manipulated and effects analysed). Typically the results are described as being either qualitative or quantitative. Scientific research should be systematic, empirical, reductive, logical and replicable (Tuckman, 1978), as described in Table 7.1.

Feature	Characteristics
Systematic	Identifying and labelling variables, and designing research to test associations between them.
Empirical	Collecting data to allow evaluation of the problem and hypotheses.
Reductive	Synthesising data to establish general associations or relationships.
Logical	Examining the procedures used to allow researchers to evaluate conclusions.
Replicable	Recording all important information to allow others to test the findings or to build on the results.

Table 7.1. Characteristics of research.

In sport and exercise biomechanics, empirical research is generally quantitative; this research forms the basis of this chapter. The research designs commonly used in biomechanics are experimental, quasi-experimental and correlational. Much laboratory-based research in sport and exercise biomechanics falls into the first two of these categories, and often involves looking at the effects of an experimental treatment on differences in variables between groups. Much field based research in sports biomechanics is not truly experimental, as no control is exercised through the use of independent or categorical variables. Such research is often more correctly described as correlational research.

7.2.1 Experimental research

Experimental research involves the experimenter manipulating treatments - the experimental or independent variable - to monitor possible effects on selected dependent variables. The experimenter tries to control all factors except the dependent variable.

7.2.2 Quasi-experimental research

Quasi-experimental research includes designs in which selection and assignment of subjects to treatments are not random. This principally includes non-probability sampling, and the use of independent variables that cannot be manipulated and are categorical (e.g. age, sex, skill level) rather than under the experimenter's control. An example, incorporating a categorical variable, is an *ex post facto* design studying groups of skilled and unskilled subjects to establish variables that distinguish between the groups.

7.2.3 Correlational research

Correlational research is a branch of descriptive research that explores relationships that exist among variables. The basic distinction between this and experimental research is that the experimenter neither manipulates independent variables nor administers experimental treatments. Such research is designed to collect data on several variables from the same subjects, to analyse the relationships that exist among these variables, and, possibly, to predict values of a dependent variable. This research is often most valuable at the descriptive stage of the research process. The data are usually analysed by correlation-regression statistics.

7.3 Planning research

Research may be considered as four primary processes: research design, conduct of the investigation, statistical analysis, and reporting of the research (Table 7.2).

Research process	Characteristics
Research design	Hypotheses formulation, group selection, experimental design, and control of extraneous variables.
Conduct of investigation	Control for experimental error, establishment of uncertainties (errors) in experimental data, objectivity and reliability checks, data processing, and error propagation.
Statistical analysis	Justifying and implementing statistics to describe (descriptive statistics), support (effect size statistics) and test hypotheses (inferential statistics).
Reporting of studies	Information needed for interpretation and replicability.

Table 7.2. The four processes of research.

The initial stages of research involve a broad consideration of all the components of the study. The four processes of research (Table 7.2) are considered together, with the aim of establishing whether the results expected would provide an answer to the research question, using the best methods available to the researcher (Figure 7.1). If this is achieved, the study can be implemented; otherwise modifications should be made within any of the four processes.

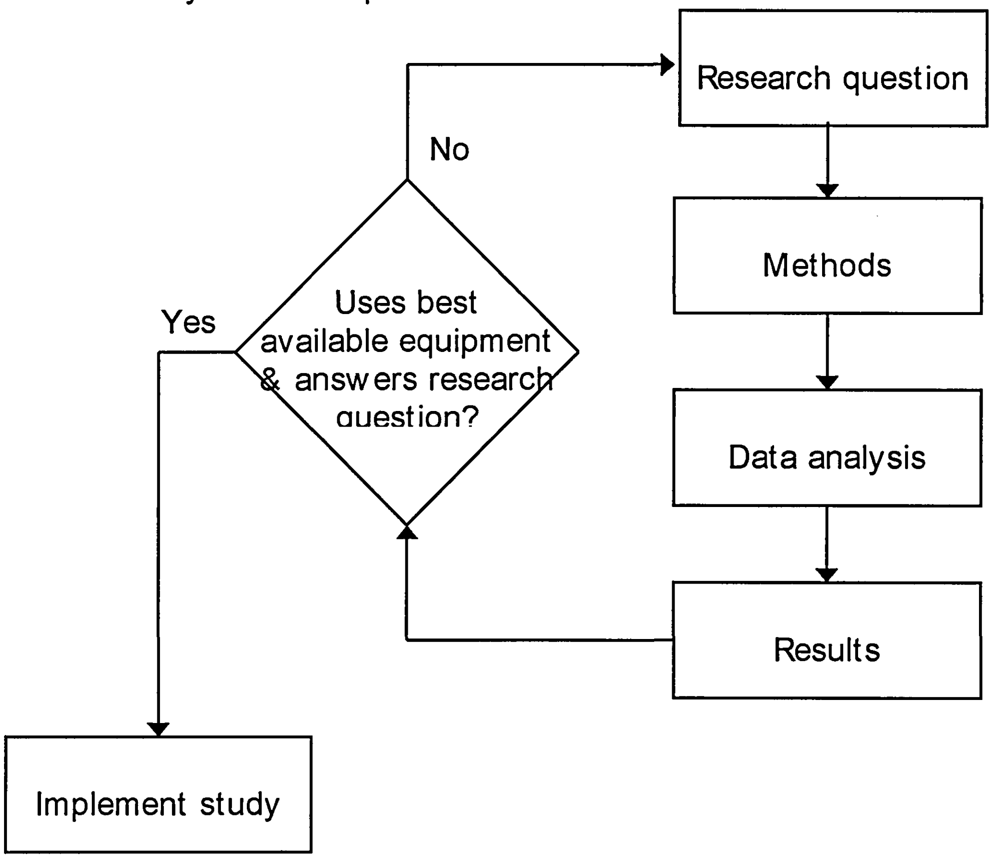


Figure 7.1. Planning research.

The planning initially involves the identification of a research question that is both worthy of attention and should contribute to the body of knowledge. Through a subsequent literature review, an appropriate research design (method) should emerge that is realistic and ethical. The design should incorporate all considerations essential for suitable and accurate descriptive, inferential and effect size statistical analysis.

Often inferential statistics are used as the penultimate component of the research process, before reporting. Such statistics have a single purpose, to identify the probability of obtaining the results if chance caused them (Carver, 1978). Inferential statistics show nothing about the odds of occurrence, generality, reliability, validity, or importance of the findings (Table 7.3). From the information that is provided by testing the null hypothesis (H_0), inferential statistics may be used to support results from descriptive and effect size statistics. The implementation of an appropriate research design, however, is of greater importance than an overemphasis on statistical issues. As the method is the most important part of the research process (O'Brien and Israel, 1987) it must be carefully planned and implemented; a good research design will, in turn, aid in the appropriate choice of any statistical analysis.

The following will highlight mechanisms for developing a good research design suitable for sport and exercise biomechanics, that should inherently result, where necessary, in appropriate descriptive, inferential and effect size statistics.

Information supposedly provided	Actually provided	Reason for or alternative method to obtain information
Probability of results	Yes	Probability of getting the results when chance is assumed to have caused them is provided.
Odds of occurrence	No	The probability that the results were due to chance cannot be calculated.
Generalisation	No	Inference or generalisation is not possible.
Reliability	No	Should be provided within the research design and by replication of the study.
Validity	No	Rigorous theorising, careful planning, and replication can assist this.
Importance	No	Discussion of the results in relation to research question, previous literature, theory, and limitations can highlight the importance of the results.

Table 7.3. Information provided by inferential statistics (partly derived from Carver, 1978).

The many considerations in planning research include sampling, sample size, trial size, statistical significance level and power desired. The following subsections focus on nine important issues in the broad topic of research design.

7.3.1 Hypothesis formulation and testing

The normal progression from the defined research question is to formulate testable research hypotheses. The hypotheses, developed from theories highlighted in previous research, should be formulated to be what the researcher logically and justifiably expects. These hypotheses will then provide the basis for the subsequent study, and the aim is to accept or reject them.

The use of previous research is vital to increase the probability, which is never determinable, of rejecting or accepting the hypotheses. In using the literature it is beneficial to access research that is unpublished and that had results that were not statistically significant, possibly by writing to key authors. This is important because of a tendency in the literature to report only statistically significant findings; this can introduce a bias into research. Both published and unpublished research can be of high and low quality; it is up to the researcher to justify the quality of such work.

The testing of hypotheses by the scientific method requires that the research hypothesis (H_1) be evaluated first, mainly by inspection and discussion of descriptive and effect size statistics. Alternative hypotheses (including the null or chance hypothesis) should then be evaluated; directional, or research, hypotheses are usually altered to the null hypotheses that are statistically tested. Simply disregarding the theory-derived research hypothesis for the convenient null hypothesis, necessary for inferential statistical analysis, can corrupt the scientific method and can hinder correct analysis. Use of *post hoc* and unsuitable theoretical propositions (particularly if H_0 is rejected at a given α level and H_1 is accepted) is problematic. It is not correct to assume that the probability of H_1 is 1-α; rejection of H_0 does not prove H_1 (see Dracup, 1995, for a lucid account).

Inferential statistics only test the null hypothesis, and do not provide any support for an alternative hypotheses. Correct research requires that the research hypothesis and data be discussed first, possibly with the aid of descriptive and effect size statistics. Other hypotheses should then be assessed, including the testing of the null hypothesis, using inferential statistics. This ensures that the data, and not the inferential statistics, drives the methods, analysis, and interpretation of the research.

7.3.2 *Research design*

Research design is often expressed by its paradigm (e.g. experimental research), links between groups (e.g. within design), and the number and timings of treatments and measurements. A clear design is important and assists readers in understanding and interpreting research. A good research design should develop from theory, with due consideration of constraints imposed by the research question, previous literature, resources, and ethics. The research design is at least as important as any subsequent statistical analysis; if the research design and methods are flawed, the value of the results is greatly reduced.

7.3.3 *Sampling*

Research questions posed in sport and exercise biomechanics often require subjects with specific characteristics. For example, a study of the techniques of elite hammer throwers will not require random sampling from the population, but purposive sampling of elite hammer throwers. To help readers to make their own inferences, it is important that the sampling used is described.

The choice of statistical analysis is affected by the type of sampling and the often small population of interest in biomechanical studies. The use of some statistics in such instances may not be appropriate, as both the sampling and small sample size may violate assumptions of the statistical tests. For instance, the mathematics of the normal probability curve underpins the tabulated sampling distribution of many statistical tests. Sampling sufficient numbers from a population with random or equivalent techniques will, by probability theory, result in a normal sampling distribution, allowing the use of these statistical tests. In biomechanical research, however, the use of non-random sampling and reliance on small numbers probably results in sampling distributions that are not normal. This can lead to misleading descriptive statistics and can cause severe problems in the control of type I errors in certain statistical tests.

7.3.4 Sample size

An early consideration in the design of any research is the determination of the required sample size (n). The probability of a biased sample increases as the sample size decreases; this increases the sampling error - the error in the estimate of a population parameter from a sample statistic. Often the sample size is unambiguously dictated by availability. The required numbers should, however, be determined *a priori* based on considerations complementary to the appropriate use of statistics. Several methods exist to calculate the appropriate sample size including statistical tables, characteristics of the sample, or requirements of the statistical test.

If the results of a research study, based on a randomly selected sample, are to be generalised to the population from which that sample was drawn, then statistical tables can be used to determine the sample size, n, for a given population size, N. For such generalisations to hold, the sample statistic should be a reliable, and unbiased, predictor of the population parameter. For the researcher to be confident that this is the case, the sample size should be in accord with tables such as Table 7.4. These are very easy to use, but are limited, as identifying the population size may be difficult in the first place. However, generalisations based on smaller sample sizes must be viewed with considerable caution. Fortunately, as discussed below and in later sections, when the findings of a research study are not to be generalised to a larger population, then the constraints on sample size can be considerably relaxed.

N	n	N	n	N	n
10	10	500	217	7000	364
15	14	650	242	10000	370
20	19	800	260	15000	375
25	24	1000	278	20000	377
50	44	1300	297	30000	379
100	80	1600	310	40000	380
200	132	2000	322	50000	381
300	169	3000	341	75000	382
400	196	4000	351	1000000	384

Table 7.4. Sample size (n) required for a given population size (N) to provide 90% confidence that the variation between the sample statistics and population parameters will be less than 5% (adapted from Krejcie and Margan, 1970).

Previous research can be used to determine the size of the effect between groups. This is important, as the power of a test is affected by the effect size as well as by the sample size at a given α (significance) level for a specific test. If a specific test power is desired, and if the effect size is known, the required sample size can be determined from published tables and equations (e.g. Cohen, 1988). Power, effect size, the α level, and the assumptions of statistical tests will be discussed in more detail below.

The requirements of specific statistical tests also affect the sample size. For example, in regression analysis the ratio of the sample size (n) to the number of predictor variables (p) is important. For small ratios of $n{:}p$ high correlations can occur by chance and the generality of the regression equation will be seriously limited; for random data, the correlation coefficient is 1 for $n = p+1$. At the opposite extreme, for a large enough n almost any multiple correlation will be statistically significant. Vincent (1995) cited a minimum ratio of 5:1 with an ideal value of around 20:1, and an even greater ratio of 40:1 for stepwise multiple regression. Different authors provide somewhat different, usually more conservative, figures (e.g. Harris, 1985; Howell, 1992). A good rule of thumb is never to have $n{:}p$ less than 10:1, and ideally between 20:1 and 40:1, with the top end of the range for stepwise multiple regression.

In testing differences between groups, a sample size of no less than 30 has been recommended per group (Baumgartner and Strong, 1994). Such sample sizes increase the probability of a normal sampling distribution on the dependent variable, an assumption on which statistical tables, such as those of the t and F-statistic, are based. However, the size of the samples, providing they are equal, is regarded by some authorities as something of a side issue (e.g. Howell, 1992). An important factor here is to seek to assess any effect of a small sample size on the probability (P) value, for example through power analysis (e.g. Cohen, 1988).

7.3.5 Trial size

Analysis of a single trial in sport and exercise biomechanics is both common and fraught with problems. These problems can often arise when one trial is assumed to be representative of a subject's technique, and when it is analysed along with the trials of other subjects. Individual performances should only be grouped after verification of similarities as, from probability theory, similarities are likely to prove to be exceptions (Bates *et al.*, 1992). Grouping may result in the characteristics of individual techniques being 'washed-out' by combining the data. This consideration lends strong support to the use of within-subject designs (Bates *et al.*, 1992).

For within-subject designs, the statistical power is influenced by the number of trials. For 90% power to be achieved, it has been recommended that 10, 5 and 3 trials should be used for sample sizes of 5, 10 and 20 respectively (Bates *et al.*, 1992). This use of repeated measures is supported by Carver (1978) who pointed out that it is desirable to build replication into designs, even though this can involve a more complex analysis.

7.3.6 Statistical significance level

All too often in biomechanics, and other research, an arbitrary statistical significance level (alpha, α) is selected with no rationale provided and with no consideration of power and type II errors. This raises several important issues relating to type I errors (reject the null hypothesis when it should have been accepted) and type II errors (accept the null hypothesis when it should have been rejected). The type I and II errors are on a balanced continuum (i.e. as the chance of making one increases, the chance of the other will decrease). The chances of making either type of error should be controlled by setting a justifiable value for α.

The chosen value of α should be based on how the cost of a wrong decision can be calculated (Morrison and Henkel, 1969); this can often be difficult. Generally, the level set depends on whether type I or type II errors are more likely and important in your research. Setting α too low is likely to lead to a low power test (unless subject numbers are very high) and might discourage further research by not rejecting a false null hypothesis. Setting α too high might send researchers on a wild goose chase. However, higher values may be more useful when developing theories, for example, than when testing theories. The value of 0.05 often chosen in the scientific literature needs careful consideration in relation to the power of the resultant test; alternative α levels can be used. Regardless of whether $\alpha = 0.05$, justification of the selected statistical significance level should be given (Franks and Huck, 1986) and P values should be reported, so that readers can make their own interpretation.

7.3.7 Control of type I error rates

Type I error rates can be increased by the use of inappropriate statistical analysis. A principal control for type I errors is to treat the set of tests as one statistical test at the chosen α level. The three primary type I error rates are as follows (e.g. Howell, 1992).

- Error rate per comparison. This (PC) is the probability of making a type I error on any single comparison (α).
- Error rate per experiment. This (PE) is the number of type I errors that would be expected in an experiment if the null hypothesis is true; note that this is a frequency and not a probability. If the comparisons (c) are of the same type (e.g. all t-tests), then PE = $c\alpha$.
- Familywise error rate. This (FW) is the probability that a set of conclusions derived from a set of comparisons (e.g. of group means) will contain at least one type I error. For independent comparisons FW = $1 - (1-\alpha)^c$. It is for the researcher to decide on the familywise error rate that is acceptable for any set of comparisons, and to report the upper bound on FW accordingly.

The issue of control of type I error rates is of great importance whenever an experiment involves several comparisons. This frequently occurs in follow-up tests to ANOVA, where control of multiple comparisons is built into the *post hoc* test (e.g. Tukey and Scheffé tests).

Large type I error rates can also arise, for example, either when comparisons are made of many dependent variables between groups using multiple *t*-tests or when correlations are made between many variables. If, for example (these figures are based on Scheirman *et al.*, 1988), consider that α is set to the commonly used value of 0.05 per comparison and that 48 comparisons are made; if the comparisons are independent and H_0 is true, PE is 2.4 (i.e. probably two or three type I errors are made) and the probability of at least one type I error (FW) is 91%, both clearly unacceptable.

Possible solutions to reducing type I error rates are to:

- Make only a few comparisons planned in advance.
- Use an adjustment technique, such as the 'Bonferroni inequality', and set PE to α and the individual comparison value to α/c. This will control type I errors but increase the probability of type II errors and reduce the power of the test significantly.
- Use an alternative statistic, such as MANOVA.
- Accept a higher FW but report it so that the reader can judge.

7.3.8 Type II errors and power

The power of a test is the probability of rejecting a false null hypothesis; it depends on the overlap of the sampling distributions of the null hypothesis and the research hypothesis. The power is $1-\beta$, where β is the type II error rate. Several factors determine the power of an inferential statistical test to identify any statistical significance. Power is affected by the sample size, α level, statistical test, and effect size index. The research design should directly control the sample size, α level and the statistical test to be used. Sample selection and research design can also influence the effect size, but the effect is not quantifiable. It is essential that factors affecting power are appropriately controlled to reduce the occurrence of type I or type II errors. The effect size can either be estimated from previous research or set at a size for the differences considered to be important by the experimenter. Effect size can be used in equations (along with n and α) for appropriate tests, as in Cohen (1988), to estimate the power of the analysis.

As a rule of thumb, conventions of small, moderate and large effect sizes equating to <0.1, 0.3, and >0.5, respectively, were proposed by Cohen (1988). Different tests, and various authors (e.g. Cohen, 1988; Vincent, 1995) give different values for these effect sizes. For example, for $\alpha = 0.05$, using a two-tailed, one-sample *t*-test, the relations between effect size, power and n are described in Table 7.5 (Cohen, 1988).

It can be seen from Table 7.5 that effect size, power and the sample size can vary greatly. The main way for the researcher to increase power is to alter the sample size (as described earlier). It should be noted that, if n is made as large as possible, it is always possible to obtain statistical significance (Hays, 1963).

Power should always be quoted to support non-statistically significant findings (Bates *et al.*, 1992).

Effect size	Power for $n = 9$	n for power of 0.80
Large	0.67	13
Moderate	0.32	32
Small	<0.07	196

Table 7.5. Power, effect size and n for a two-tailed, one-sample t-test ($\alpha = 0.05$).

7.3.9 Choice of test

During the planning stage of research, the choice of test needs to be considered. This is particularly important for considerations of type I errors, power, and discussion of the results. This will be referred to in more detail in section 7.5.2.

7.4 Conduct of an investigation

When conducting an investigation, it is paramount that errors are controlled and that the effectiveness of the study is assessed. This assessment can be performed in several ways, including estimating errors and checking reliability. The following subsections will broadly describe methods both to control errors and to assess their impact. Consideration should always be given, during the planning and conduct of research, to ethical clearance, obtaining informed consent, and subject care; further reference to these issues will not be made in this chapter (but see, for example, appendix 3).

7.4.1 Control of errors

Many factors can introduce errors into the results of an experiment. These can arise from the research design, the apparatus, the experimenter, the subjects, and interactions between these. Methods can be incorporated into the research design to reduce these errors. Unfortunately, the variety of research means that no general rules exist for control of errors; however, there are general principles to consider. Examples of the sources and the control of error are described in Tables 7.6 and 7.7 and below. We also recommend reference to Baumgartner and Strong (1994) and other detailed texts (e.g. Thomas and Nelson, 1996).

Error source	Description
History	Additional activity over the period of the experiment.
Maturation	Subjects develop over the experiment period.
Testing	Subjects learn from pre- to post-tests.
Individuality	Subjects' responses to error sources are random.
Experimental mortality	Loss of participants in the experiment.
Hawthorne effect	Experimental group perform better as they know they are being tested.
John Henry effect	Control group tries harder to be better than the experimental group, and succeeds.
Interaction of testing	Subjects' responses are different from anything natural.
Experimental setting	Subjects' responses are altered from normal.
Multi-treatment interference	Pre-tests make post-tests unrepresentative.

Table 7.6. Sources of error though the subjects.

Error source	Description
Selection bias	Results 'not general' (e.g. from a convenience sample).
Assignment bias	Biased assignment of subjects to groups.
Rating or halo effect	Prior results influence results awarded by the experimenter.
S-researcher interaction	Conflict or rapport between the subjects (S) and the researchers.
Experimenter bias	Results are found to be, or reported, as expected.
Instrumentation	Quality, reliability, and validity of measuring devices.
Statistical regression	Tendency for extreme pre-test scores to be less extreme on post-tests.
Post hoc error	Interpretation incorrect.

Table 7.7. Sources of error through the experiment.

Methods to control these errors are many, and include:

- Trying to keep the testing as natural as possible.
- Minimising the information that needs to be provided to the subjects.
- Physically manipulating subjects' actions before and during test period.
- Matching subjects between groups to control for extraneous variables (or using a within-subject design).
- Counterbalancing for subjects receiving several treatments by varying the order of administration for different subjects.
- Introducing control groups (e.g. placebo) and using single or double blind designs.
- Ensuring experimenters are competent in using the equipment and the procedures involved.
- Calibrating equipment.
- Using the best available equipment.
- Using appropriate statistical analysis and methods to factor out extraneous variables (e.g. ANCOVA).

There are many methods for controlling error. Much experience can be gained through a thorough literature review. Good, comprehensive research, and attention to detail during the conduct of the research, will greatly help to reduce error.

7.4.2 Estimation and propagation of experimental errors

If the errors in measurements (which are estimates of the true values) were known, then these could be removed to find the true values. Unfortunately, except for calibration, which can remove some systematic error sources, this is not possible. What is important is to provide a reasonable estimate of the uncertainty (or error) in the measurements and to assess how these propagate in any calculations based on those measurements. Fortunately, this is now becoming far more common in the sport and exercise biomechanics literature. More in-depth information is contained in chapter 8; the following provides a few examples of error estimation and propagation.

- The estimate of error can be obtained for example from the repeated digitising of analysed sequences. Then the estimate of error can be based on $MS_E^{½}$ (where MS_E is the error variance obtained from ANOVA); the estimate of measurement error can, for example, be taken as 1.96 $MS_E^{½}$ to provide 95% confidence limits. The estimate of measurement error can also be made in other ways (see, for example Taylor, 1982). Unbiased estimates of error are preferable.

- For all smoothing and differentiating routines, Lanshammer (1980) has provided formulae that can be used to estimate the maximum attainable precision in the values of the derivatives. The error (noise) removal capabilities of all filtering, smoothing, and differentiating routines used should be assessed using this formula: $\sigma_k^2 \geq (\sigma^2 T\omega_s^{2k+1})/(\pi(2k+1))$. In this equation, σ_k is the standard deviation of the normally distributed, random noise in the estimated k^{th} derivative, σ is the standard deviation of the noise in the measured displacement data, T is the sampling interval (the inverse of the sampling rate in Hz). The term ω_s is the 'bandwidth' of the signal (the range of frequencies that it encompassed) in $rad{\cdot}s^{-1}$. This would be 10π $rad{\cdot}s^{-1}$ for a movement with a maximum signal frequency of 5 Hz. If such a signal was sampled at 100 Hz ($T = 1/100$s) - this is twenty times the maximum frequency - the standard deviation of the noise in the filtered and differentiated acceleration data ($k = 2$) would be 140 times that of the noise in the raw displacement data. It would be even worse for lower sampling rates (see also chapter 8).

- Error propagation should be investigated for all calculations based on error contaminated data, such as calculations of joint moments (see chapter 8 for more details and examples) and segmental energy levels. The error propagation formulae based on the chain rule of differential calculus are well known and widely published (e.g. Taylor, 1982; Holman, 1989).

7.4.3 Reliability and validity

Two important characteristics of a test or measurement are validity and reliability. The former relates to the degree to which a test or instrument measures what it purports to; the latter relates to the consistency and dependability of the measures. It can be inherently difficult to assess validity, but a sound research design evolved from theory can minimise uncertainty. Validity can further be supported by assuring reliability, if this can be measured. In some instances, reliability cannot be assessed because of unique and unrepeatable testing; however, this rarely occurs in sport and exercise biomechanics.

Reliability is essential for validity. When the measurement depends on human intervention, such as manual digitising, two types of reliability are defined: intra-operator (often referred to simply as reliability) and inter-operator reliability (referred to as objectivity). These are calculated, for interval or ratio level data, either by the use of a 'coefficient of reliability' or by 'boundaries of agreement'.

The coefficient of reliability is the ratio of the true measurement variance to the observed measurement variance. The former is obtained by subtracting the error variance from the observed measurement variance. Possible ways of obtaining this include interclass (e.g. Pearson product moment coefficient) or intraclass correlations. Interclass correlation is totally inappropriate as it requires independent variables, is limited to two sets of measurements, and does not detect systematic errors (Thomas and Nelson, 1996).

The analysis of variance (ANOVA) has often been accepted as an appropriate way of calculating the reliability (intraclass correlation) coefficient. The reliability coefficient, $R = (MS_B - MS_W)/MS_B$, where the variances are partitioned into components: trial or operator ($MS_W = (SS_T + SS_E)/(df_T + df_E)$, using trial (T) and error (E) variances); repeated measures or frames (MS_B). Interpretation of the reliability coefficient as indicating whether reliability is good or otherwise varies between authors. Fleiss (1986) described reliability as poor ($R < 0.40$), fair to good ($0.40 < R < 0.75$), and excellent ($R > 0.75$). A more conservative interpretation was provided by Vincent (1995) with acceptable but questionable ($0.70 < R < 0.80$), moderate ($0.80 < R < 0.90$), and high ($R > 0.90$). For ordinal level data, the equation $R = (MS_B - MS_E)/MS_B$ can be used for assessing reliability (Vincent, 1995).

Unfortunately, far too much emphasis has been placed on the calculation of the reliability coefficient. It should be the least-biased estimate of the population ratio of the true measurement variance to the observed measurement variance. The equations most commonly used to calculate this coefficient, given above, are not unbiased (see Winer *et al.*, 1991), and it should be noted that R tends to $-\infty$ as MS_B tends to zero, a somewhat undesirable property. Problems also arise from the reliability coefficient being severely affected by the variability within frames (effects); if this variance (MS_B) is large, it will give a larger reliability coefficient than for a smaller variance. More important by far than the reliability coefficient is the estimation of the error in the (repeated) measurement. This can be established from ANOVA, as discussed in the previous subsection, or from the method of Bland and Altman (1986).

Bland and Altman (1986) proposed that, as the intraclass correlation is an inferential statistical test, it should not be used to assess reliability. It does not identify the degree of compatibility or agreement between data sets (see also Ottenbacher and Stull, 1993; Mullineaux *et al.*, 1994). These authors propose that an alternative test should be used, that of 'limits of agreement', as this overcomes the difficulties mentioned. Such a test is more appropriate, as it is not an inferential statistical test and provides more meaningful information. Agreement for two sets of measurements is calculated by firstly subtracting the values in one set from those of the other set, and then calculating the mean (δ) and standard deviation (σ) of the differences. The boundaries of agreement are calculated as $\delta \pm 1.96\sigma$ (Bland and Altman, 1986) with 95% confidence. For more than two groups, a fixed boundary value can be calculated as $\pm 1.96\sqrt{(2MS_E)}$ with 95% confidence (BSI, 1979; Mullineaux *et al.*, 1994; Bland, 1995). This method is not, as sometimes suggested, limited to two sets of data, and it can be extended to any number of repeated measures or trials.

Both repeated measures ANOVA and the Bland and Altman method are acceptable ways of assessing reliability, providing that inappropriate conclusions are not drawn, for example, by exclusive reliance on the use of the intraclass correlation coefficient. The important factor is that an assessment is made that reliability, and subsequently validity, can be supported.

7.5 Data analysis

In analysing research data, three main statistical methods are available: descriptive, inferential, and effect size. The use of each is largely dependent on the characteristics of the sample and the aim of the analysis. Generally, they should be used as follows.

- Descriptive statistics describe the characteristics of the sample. When the sample is not representative of the population then no other tests are needed.

- Inferential statistics provide the probability of obtaining the results if chance was assumed to have caused them. They do not indicate anything else about the data.

- Effect size statistics quantify the difference between groups, or the strength of a relationship.

The use of each of these can substantially ease the interpretation of analysed data. Unfortunately, inappropriate use can waste time and complicate an analysis, masking important information or, worse, misinforming the reader. Misuse may arise from poor research designs, inappropriate editorial prerogative, or from research that cannot justify inclusions of a particular test. The last of these can be avoided by always 'justifying' the use of particular statistical tests, thereby assuring that the assumptions of the test are sufficiently met (a check list of these is contained in Table 7.8). Scientific journal editors vary in their insistence on the use of inferential statistics; this has caused many problems and there is now a shift towards effect size becoming mandatory. Insistence on an appropriate statistical analysis would be a far better policy and would mark a positive advance for research in sport and exercise biomechanics. The effect of these problems was partly illustrated by Armstrong (1987), who noted that generalisation is often incorrect; inference should be performed with great care to compensate for this. Carefully observation of the data, possibly using descriptive statistics, may be more important in research than the often misleading use of inferential (and other) statistical tests (Armstrong, 1987).

In choosing a statistical test, it should be noted that some tests make assumptions about the population from which the sample was drawn (such as its normality). These assumptions may also influence the selection of subjects (e.g. sample size, random selection and assignment). Such tests, that involve estimation of one or more population parameters, are known as parametric tests; nonparametric tests do not rely on parameter estimation or distribution assumptions. If the assumptions of a specific parametric test are met, then data of ordinal, interval and ratio levels of measurement are appropriate for such tests (Howell, 1992; Safrit and Wood, 1995).

Parametric tests are robust to violations of many assumptions (e.g. Howell, 1992), and the use of parametric tests is generally preferable as they have advantages over non-parametric tests. These include: that they are more powerful (i.e. more likely to reject a false null hypothesis), and that, in tests of more that two groups, the location of the differences can be found statistically. For severe violations, however, nonparametric tests are more powerful and valid in that they reduce type I error rates. There are many examples in biomechanics where parametric statistics were used without the authors making the appropriate justifications. Examples can also be given where parametric statistics could have been justified and would have provided a more powerful test with implications for a correct analysis and better interpretation. Implications for analysis, from the violation of assumptions for inferential statistics, also apply to descriptive statistics.

7.5.1 Descriptive statistics

Descriptive statistics are probably the most important component of data analysis; they enable the findings to be described and indicate trends. These trends form the basis for subsequent analysis, interpretation and discussion. As with other statistics, descriptive statistics should be used under prescribed conditions (Vincent, 1995). The mean and standard deviation are preferred as they are sufficient, unbiased and efficient estimators (see Howell, 1992 for an explanation of these terms). However, the mean and standard deviation are unduly affected by outliers. As a result, the median and interquartile range are generally more appropriate for data that do not satisfy the requirements for parametric tests.

7.5.2 Inferential statistics

The choice of the appropriate statistical test should be made according to a 'decision tree' (see, for example, Vincent, 1995, and Table 7.8 for clarification).

Choice of test. There are many considerations in choosing an appropriate statistical test. Before the initiation of the study, the choice of test should be narrowed down by known characteristics. These are the research question (whether it addresses an association or correlation), the level of measurement (location of the dependent variable in one of the four main categories on the measurement continuum), and the sampling and assignment of subjects. The next stage is to check that the assumptions for parametric tests are met.

Testing normality. Parametric statistical tests usually assume that the sampling distribution (of the mean, for example) is normally distributed. This can be best approximated by the use of large samples and random sampling; the sampling distribution then approaches normality even for non-normal populations (the central limit theorem): this is the main reason for requiring certain sample sizes before using parametric statistical tests. Any single sample can provide an estimate of the sampling error; for example the standard error of the mean is the ratio of the sample standard deviation to the square root of the sample mean. For this, and other sample statistics, to give unbiased estimates of the population parameters, it is assumed that the characteristics of the normal curve apply to the sample data. This is less likely, and the sampling error is larger, for small non-random samples. It is, therefore, important to describe the distribution of the sample's dependent variables. To assess normality, Vincent (1995) suggests that the Z scores for kurtosis (scores congregating in the tails of the distribution) and skewness (one tail more pronounced than the other) should be within a range of $\pm$ 2. These Z scores are obtained, for example, for the kurtosis by dividing the kurtosis value by the standard error of the kurtosis. These values are often produced by software packages (e.g. SPSS for Windows ®).

Test specific assumptions. There can be many assumptions that should be met for the justification of specific statistical tests. A selection of tests and their principal assumptions are given in Table 7.8. Violations principally result in the increase in the occurrence of type I errors: adjustments must therefore be made to correct for this. It should also be noted that robustness to assumptions often exists and can allow test results to be accurate. Reference to suitable texts (e.g. Kinnear and Gray, 1994) should be made to determine the meaning and assessment of assumptions and how to correct for their violation. Three main assumptions will be discussed below: homogeneity of variance for several parametric tests of association, cause and effect for tests of relationships, and linearity for linear correlation and regression analyses.

- *Homogeneity of variance.* Homogeneity of variance is the assumption that the variances of groups are similar. It can be assessed in many ways, such as by using the Levene's test or the test proposed by O'Brien (1981). Homogeneity of variance should not be tested using *F*-ratios, however, as these are badly affected by non-normality. A simple check is to assume homogeneity of variance if the ratios of variances between groups are not greater than four (Howell, 1992) or, more strictly, two (Vincent, 1995). The position regarding the violation of homogeneity of variance (and normality) is summarised below for *t*-tests.

  - ♦ If sample sizes are equal, violation of homogeneity of variance produces an α value within 0.02 of the nominal value, which is tolerable. A similar comment applies for violations of normality assumptions if the populations are roughly the same shape or both symmetric. Tests are robust to normality and homogeneity of variance violations for distributions of different shape if the sample sizes exceed 25 (Box 1953; Boneau, 1960). If the populations are markedly skewed, particularly in opposite directions, serious problems arise unless variances are fairly equal, and *t*-tests are then not recommended.

  - ♦ If sample sizes are not equal and heterogeneity of variance exists, the actual and nominal α values differ considerably (Boneau, 1960). However, the Welch-Satterthwaite solution (see Howell, 1992) may reinstate robustness even under these conditions.

Test	Design							Assumptions															
	Level of measurement	Between or independent	Within or correlated	No. of groups, factors or IVs	No. of levels within groups or factors	No. of DVs	No. of covariances	Random sampling and normal distribution	Homogeneity of variance	Compound symmetry (sphericity)	Covariates independent of IVs	Covariates linear with DV	Homogeneity of regression coefficients	Relevant DVs only included	Non-multicolinearity of DVs	Uni- and multi-variate outliers removed	Minimum subjects per group [or ratio to each IV]	Linearity between DV and IV(s)	Homoscedasticity	Singularity	Non-multicolinearity	Normal distribution of errors	Inclusion of all relevant IVs
Association																							
Chi^2	N	x		1+	2+	1											5						
McNemar	N		x	2		1																	
Cochran's Q	N		x	3+		1																	
Mann-Whitney U	O	x		2		1																	
Wilcoxon matched pairs	O		x	2		1																	
Kruskall-Wallis	O	x		3+		1																	
Friedman's 2-ANOVA	O		x	2	2+x2+	1																	
1 group t-test	I/R	x	x	1		1		x	x								5						
Independent t-test	I/R	x		2		1		x	x								5						
Paired t-test	I/R		x	2		1		x	x								5						
1-ANOVA (B)	I/R	x		1	2+	1		x	x														
1-ANOVA (W)	I/R		x	1	2+	1		x		x													
1-ANCOVA(B)	I/R	x		1	2+	1	1+	x	x		x	x	x										
2-ANOVA (BB)	I/R	x		2	2+x2+	1		x	x														
2-ANOVA (WW)	I/R		x	2	2+x2+	1		x		x													
2-ANOVA (BW)	I/R	x	x	2	2+x2+	1		x	B	W													
MANOVA	I/R	x	x	2	2+x2+	2+		x		x				x	x	x							
Relationship																							
Phi	N		x	2	2x2	1											5						
Cramer's V	N		x	2	3+x2+	1											5						
Spearman's	O		x	2		1																	
Pearson's	I/R		x	2		1		x															
Linear regression	I/R		x	1		1		x									[10]	x					
Multiple reg. (Stepwise)	I/R		x	2+		1		x									[10(40)]	x	x	x	x	x	x

Figure 7.8. Assumptions required for the use of various statistical tests. Each notation indicates an assumption that should be fulfilled by the respective test, where: N is nominal data; O is ordinal data; I/R is interval or ratio data; B is a between design; W is a within design; x is a criteria to meet; each number represents its value (and upwards by a +) and x separates the required number for each group or factor or IV for a test.

It is also worth noting that larger sample sizes will tend to give a better chance of meeting the two main assumptions (homogeneity of variance and normality). Sample sizes of less than five are not suitable. With small sample sizes it is particularly important to ensure that the necessary checks have been carried out into whether violations have been made and if they are important. Non-parametric tests, as mentioned previously, may have greater power when parametric assumptions are violated (Blair and Higgins, 1985). The violation of homogeneity of variance (and normality) for ANOVA can be summarised as follows.

- The test is very robust regarding normality violations if sample sizes are equal. If the populations are symmetric or similar in shape (such as skewed in the same direction), and if the ratio of the largest to the smallest variance is less than four, the results are likely to be valid. A box plot or stem-and-leaf diagram (see Howell, 1992) can be used to check the shapes of the distributions of the samples quickly.

- Unequal sample sizes and heterogeneity of variance should not be mixed. However, use of the Welch (1951) solution may reinstate robustness under these conditions.

- Violations of the independence assumptions can seriously affect an analysis (Kenny and Judd, 1986).

- For repeated-measures designs, a further assumption is the 'compound symmetry of the covariance matrix'. This requires homogeneity of both variance (the leading diagonal of the matrix) and covariance (the off diagonal terms of the matrix). When compound symmetry is not demonstrated, several adjustments are available, such as that of Huynh and Feldt (1970). One of these should be applied, or alternative methods that do not require this assumption (such as MANOVA) should be used (e.g. Howell, 1992).

- *Cause and effect*. Correlation and regression analyses are often assumed to prove cause and effect: they do not, as only an underlying theory can do this, supported by a well-designed experiment. As in all statistics, what is provided is the reliability or an effect size that is either acceptable or as large as predicted. The effect size quantifies the strength of meaningfulness of the effect, that is the variance in the dependent variable that can be accounted for by the variance in the independent variables. The meaningfulness of the effect of interest can, for example, be quantified (as a ratio or percentage) by the use of the coefficient of determination (R^2). However, this does not necessarily tell us much about the real importance of the finding, which should be assessed by reference to the underlying theory that it is proposed to support.

- *Linearity*. Many examples occur in the literature of the use of linear correlations on non-linear data, such as between release speed and distance thrown when, in the absence of aerodynamic forces, the relationship is nearly quadratic. The following advice can help in avoiding this problem.

- Never perform a correlation without first drawing a scattergram to investigate the underlying relationship.
- If an underlying theoretical relationship is non-linear, find a transformation to linearise the data, or use a non-linear regression package.
- In the absence of an underlying theory, the 'principle of parsimony' can be evoked to support linear relationships, if a non-linear relationship makes only a modest improvement to the correlation coefficient. If the dependent variable must be zero when the independent variable is zero, then this might be considered as an underlying theory that should be enforced.

7.5.3 Effect size statistics

Effect size, also referred to as magnitude of effect or the effect size index in statistical tests, can be calculated for given associations or correlations between data sets. The effect size can provide a percentage value of the association or correlation between groups that is due to, or explained by, the experiment. In correlation analyses this is represented simply be the R^2 value; in association tests, specific formulas are used to calculate effect size. Reference to two tests (*t*-test and ANOVA) will be made below for the calculation of effect size.

For two groups, effect size = $(\bar{x}_1 - \bar{x}_2) / \sigma$, where $\bar{x}_1$ and $\bar{x}_2$ are the means of groups 1 and 2 respectively, and σ is the pooled standard deviation of the two groups. Effect sizes of <0.2 represent small differences, 0.5 moderate differences, and >0.8 large differences (Cohen, 1988).

For ANOVA, two methods of calculating effect size are common: eta squared (η^2), and omega squared (ω^2). The calculation of η^2 is simple; $\eta^2 = SS_B / SS_T$. A more accurate measure tries to account for the unexplained variance (and will probably produce a smaller value) as $\omega^2 = (SS_B - (k\text{-}1)(MS_E))/(SS_T + MS_E)$. For both these, an effect size of <0.05 is small, 0.10 is medium, and >0.20 is large (Cohen, 1988). Refer to section 7.4.3 for a clarification of the notation (where k is the number of groups).

The benefit of reporting the magnitude of effect is highlighted by the probability of either a type I or II error occurring. For example, through a small sample size and large variance a non-statistically significant finding can result (type II error). Alternatively, for a large n and small variances, a statistically significant result can be found (type I error). Both of these are problematic, but reporting effect size can indicate whether the statistical findings are meaningful.

Increasing the sample size has been proposed as a simple solution to many sampling problems, if time and resources permit. Reporting the effect size is much simpler, and provides readers with a means to interpret the importance of the findings.

7.6 Reporting of a study

Reporting of studies is covered in the various chapters of these guidelines, and some of the other issues are addressed by Yeadon and Challis (1994). Accuracy of reporting (e.g. referencing) is of great importance, as it reflects scientific rigour and quality (Morrow, 1991; Stull *et al.*, 1991). In general the reporting of research should include the following (where relevant).

- Sufficient details for the replication of the experiment, including sample characteristics and selection.
- The research design, including timing and numbers of measurements and treatments.
- Methods of controlling errors.
- Justification for chosen data analyses.
- Informative results, particularly descriptive statistics.
- Justification of the level of inferential statistical significance (α), probability values (P) for each comparison made, and the power of the tests used.
- Effect size, if only to support inferential statistical analyses.
- Reliability checks.
- Establishment and quantification of uncertainties in experimental data.
- Uncertainty propagation.
- Useful sections and content, such as discussion, conclusions and recommendations; these should not ‘stretch’ the implications of the research, and should consider the research question, theory, previous literature, results, limitations and future research.

7.7 Conclusions

This chapter has tried to highlight major issues associated with performing and reporting ‘quality’ research in sport and exercise biomechanics. Research design is a complex area, but comprehensive planning and careful implementation of a ‘good’ method can overcome many pitfalls. One particular problem, that of the misuse of statistics, can be minimised by the use of descriptive statistics, careful evaluation of the use of any inferential statistics and confirming their implications through effect size statistics.

For further advances in sport and exercise biomechanics, more innovative and correctly conducted research needs to evolve. The review paper of Yeadon and Challis (1994) provides a good starting point, and many useful references, for further study of the area. In addition, accurate and comprehensive reporting of this evolving research will greatly aid these advances.

Further texts for research methods in the sport and exercise sciences (e.g. Thomas and Nelson, 1996), statistics (e.g. Howell, 1992; Vincent, 1995), software package use (e.g. Kinnear and Gray, 1994, for SPSS for Windows ®), and error analysis (Taylor, 1982) are recommended. The examples given here are very readable.

We would ask the reader to note that, in the contents of this chapter, we have tried not only to provide guidance on the correct use of inferential statistics, but also to reflect current trends in the use of statistics in research. These trends include, for example, tendencies to move away from routine use of inferential statistical tests and to make more use of effect size statistics.

7.8 References

Armstrong, R.L. (1987). The use of statistical significance: a commentary related to Franks and Huck. *Research Quarterly for Exercise and Sport*, **58**, 81-82.

Bates, B.T., Dufek, J.S. and Davis, H.P. (1992). The effect of trial size on statistical power. *Medicine and Science in Sports and Exercise*, **24**, 1059-1068.

Baumgartner, T.A. and Strong, C.H. (1994). *Conducting and Reading Research in Health and Human Performance*. Dubuque, IA: Brown and Benchmark.

Blair, R.C. and Higgins, J.J. (1985). Comparison of the power of the paired samples t-test to that of Wilcoxon's signed-ranks test under various population shapes. *Psychological Bulletin*, **97**, 119-128.

Bland, J.M. (1995). *An Introduction to Medical Statistics*. Oxford: OUP.

Bland, J.M. and Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, **1**, 307-310.

Boneau, C.A. (1960). The effects of violation of assumptions underlying the t-test. *Psychological Bulletin*, **57,** 49-64.

Box, G.E.P. (1953). Non-normality and tests on variance. *Biometrika*, **40**, 318-335.

BSI (1979). Precision of test methods I: Guide for the determination and reproducibility of a standard test method (BS 5497, part 1). London: BSI (British Standards Institute).

Carver, R.P. (1978). The case against statistical significance testing. *Harvard Educational Review*, **48**, 378-399.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioural Sciences.* Hillsdale, NJ: Lawrence Erlbaum.

Dracup, C. (1995). Hypothesis testing - what it really is. *The Psychologist*, **8**, 359-362.

Fleiss, J.L. (1986). *The Design and Analysis of Clinical Experiments*. London: Harper and Row.

Franks, B.D. and Huck, S.W. (1986). Why does everyone use the .05 significance level? *Research Quarterly for Exercise and Sport*, **57**, 245-249.

Harris, R.J. (1985). *A Primer of Multivariate Statistics.* New York: Academic Press.

Hays, W.L. (1963). *Statistics.* New York: Holt, Rhinehart and Winston.

Holman, J.P. (1989). *Experimental Methods for Engineers.* London: McGraw-Hill.

Howell, D.C. (1992). *Statistical Methods for Psychology.* Belmont, CA: Duxbury Press.

Huynh, H. and Feldt, L.S. (1970). Conditions under which mean square ratios in repeated measurement designs have exact F-distributions. *Journal of the American Statistical Association*, **65**, 1582-1589.

Kenny, D.A. and Judd, C.M. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin*, **99**, 422-431.

Kinnear, P.R. and Gray, C.D. (1994). *SPSS for Windows Made Simple.* Hillsdale, NJ: Lawrence Erlbaum.

Krejcie, R.V. and Margan, D.W. (1970). Determining sample size for research activities. *Education and Psychological Measurement*, **30**, 607-610.

Lanshammer, H. (1980). Precision limits on derivatives obtained from measurement data. In *Biomechanics VII* (edited by K. Fidelius and A. Morecki), pp. 586-592. Baltimore, MD: University Park Press.

Morrison, D.E. and Henkel, R.E. (1969). Significance tests reconsidered. *The American Sociologist*, **4**, 131-140.

Morrow, J.R. (1991). Editor's viewpoint: shocking. *Research Quarterly for Exercise and Sport*, **62**, v.

Mullineaux, D.R., Scott, M.A. and Batterham, A.M. (1994). The use of 'agreement' for assessing digitiser reliability/objectivity. In *Proceeding of the Biomechanics Section Meeting of the British Association of Sport and Exercise Sciences* (edited by J. Watkins), no. 20, pp. 17-20. Leeds.

O'Brien, K.F. and Israel, R.G. (1987). Response to: Why does everyone use the .05 level of significance? *Research Quarterly for Exercise and Sport*, **58**, 85-86.

O'Brien, R.G. (1981). Comment on 'Some problems in the nonorthogonal analysis of variance.' *Psychological Bulletin*, **83**, 72-74.

Ottenbacher, K.J. and Stull, G.A. (1993). The analysis and interpretation of method comparison studies in rehabilitation research. *American Journal of Physical Medicine and Rehabilitation*, **72**, 266-271.

Safrit, M.J. and Wood, T.M. (1995). *Introduction to Measurement in Physical Education and Exercise Science.* St. Louis, MO: Mosby.

Scheirman, G.L., Smith, S.L. and Dillman, C.J. (1988). Three-dimensional kinetic and kinematic relationships in the long jump. In *Biomechanics XI-B* (edited by G. de Groot, A.P. Hollander, P.A. Huijing and G.J. van Ingen Schenau), pp. 561-564. Amsterdam: Free University Press.

Stull, C.A., Christina, R.W. and Quinn, S.A. (1991). Accuracy of references in research quarterly for exercise and sport. *Research Quarterly for Exercise and Sport*, **62**, 245-248.

Taylor, J.R. (1982). *An Introduction to Error Analysis*. Mill Valley, CA: University Science Books.

Thomas, J.R. and Nelson, J.K, (1996). *Research Methods in Physical Activity.* Champaign, IL: Human Kinetics.

Tuckman, B.W. (1978). *Conducting Educational Research*. New York: Harcourt Brace Jovanovich.

Vincent, W.V. (1995). *Statistics in Kinesiology.* Champaign: IL: Human Kinetics.

Welch, B.L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, **38**, 330-336.

Winer, B.J., Brown, D.R. and Michels, K.M. (1991). *Statistical Principles in Experimental Design*. New York: McGraw-Hill.

Yeadon, M.R. and Challis, J.H. (1994). The future of performance-related sports biomechanics research. *Journal of Sports Sciences*, **12**, 3-32.

Appendix 14

Mullineaux, D.R., Bartlett, R.M. and Bennett, S. (2001b). Research methods and statistics in biomechanics and motor control. *Journal of Sports Sciences*, **19**, 739-60.

Journal of Sports Sciences, 2001, 19, 739–760

Research design and statistics in biomechanics and motor control

DAVID R. MULLINEAUX,[1]* ROGER M. BARTLETT[2] and SIMON BENNETT[3]

[1]*School of Sport and Leisure Management and* [2]*Centre for Sport and Exercise Science, Sheffield Hallam University, Sheffield S10 2BP and* [3]*Department of Exercise and Sport Science, The Manchester Metropolitan University, Crewe + Alsager Faculty, Alsager ST7 2HL, UK*

Accepted 30 May 2001

Biomechanics and motor control researchers measure how the body moves and interacts with its environment. The aim of this review paper is to consider some key issues in research methods in biomechanics and motor control. The review is organized into four sections: proposing, conducting, analysing and reporting research. In the first of these, we emphasize the importance of defining a worthy research question and of planning the study before its implementation to prevent later difficulties in the analysis and interpretation of data. In the second section, we cover selection of trial sizes and suggest that using three trials or more may be beneficial to provide more 'representative' and valid data. The third section on analysis of data concentrates on effect size statistics, qualitative and numerical trend analysis and cross-correlations. As sample sizes are often small, the use of effect size is recommended to support the results of statistical significance testing. In using cross-correlations, we recommend that scatterplots of one variable against the other, with the identified time lag included, be inspected to confirm that the linear relationship assumption underpinning this statistic is met and, if appropriate, that a linearity transformation be applied. Finally, we consider important information related to the issues above that should be included when reporting research. We recommend reporting checks or corrections for violations of underpinning assumptions, and the effect of these checks or corrections, to assist in advancing knowledge in biomechanics and motor control.

Keywords: cross-correlations, effect size, trial size, variability.

Introduction

The science of research methods is continually developing, due largely to advances in technology and computer power. For biomechanics and motor control, the collection and analysis of data on muscle recruitment, force generation and movements by a sport performer and of the resulting outcomes is becoming easier. Measurements of these types of data share common ground in biomechanics and motor control, but the research questions that need to be answered are quite diverse both within and between these areas. The aim of this review is to clarify some issues of research design and statistics related to the measurement of data in biomechanics and motor control. These issues will be considered under four stages of research: proposing, conducting, analysing and reporting (Mullineaux and Bartlett, 1997). The main focus of the review is on issues in conducting research (i.e. trial size) and the analysis of data (i.e. effect size statistics, qualitative and numerical trend analysis and cross-correlations); a few pertinent issues on proposing (i.e. planning) and reporting (i.e. related to issues in the other three sections) research will also be included. This review should be read in conjunction with the other reviews in this issue, which cover issues relevant to research in the sport and exercise sciences. These include multivariate statistics and qualitative approaches in psychology (Biddle *et al.*), regression and analysis of variance in physiology (Winter *et al.*) and experimental control and reliability testing in sports performance (Atkinson and Nevill).

* Author to whom all correspondence should be addressed. e-mail: d.mullineaux@shu.ac.uk

Proposing research

The unique nature of scientific research was illustrated by Gould (1981, p. 22), who stated that science

Journal of Sports Sciences ISSN 0264-0414 print/ISSN 1466-447X online © 2001 Taylor & Francis Ltd
http://www.tandf.co.uk/journals

'progresses with hunch, vision and intuition'. In planning research, Kraemer and Thiemann (1987, p. 96) consider that 'a considerable degree of expertise in the field of application, experience, instinct, and creativeness' is required. It is difficult to cater for these elements in planning scientific research because of their subjective nature and, as there is a vast diversity of potential research questions, only common elements can be considered. We consider two common elements that should be fulfilled.

First, we must define a research question. Tuckman (1978) suggested that a research question should fulfil five criteria: provide practical value, have sufficient scope for a study, interest the researcher, possess theoretical worth and possess workability. For it to have practical value, the question must be currently unanswered and we must expect that the answer would contribute to knowledge. The scope for the study includes whether there are sufficient variables and potential results, and enough to write about. The researcher's interest can enhance the contribution to quality research from the subjective elements identified earlier and through the probable greater effort devoted to the task. Most importantly, the research question should possess theoretical worth – that is, it should be underpinned by a theory.

A worthy research question may fall into one of three categories: try to develop a theory to explain observations, lend support to a theory (e.g. repeatability) or improve a theory. A good theory 'must accurately describe a large class of observations on the basis of a model that contains only a few arbitrary elements, and it must make definite predictions about the results of future observations' (Hawking, 1996, p. 15). Consequently, a theory should be testable through the implementation of a study. For example, kinematic data showing that non-long jumpers ($n = 11$) decrease their variability in foot placement positions in the run-up to the take-off board (Scott *et al.*, 1997) were used to support the theory that gait is visually regulated. In contrast, the breaststroke turns by age-group swimmers ($n = 23$) were described using descriptive statistics on 21 variables, and the relationships between 19 of these were reported using a Pearson correlation matrix (Blanksby *et al.*, 1998), subsequently providing information that can be used to develop theory. The restriction on the length of journal papers prevents a theoretical explanation to be provided for all these variables and statistical analyses; such large numbers of analyses also introduce problems with respect to statistics (e.g. statistical power; see, for example, Kraemer and Thiemann, 1987). In some research, it may be appropriate to present the predictions from the theory as hypotheses. If hypotheses are presented, then these should be stated in terms of what is logically and justifiably expected. If any statistical significance testing is required, then it is paramount that the hypotheses are stated in this way, as significance testing is influenced by the context, prior knowledge and plausibility of the results (Matthews, 1998a). The importance of plausibility is highlighted in a report to the European Science and Environment Forum that concluded, 'By failing to take into account the intrinsic plausibility of the hypothesis under test, frequentist methods (i.e. *statistical significance testing*) are capable of greatly exaggerating both the size and the significance of effects which are in reality the product of mere chance' (Matthews, 1998b; emphasis added). Although plausibility is never absolutely determinable, prior knowledge and theory can assist in estimating the plausibility of the results. Prior knowledge can be of varying quality, whether published or not, and it is up to the reader to decide on its worth.

The final criterion offered by Tuckman (1978) in defining a research question, the issue of workability, brings us to the second element in planning research. Workability refers to the logistical considerations in planning research and has been addressed by several authors. For example, Coolican (1999, p. 19) considers that the variables, sample, design and analysis should all be reviewed in the planning stage. More broadly, Mullineaux and Bartlett (1997) consider that all the components of a research study (i.e. proposing, conducting, analysing and reporting) should be considered together to establish whether the results expected would provide an answer to the research question using the best methods available to the researcher. Much of this preparation can be enhanced through the influence of the researcher and through a literature review, from which should emerge an appropriate research design (i.e. structure of the data collection; see section on 'Conducting research') and method (i.e. process of conducting the study). Careful planning and implementation of the method is desirable, as O'Brien and Israel (1987) considered that this is the most important part of the research process. In biomechanics and motor control, a variety of research designs are used, although in general the results obtained are quantitative, requiring that the research design should incorporate all considerations essential for suitable and accurate descriptive, inferential and effect size statistical analyses. Nevertheless, the implementation of a good research design is of more importance than an over-emphasis on statistical issues to ensure that the results, and not the statistics, drive the analysis and interpretation of the research. The research question and method should be worthy, realistic and ethical and should be kept simple to enable ease of analysis, reporting and understanding by the reader. Although the researcher should already have considered the methods during the formulation of hypotheses, the next step is to formulate explicitly the methods for conducting the research.

Conducting research

When conducting research, with the aim of obtaining high-quality results that are of use in trying to answer the research question, consideration should be given to ethics, research design, and sources and control of errors. Researchers may often be bound by ethical stipulations set by funding, medical, ethical, legal and professional bodies. For example, members of the British Association of Sport and Exercise Sciences are bound by the profession's code of conduct (BASES, 1995) governed by three main principles. These emphasize: that clients should receive the highest standards of professionalism, consideration and respect; that research and testing should be carried out with utmost integrity and that working practices are safe; and that the welfare of the client is paramount. These principles are met by working within the guidelines of obtaining ethical clearance and informed consent, and ensuring care of the participants throughout.

A good research design – and research question – should develop from theory, with due consideration of constraints imposed by the research question, previous literature, resources, ethics and methods. The research design and method are very important because if they are flawed the value of the results is greatly reduced. A clear design assists readers in understanding and interpreting research. Describing the design clearly in terms of the paradigm – that is, the conceptual framework or type of research (e.g. experimental research) – links between groups (e.g. within-individuals), and the number and timings of treatments and measurements is essential if results are to be properly interpreted.

When conducting an investigation, it is important that errors are controlled and that their effect on the study is assessed. This assessment can be performed in several ways, including estimating errors and checking reliability. Errors in the results of an experiment can arise from such sources as the research design, apparatus, experimenter, participants and interactions between these factors (Mullineaux and Bartlett, 1997). Methods can be incorporated into the research design to reduce these errors. Because research is so varied, no general rules exist for controlling errors; however, some basic principles can be considered (see Table 1). For details on these, readers are directed to Thomas and Nelson (1996), Mullineaux and Bartlett (1997) and Baumgartner and Strong (1998). Applications of these principles are often implicit within a research design, for which the most pertinent ones to the specific topic can be identified through a thorough literature review.

In conducting research with humans, many research design issues need to be considered to ensure that the results obtained are valid and can be analysed to answer the research question. The cost and time commitment of testing introduces several issues, including the trade-off between the chosen and required sample size versus trial size. This issue will be further explored below, as the time-consuming nature of collecting, analysing and interpreting data is becoming less restrictive with technical developments. With more time available, should researchers use a larger sample or more trials?

Trial sizes required for human performance research

The validity of results obtained in an experiment include whether the results are representative of the 'typical' response for the experimental conditions. For example, this response can be considered to be a measure of performance (e.g. outcome measure) or a measure of technique (e.g. movement kinematics measure). One definition of a representative response is that it is the central tendency score (e.g. Kroll, 1957), which subsequently requires that variability – 'The degree of difference between each individual score and the central tendency score' (Thomas and Nelson,

Table 1. Example methods to control for experimental errors (from Mullineaux and Bartlett, 1997, p. 92)

Attempts to keep the testing as natural as possible
Minimize the information that it is necessary to provide to the participants
Physical manipulation of participants' actions before and during test period
Matching participants between groups to control for extraneous variables (or use a within-individual design)
Counterbalancing for participants receiving several treatments by varying the order of administration for different participants
Introduction of control groups (e.g. placebo) and using single- or double-blind designs
Ensure experimenters are competent in using the equipment and the procedures involved
Calibration of equipment
Use of the best available equipment
Appropriate statistical analysis and use of methods to factor out extraneous variables (e.g. ANCOVA)

1996, p. 100) – is measured. To obtain these measures, several trials need to be recorded. Trial size is influenced by experimental errors, robustness of the assumptions of the analysis method, research design, consistency of participants' responses and definition of variability used. Below, we discuss some factors that influence variability (consistency of participants' responses and sources of variation), methods for estimating the number of trials required within a research design, and methods of assessing variability and analysing multiple trials.

Consistency of participants' responses

Because of the complexity of human beings – from the vast array of neural connections to the multidimensionality of joints, and the unique experiences, perceptions, intentions and expectations that are brought to the experimental setting (Dufek *et al.*, 1995) – the same goal (i.e. outcome measure) can be achieved using many different techniques (i.e. motor equivalence). This inherent variability has been seen as being functional, in that it permits the flexibility to adapt to a non-stationary environment and hence changing task demands (for a review, see Newell and Corcos, 1993). Bearing this variability in mind, it can be implied that, when data collection involves humans, the analysis of a single trial may be inappropriate, as the assumption of a single performance or technique measure being representative of the typical response should be viewed with caution (Bates *et al.*, 1992). Subsequently, the use of single-individual designs in which multiple trials are analysed have been proposed by Bates (1996), although Reboussin and Morgan (1996) suggest these are only beneficial in the early stages of research for hypothesis generation partly because of problems of generalizability (see James and Bates, 1997, for a review on the analysis of single-individual designs, i.e. the analysis of several trials of one participant in a study). In multiple-individual designs, for reasons such as time and cost restrictions, ease of sampling and statistical analysis, a single trial is still often used. If within-trial consistency in response measures is shown to be high, as seen for example in highly skilled performers (Hubbard and Seng, 1954; Bootsma and Van Wieringen, 1990), then analysing a single trial in multiple-individual designs may be justified. Nevertheless, there are still potential problems with using multiple-individual designs because the distribution of within-trial variability is mediated by the specific task demand.

In a study of elite table tennis players performing the attacking forehand drive, Bootsma and Van Wieringen (1990) found that the variability of the direction of travel of the bat declined from the moment of initiation to the moment of bat on ball contact. Therefore, rather than exhibiting an increase in variability throughout the drive as a result of noise in the system (i.e. magnification of error over time in performing the specific task), elite players used a strategy in which the variability was reduced throughout the drive to a minimum at bat on ball contact (see also Bootsma *et al.*, 1991). A further example of a reduction in variability in certain key variables has been reported by Arutyunyan *et al.* (1968). Based on their work on pistol shooting, they suggested that compensatory movements of the upper arm enabled expert marksmen to achieve low variability in the position of the pistol barrel. Novice marksmen, in contrast, were unable to demonstrate such compensatory movements and, therefore, exhibited more variability in the position of the pistol barrel. Clearly, in these tasks an analysis at certain temporal parts of the movement (e.g. initiation of drive compared to bat on ball contact) or at specific locations (e.g. upper arm compared to the pistol barrel) could have indicated different amounts of variability. If inappropriate variables are selected, then the results could be misleading in suggesting that elite performers are less consistent than novice performers.

Sources of variation

As well as intra-individual differences, there are often large inter-individual differences in participants' responses. With a research design that examines differences in responses within a group of performers, individual differences in performance or technique may be masked (Michaels and Beek, 1996). For example, Temprado *et al.* (1997) found that although both expert and novice groups of volleyball players exhibited three types of joint pair coordination, in which there were group differences in the frequency of occurrence, this tri-modal pattern was not found in two of the six experts. In recognition of this, Bates *et al.* (1992) suggested that individuals' responses should only be grouped after verification of any similarities or trends in the data, as similarities are likely to prove to be exceptions. Burden *et al.* (1998) studied the hip and shoulder rotations of seven golfers. Despite the golfers' similarly high standard (mean $\pm s$: golf handicap of 7 ± 1), Burden *et al.* found large inter-individual variations (e.g. maximum shoulder angle of $102 \pm 16°$, range 80–126°). Informative and understandable results were provided by simple data analyses including descriptive statistics of the group and individuals' deviations from these. The use of simple analyses has merit, as combining such data (e.g. in an analysis of variance, ANOVA) could be problematic despite similar abilities, and lends support for other less advanced statistical or data analyses (e.g. angle–angle diagrams; see section on 'Trend analysis').

Clearly, both inter- and intra-individual variation are possible. These variations influence reliability and statistical significance testing. Classical test theory (see Safrit and Wood, 1995) considers that variance in the observed score (σ_x^2) equals the sum of the variances in the true (σ_t^2) and error (σ_e^2) scores. Reliability is subsequently defined as the ratio of the true to observed variance:

$$\text{reliability} = (\sigma_x^2 - \sigma_e^2)/\sigma_x^2 \qquad (1)$$

In an ANOVA, this can be presented as a version of the intra-class correlation coefficient:

$$R = (\text{MS}_\text{B} - \text{MS}_\text{W})/\text{MS}_\text{B} \qquad (2)$$

where MS_B and MS_W denote the mean-square variance between-groups (variance of the group means from the grand mean) and within-groups (variance of the individual scores from the mean of its group), respectively. Assuming that only MS_B changes, if this increases then reliability increases but statistical power decreases because individual heterogeneity increases. Alternatively, if only MS_W increases then, as this is assumed as error, the reliability and the power of the test decreases because MS_B forms part of the denominator for effect size, which influences statistical power. This example emphasizes that it is important to identify the source of variation in an experiment. For example, the intra-class correlation coefficient is recommended in many textbooks (e.g. Thomas and Nelson, 1996; Vincent, 1999). However, because of limitations – including the conflicting effects of variability – others have criticized it and proposed different techniques. These include limits of agreement (Bland and Altman, 1986) and least-products regression (Ludbrook, 1997). Often the limitations are specific to the data. Hence, more recently, some authors have proposed the use of either a combination of tests (e.g. Rankin and Stokes, 1998, recommend the intra-class correlation coefficient plus limits of agreement) or recommend a test that meets the underpinning assumptions (e.g. Mullineaux *et al.*, 1999, compare limits of agreement and least-products regression). For a review on reliability testing, see Atkinson and Nevill (1998).

An extension of classical test theory is generalizability theory, in which sources of variation in a measurement can be quantified (see Safrit and Wood, 1995). The principal benefits of this theory are that it: enables sources of error to be quantified; provides future research with information to use in designing methods to minimize error and increase reliability and power; and aids in interpreting future results by providing a quantification of variability. Salo and Grimshaw (1998) have applied generalizability theory to the kinematic analysis of sprint hurdles. They considered that total variance of a single observation in digitized data equalled the sum of variance from three sources: between-individuals (4 females, 3 males), within-individuals repeated trials ($n = 8$) and redigitization ($n = 8$ for one male and one female). They reported percentage coefficients of variation (%CV) ranging from 1.0% for the women's centre of mass height at landing to 209.7% for the men's centre of mass horizontal distance from the hurdle at maximum clearance height. For redigitization, the variability in results for male participants was better than for female participants. For the males, the mean %CV of 9.1% ranged from 0.1% for the centre of mass mean horizontal velocity to 93.1% for the centre of mass horizontal distance from the hurdle at maximum clearance height. The authors commented that the more computations required, the greater the variability (e.g. displacement versus centre of mass variables) and, subsequently, that more trials would be required to obtain a representative measure. The authors did recognize that, as the %CV contains the mean ($\bar{x}$) as the denominator,

$$\%\text{CV} = 100\ s/\bar{x} \qquad (3)$$

then as the mean approaches zero, the %CV will increase even for low absolute standard deviations (s). For example, as the women's mean centre of mass horizontal velocity and vertical take-off velocity were 6.9 and 1.6 $\text{m}\cdot\text{s}^{-1}$, respectively, both with a standard deviation of 0.2 $\text{m}\cdot\text{s}^{-1}$, then the coefficients of variation are different at 2.9% and 12.5%, respectively. This consideration of the delimitations of statistics is invaluable in analysing results, as knowledge of these can also help in designing and interpreting a study. Hence, using %CV in the example above would suggest that the take-off velocity is more variable in the vertical than in the horizontal, but this is partly attributable to a limitation of the statistic being its inability to cater for different sizes in the mean score. All statistics have delimitations, a few further examples of which will be illustrated later.

Estimate of the number of trials required

Generalizability theory is limited to the sources of variances investigated and the context in which the theory is studied. Salo and Grimshaw (1998) considered between-individuals, within-individuals repeated trials and redigitization variances; further research is required to identify variances from other factors, including sampling frequency, the ability of the performer and an individual's years of experience at the task. The context of this study applies to the particular research design involving manual digitizing, a sampling frequency of 50 Hz, a large field of view and analysis of variables requiring various amounts of data processing (e.g. differentiated versus raw data). As a measure of consistency, Salo *et al.* (1997) calculated the number of trials required in sprint hurdles to obtain a reliability of 0.90 using the intra-class correlation coefficient on a variety of kinematic variables. Depending on the consistency of certain variables, between 1 (e.g. maximum knee angle of the lead leg during clearance for males and females) and 78 trials (e.g. decrease in centre of mass horizontal velocity from touchdown to landing for males) can be needed for a reliable and representative measure of the hurdler's average technique to be obtained. Although the context of this experiment is quite limited, it does emphasize that a single measurement can be unrepresentative of the average movement pattern.

This type of research on quantifying variability is useful in determining factors that influence variability, but few studies in biomechanics and motor control provide such information, and even fewer use such findings for designing studies or interpreting results. Interpreting results with respect to sources of variability identified in the literature is simple. The importance of this simple approach is illustrated below, as determining the required trial size and methods of analysis of multiple trials for an experiment to quantify variability is not straightforward.

As outlined above, there are many sources of variability that can influence data validity and, subsequently, that more than one trial will probably be required to obtain a representative measure of a performer's technique or outcome measurement. However, including more than one trial in an experiment introduces new problems such as fatigue and learning, and time and cost restrictions. These potential problems can be addressed by appropriate research designs, solutions that are often described in the literature (see Table 1). To provide some insight into aspects of the research designs adopted in biomechanics, Table 2 describes some relevant features of research papers published in the *Journal of Applied Biomechanics* in 1998.

In biomechanics research, the sample and trial sizes used are often small. Table 2 shows that more trials were recorded (mean of 3.9) than were analysed (mean of 2.9), and in 50% of papers only one trial was analysed. The use of multiple trials is rare, yet earlier we illustrated the potential problem of using only one trial as a representative measure of a performer's technique and, subsequently, the effect on the validity of the results. There are also benefits of more than one trial for statistical significance testing, which is discussed later. It is on this basis that Bates *et al.* (1992) recommended the number of trials that should be included in a repeated-measures design. They suggested that, for statistical power of 90%, the ability to detect a one standard deviation difference between the group means, trial sizes of 10, 5 and 3 should be used for sample sizes of 5, 10 and 20, respectively. Interestingly, their results showed that if trial size is increased, then the required sample size decreases at a proportionally greater rate. This decrease is beneficial, as fewer data are required, but this has yet to be explored fully. However, as Carver (1978) pointed out, even though it is desirable to build replication into designs, this can involve a more complex process of analysis. The criteria used for selecting which trials to analyse from those recorded and the methods of data treatment of more than one trial used in studies published in the *Journal of Applied Biomechanics* in 1998 are described in Table 3.

Analysing multiple trials and quantifying variability

Generally, three approaches exist in the methods for selecting the number of trials to be analysed from those recorded (see Table 3): only one trial collected (26.9% of studies), all trials are used (38.5%) or the best of several trials is used (23.1%). In a few cases, it was not clear what the criteria had been for selecting trials (11.5%). In 50.0% of studies, one trial was used as

Table 2. Summary of selected research design features from research papers ($n = 22$) published in the *Journal of Applied Biomechanics* in 1998

	Mean ± s	Range*	Notes
Sample size	14.5 ± 13.5	3–67	Most participants were human (rather than objects)
Groups/conditions	2.4 ± 2.6	1–14	Most used one group and several conditions
Trials recorded	3.9 ± 3.0	1–10	26.9% recorded only one. One study selected the number of trials on theoretical grounds derived from the literature
Trials analysed	2.9 ± 2.8	1–10	50% analysed only one trial
Total trials	80 ± 126	5–504	This is the product of sample size, group/conditions and trials analysed

Note: The complexity of designs has necessitated some simplification. Five papers as technical notes and one invited review paper have been excluded from the summary. * The normal distribution assumption required of the mean and standard deviation statistics is violated, hence the range has been included to clarify the data dispersion.

Table 3. Criteria for selecting trials from those recorded and data treatment methods of trials described in papers ($n = 22$) published in the *Journal of Applied Biomechanics* in 1998

Criteria for selecting trials	Percentage	Data treatment of trials	Percentage
Only one recorded	26.9	None (only one available)	50.0
Best one of several	23.1	Averaged	30.8
All trials	38.5	Unknown	19.2
Unknown	11.5		
Total	100.0	Total	100.0

either it was the only one collected or the 'best' trial was selected. Where more than one trial was selected, the average of these was used in subsequent analyses (30.8%); in the remaining studies, the methods for treating trials was not clear (19.2%).

This simple summary of the criteria for selecting the number of trials and data treatment of trials in biomechanical research is quite informative (Table 3). In this instance, only two techniques for treating multiple trials on the same condition have been identified: taking the 'best' or taking the 'average'. Either technique can produce 'better' results, but the choice should be based on a sound foundation. Kroll (1957) suggested that if the trial-to-trial error variance is random and uncorrelated, confirmed by a non-significant between-trials F-statistic on a one-way repeated-measures ANOVA, then choosing the average score is the only correct method. Other methods, such as taking the 'best', violate the operating assumption of random, uncorrelated error variance. However, if the error variance is not random or uncorrelated – examples of which were described earlier (e.g. the variability in some joint pairs is used to reduce variability in the pistol barrel by elite performers; Arutyunyan *et al.*, 1968) – then the choice of appropriate methods is not so definitive. Kroll (1957) considered that methods including the average or best could be defensible, but preferably the researcher should identify a measurement schedule for random and uncorrelated error variance. This again highlights that appropriate analysis based on *a priori* theorizing should be used to obtain valid results.

One other feature of these papers was that variability had occasionally been quantified for data interpretation or controlled for statistical purposes. For example, Caldwell *et al.* (1998) used the coefficient of variation and Young and Marteniuk (1997) used the root mean square to quantify variability. Also, Barrentine *et al.* (1998) used repeated-measures ANOVA to control for variability between individuals. Other techniques available to control variability, in instances where this is required, include analysis of covariance (ANCOVA), matched-individuals designs and further methods described later (see section on 'Trend analysis' and Table 6). Shultz and Sands (1995, p. 272), in their review of measurement and statistics for exercise specialists, supported the use of reducing variance owing to individual heterogeneity. They considered that the resulting increase in effect size is the most cost-effective method of increasing statistical power, especially instead of increasing the sample size. All of these methods have their benefits, notably that they provide some quantification or control of variability; however, they also have their limitations. A limitation of the coefficient of variation was illustrated earlier (see equation 3). To use another example, as ANCOVA uses regression to adjust the values of the dependent variable for each group so that there is an equal proportional relationship with a covariate, this adjustment introduces error into the data. Although this will increase the statistical power, it decreases data validity if care is not taken in its use. The importance of using any of these statistical techniques is that they can help determine how many trials might be required in future research so that more valid results can be obtained.

Summary

We hope we have highlighted the conceptual importance of increasing trial size, particularly in obtaining more valid results by being more representative of performance or technique. To determine an appropriate trial size, Caldwell *et al.* (1998) used the coefficient of variation and Salo and Grimshaw (1998) used generalizability theory to assess variability. These measures or assessments can provide an insight into methods of experimental control and they stress the need to increase trial sizes that will subsequently increase reliability and statistical power. Simpler and more apparent methods of determining trial size are limited. We noted earlier the recommendations of Bates *et al.* (1992) on the trial sizes required for different sample sizes to obtain a 90% power to detect statistically a one standard deviation difference between the groups' means. As statistical significance is affected by many factors, this interaction between trial and sample sizes was extended by Dufek *et al.* (1995) to assess the influence of single-individual versus group analyses, effect size and variability on statistical power. In general, larger trial sizes are required to detect statistically significant findings for inter-individual analyses, when effect sizes are small and when variability is large. Their tabulated and graphical results can be of further use in identifying the number of trials required in designing an experiment.

As resources are limited, increasing the trial size is often at the expense of sample size. Sample size is often increased to increase statistical power; however, experimental control of variation and increased trial size can perform this task more economically. Also, increasing trial size may be more suitable when the available sample sizes are small. In analysing multiple trials, trials are commonly averaged, as this decreases the within-individual variation and increases statistical power, or they are analysed in repeated-measures analyses of variance. In summary, we recommend that some assessment of variability should be included or reported and, where possible, more than one trial should be used. If possible, trial size should be determined *a priori* based on considerations complementary to the inherent variability in the movement and appropriate use of statistics. The simple recommendation of using the average of three trials may have some merit as a practical way of meeting the theoretical basis for obtaining more valid results both experimentally and statistically. In addition, rather than just controlling variability for statistical purposes, it can be meaningful to discuss variability in the light of its importance in the successful control and outcomes of movements (see, for example, Arutyunyan *et al.*, 1968).

Data analysis

The many methods of analysing data include the use of qualitative analysis (e.g. describing the coordination of two segments as 'tightly coupled'; see section on 'Trend analysis') and the three major forms of statistics (i.e. descriptive, inferential and effect size). The choice is dependent on the aim of the analysis, characteristics of the sample, type of data, research design, political interventions and the experimenter's statistical training. The correct use of each of these can have substantial benefits for the interpretation of analysed data. Unfortunately, their inappropriate use can waste time and over-complicate an analysis. This, in turn, can mask important information or misinform the reader. Data in biomechanics and motor control research often possess two features: the sample size is small and the data form time series. The former is usefully analysed by effect size statistics, the latter by trend analysis.

Effect size

Effect size forms part of the mathematics underpinning many inferential statistics and influences the significance value obtained. It has most often been observed explicitly in research forming the basis for meta-analysis. Effect size statistics provide a quantification of the magnitude of the association between data. Alternatively, Shultz and Sands (1995, p. 266) define effect size as 'an index of the degree of departure from the null hypothesis and is mathematically related to the noncentrality parameter of most test statistics'. These two definitions support two different applications of effect size as descriptive or inferential statistics, respectively. This section primarily emphasizes the use of effect size as a descriptive statistic, as many of the limitations of inferential statistics described in the literature could also apply.

The recommendation of effect size as a descriptive statistic in research has mostly arisen from the debate over the usefulness of inferential statistics. This debate appears to contain one universally accepted component: significance values (P) only provide the probability of obtaining the result *assuming* it were due to chance. In a recent paper, Chow (1998) set out the arguments for using inferential statistics, principally in theory-corroboration research. The many responses to his paper illustrate the varied support and opposition to this position (see Open Peer Commentary in Chow, 1998, pp. 194–228). In theory-corroboration, 'Statistical significance means only that chance influences can be excluded as an explanation of data; it does not identify the nonchance factor responsible' (Chow, 1998, p. 169). Many of the opponents of significance testing agree with this statement (e.g. Carver, 1978), but the problem is that, if the null hypothesis (H_0) is rejected, most researchers automatically accept the research hypothesis without trying to discount the many other alternative hypotheses. Others believe that the null hypothesis should be considered with respect to the 'plausibility' of the result (e.g. Matthews, 1998a,b), otherwise many of the significant results are meaningless flukes. Many authors have highlighted other limitations of inferential statistics (e.g. Carver, 1978; Armstrong, 1987). One of the practical limitations is the binary output (accept or reject) produced by rejecting a hypothesis when a specific level of significance (α) is exceeded, and that no quantification of the finding by a unit of measurement is provided. It is also possible to obtain either a type I or II error. A large sample size and small variability predispose towards the former, where differences are small but the null hypothesis is rejected. A small sample size and large variability predispose to type II errors, where differences are large but the null hypothesis is accepted. Armstrong (1987) noted that generalization is often incorrect and that inference should be performed with great care to compensate for such problems. Some of this dissatisfaction has led researchers to propose alternatives, including confidence intervals (e.g. Borenstein, 1997; Sim and Reid, 1999), graphical representation (e.g. Wainer and Thissen, 1981) and effect size statistics (e.g. Cohen, 1988).

A particular limitation in biomechanics and motor control research is that the reliance on small sample sizes reduces the statistical power and may result in non-significant findings. These, in turn, can lead to misleading inferential statistics and can cause major problems in the control of type I errors in certain statistical tests. Several solutions exist. Increasing the sample size is an obvious simple solution if time and resources permit; however, although this would increase the power, it would not negate the above criticisms. Simpler still would be to report effect sizes and provide readers with a means to interpret the importance of the findings (Mullineaux and Bartlett, 1997). It is unlikely that inferential statistics will become obsolete; therefore, it would be beneficial to use effect sizes to support rejection or acceptance of the null hypothesis (e.g. to justify there is no type I or II error) or to support results from descriptive statistics.

Effect size has been proposed as an alternative to inferential statistics because it is easy to calculate and it provides more meaningful results. The notion of effect size statistics has become more widespread in recent years (e.g. Vincent, 1999). They have been used in a few studies in biomechanics (e.g. Goosey *et al.*, 1998), often together with inferential statistics. But how common is their use generally? The results of a simple search for 'effect size' on the Science Citation Index (Web of Science, 2000), in comparison to '*t*-test' and 'ANOVA', are reported in Table 4.

In all years, 'ANOVA' has generally been reported the most, followed by '*t*-test' and then 'effect size'. Between the start of the Science Citation Index database in 1981 and 1990, the frequency of occurrence of 'effect size', '*t*-test' and 'ANOVA' remained small and relatively constant. Since 1991, the frequency of the reporting of each of these terms has steadily increased, but the size of the database also almost doubled each year up to 1998. Despite some early publications (e.g. Cohen, 1969) and later ones (e.g. Ottenbacher and Barrett, 1991) that proposed the benefits of effect size statistics, it appears that they have not been widely adopted. Possibly, the use of inferential statistics still provides researchers with 'objectivity' and the 'gatekeepers' (e.g. journal editors and funding bodies) still seek this supposed objectivity. Researchers may find comfort in this 'objectivity' (Matthews, 1998a). However, inferential statistical analysis is complex and requires underpinning knowledge for correct use. Some researchers do not possess that knowledge and, therefore, use inferential statistics incorrectly. It is possible to obtain more useful results by simplifying the analysis, an example of which was described earlier (see Burden *et al.*, 1998, on p. 742).

Although effect size statistics have primarily been proposed as an alternative to inferential statistics, care must still be taken to ensure that they are used correctly. Authors rarely report whether the assumptions underpinning inferential statistical tests have been met. This may be due to their complexity, the availability of computer statistical packages and the views of some journal editors. Effect size may fall into the same trap, as few textbooks mention the underpinning assumptions required for a valid calculation. 'Justifying' the use of any statistical test by ensuring that the assumptions of the test are sufficiently met or controlled has implications for a correct analysis and better interpretation. The most common effect size statistics are parametric, requiring random sampling and normally distributed data. Each test also has specific assumptions. For example, omega-squared, which can be used to support a between-groups ANOVA, also requires homogeneity of variance between the groups. Effect size suffers from the same problem as inferential statistics in that checks and corrections for the assumptions are not routinely performed.

Table 4. Frequency of statistical tests reported in the title, abstract or keywords of articles contained in the Science Citation Index (Web of Science, 2000) from 1981 to 1999

Year	Effect size	ANOVA	*t*-test	Notes
1981–90	0–3	2–15	2–9	The size of the database remained constant at *c.* 1200 articles per year
1991	30	298	321	In 1991, the database contained *c.* 2400 articles and almost doubled each year up to *c.* 180,000 articles in 1998 (derived from BIDS ISI Data Service, 1998)
1999	130	1082	956	Frequency of all tests gradually increased from 1991

Calculations of effect size can be performed for many research designs. There are two main types, explained variance and effect size statistics, although additional names include magnitude of effect, effect size index and the individual names of tests such as eta-squared. Effect size is calculated in the original units of measurement. The most common is a standardized score (i.e. Z) indicating by how many standard deviations the group means differ. In general, an effect size of 0.2, 0.5 and >0.8 represents small, moderate and large differences (Cohen, 1988), respectively, where a moderate difference is considered visible to an experienced researcher (Cohen, 1992, p. 156). The explained variance, in contrast, indicates the amount of association or correlation between groups that is due to, or explained by, the experiment or treatments. The output is provided as a decimal (e.g. 0.60) or percentage (e.g. 60%) where, for example, a coefficient of determination (r^2) of 0.6 indicates that 60% of differences or variance are due to the variability in the treatment and 40% are due to extraneous variables. Although the explained variance is typically presented in this way without units, it can be converted to an effect size (see Cohen, 1988). However, Shultz and Sands (1995, p. 268) suggest that effect size is underestimated by the explained variance. For example, an effect size of 0.8, although considered large (Cohen, 1988), only equates to an explained variance of 14% (see Shultz and Sands, 1995, for the conversion formula). Such an explained variance appears small and might be misleading, hence such a conversion is not recommended.

As effect size can be beneficial in supporting significance tests, several tests are provided in Table 5 for supporting Pearson rho correlation coefficient, t-test and one-way ANOVA inferential statistical tests. To identify effect size statistics (e.g. a two-way between-group ANOVA) that are of use in supporting other inferential statistics, see, for example, Howell (1997) and Thomas and Nelson (1996, p.166). Matching effect size statistics with inferential statistics is beneficial, as both generally share the same assumptions. If the data violate the respective assumptions for the inferential statistic, then either a correction should be applied or an appropriate non-parametric statistic should be used. These assumptions are described in many textbooks and many statistical packages have corrections as features.

Table 5. Examples of explained variance, effect size or equivalent tests (*) suitable to support analyses with selected inferential statistics

Test[a]	Explained variance[b]	Effect size[c]
Pearson	r^2	*$SE_E = s_Y \cdot \sqrt{(1-r^2)}$
t-test (B)	$\omega_t^2 = (t^2-1)/(t^2+n_1+n_2-1)$	‡$ES_B = (\bar{x}_1 - \bar{x}_2)/s_C$
t-test (W)	*$PC = 100 \cdot ((\bar{x}_{POST} - \bar{x}_{PRE})/\bar{x}_{PRE})$	$ES_W = (\bar{x}_{POST} - \bar{x}_{PRE})/s_{PRE}$ *$LOA = \delta \pm 2\sigma$
ANOVA (B)	$\eta^2 = SS_B/SS_T$ $\omega_F^2 = (SS_B - (k-1) \cdot MS_E)/(SS_T + MS_E)$	ES_B for pairs of means
ANOVA (W)	N/A	$ES_A = (\bar{x}_{MAX} - \bar{x}_{MIN})/s_W$ *$LOA_B = \pm 2 \cdot \sqrt{2MS_E}$

Note: See Bland and Altman (1986) for LOA, Bland (1995) for LOA_B, Park and Schutz (1999) for ES_A and Vincent (1999) for remaining formulae. N/A = not applicable.

[a] Pearson product-moment correlation coefficient (Pearson); between-groups factor (B), within-groups factor (W).

[b] Coefficient of determination (r^2); omega-squared for a between-groups design t-test (ω_t^2); t-test statistic (t); group 1 sample size (n_1); group 2 sample size (n_2); percentage change (PC); post-test mean ($\bar{x}_{POST}$); pre-test mean ($\bar{x}_{PRE}$); eta-squared (η^2); sum of squares between groups or treatment variance (SS_B); sum of squares total or total variance (SS_T); omega-squared for a between-groups F-statistic (ω_F^2); number of groups (k); mean-square error variance (MS_E).

[c] Standard error of the estimate (SE_E); standard deviation of dependent variable (s_Y); coefficient of determination (r^2); effect size of between-factor (ES_B); group 1 mean ($\bar{x}_1$); group 2 mean ($\bar{x}_2$); standard deviation of control group (s_C); effect size of within-factor (ES_W); post-test mean ($\bar{x}_{POST}$); pre-test mean ($\bar{x}_{PRE}$); standard deviation of pre-test (s_{PRE}); levels of agreement (LOA); mean of differences (δ); standard deviation of the differences (σ); effect size of within-factor for ANOVA (ES_A); largest within-factor mean ($\bar{x}_{MAX}$); smallest within-factor mean ($\bar{x}_{MIN}$); mean within cell standard deviation (s_W); levels of agreement boundary (LOA_B); mean-square error variance (MS_E).

‡ When s_C is not available, use the pooled standard deviation, $s_P = \sqrt{(s_1^2 \cdot n_1 - 1 + s_2^2 \cdot n_2 - 1)/(n_1 + n_2 - 2)}$, including the standard deviations of groups 1 (s_1) and 2 (s_2), and sample sizes of groups 1 (n_1) and 2 (n_2).

For a correlation between two sets of data, the explained variance provides the most appropriate measure of the size of the effect. If the data meet the assumptions of a Pearson correlation, then the explained variance is simply the coefficient of determination – that is, the correlation coefficient squared (r^2). In the instance where there is a dependent variable and an independent variable, then the coefficient of determination indicates the amount of variance in the dependent variable that can be explained by the variance in the independent variable. Interpretation is specific to the circumstances. For example, a low coefficient of determination may be acceptable if you anticipate a relationship, but a higher value may be required if you need to use the independent variable to predict the dependent variable. There are no effect size statistics for correlated data; however, the standard error of the estimate or confidence intervals can be useful. Confidence intervals can provide added benefits and can be obtained through parametric confidence intervals or non-parametric bootstrapping (see Zhu, 1997, for a tutorial on bootstrapping). It has been suggested that confidence intervals can be used as an alternative method to effect size for supporting inferential statistics. Care should be taken in their interpretation as, like inferential statistics, confidence intervals are based on probabilities and suffer from some of the above problems.

To support an ANOVA (between-group design), two methods of calculating the explained variance are common: eta-squared (η^2) and omega-squared ($\omega_F{}^2$). Eta-squared (Vincent, 1999, p. 165) is the ratio of the treatment variance (SS_B) to total variance (SS_T). A more accurate measure is omega-squared, which will probably produce a smaller value (Vincent, 1999, p. 166), as it tries to account for the unexplained variance. For these tests, an effect size of 0.05 is small, 0.10 is intermediate and >0.20 is considered large (Cohen, 1988).

For a within-group ANOVA, although there are no formulae for explained variance, Park and Schutz (1999) provide – in their paper on calculating power for repeated-measures ANOVA – a formula for estimating effect size (see Table 5). In addition, as agreement quantifies the differences between repeat measures, we propose that agreement can be used as an equivalent test to effect size. The agreement boundary (LOA_B) provides a comparison for all measurements (BSI, 1979; Bland, 1995). Between pairs of means, the limits of agreement (LOA) can be used (Bland and Altman, 1986). Two benefits of agreement are that the output is in the same units of measurement as the data and that they are also easy to calculate. Results from agreement should be interpreted with respect to the specific circumstances and what is considered a meaningful magnitude (e.g. based on previous literature) or greater than measurement errors (e.g. identified through reliability testing).

We outlined earlier some effect sizes reported in the literature that indicate whether the differences between the data are large, medium or small (e.g. Cohen, 1988). Hopkins (1997) describes alternative interpretations of some effect size statistics: small (0.2), moderate (0.6), large (1.2), very large (2.0) and nearly perfect (4.0). The use of all these criteria has some limitations in that they are not individualized to the experiment. Hopkins *et al.* (1999) suggested that differences between elite performers are very small. As such, much smaller effect sizes might be meaningful in such circumstances. This example supports the view that, generally, interpretation should be based on *a priori* theorizing. This may involve identifying from previous literature what is an important difference theoretically, ethically, economically and practically, and considering whether the differences are bigger than limits identified for error, reliability and variability analyses. Using the values reported in the literature as a guide can be useful in identifying the size of the effect. Using effect size statistics should involve four steps: (1) identify the appropriate test based on the research design; (2) use *a priori* theorizing to identify acceptable results for interpretation; (3) collect data and test assumptions underpinning the effect size statistic, and correct for any violation; and (4) perform the correct test and interpret.

To summarize, in describing data, effect size statistics quantify the magnitude of the association between data in an experiment. The limitations of these statistics include: (a) there is no test for every experimental design; (b) the complexity of tests will increase with a perceived need for more tests and increased computer power; (c) the lack of literature suggests they will be used infrequently and, because the assumptions required are not detailed, they are likely to be misused; and (d) they are only descriptive, hence external validity and extrapolation is limited. The benefits, however, are that they: (a) are simple to calculate; (b) quantify the difference between groups; (c) provide results in the original units of measurement that are easier to interpret and can be applied practically; (d) can be used to support P-values, which alone can be misleading; and (e) can support qualitative data. Graphing and observing data can also be used to support all statistics, which alone can be misleading (Armstrong, 1987).

Trend analysis: kinetic, spatial and temporal data analysis

The previous section outlined the many factors that influence the variability of human movement and some of the implications for statistical design were discussed.

However, once it has been accepted that variability is inherent in all human movement, the question still remains: What techniques are available to the biomechanics and motor control researcher for analysing time-series data (e.g. kinematic, kinetic, electromyographic)? Traditionally, much research in these disciplines has been concerned with the relative timing and magnitude of discrete kinematic and kinetic variables. For example, work on javelin throwing has shown that the peak speed of distal segments is higher than that of proximal segments, elite athletes achieve higher peak speeds of distal segments than novices, and 50% of the javelin's final speed is generated during the 50 ms before release (Best *et al.*, 1993; Bartlett *et al.*, 1996; Morriss and Bartlett, 1996). However, since the generation of peak speeds in the javelin throw requires the transfer of momentum from the proximal to distal segments, it follows that coordination in either inter- or intra-segments is of crucial importance to successful performance. Coordination is the bringing together of the movement system components into proper relation with each other (see Turvey, 1990). Various qualitative and quantitative methods are used to analyse intra-segment (e.g. thigh with calf) and inter-segment (e.g. trunk with thigh) coordination.

For qualitative analysis, two methods involve the inspection and interpretation of position–time graphs and position–position graphs. With the former, the positions of pairs of segments are plotted as a function of time. Then, the researcher visually inspects the graph for signs of locking (i.e. a specific element of coordination with movement in the same or opposite direction, but not necessarily with equal magnitude) between the pairs of segments. Although the pair of segments to be plotted is based typically on intuition about which would be expected to be coordinated, this approach requires care as the coordination strategy is not always obvious (e.g. linearly related). An example of this approach is shown in Fig. 1, in which the lower and upper arm segment angles (calculated relative to the vertical axis) during an underarm throwing task are plotted as a function of time. It can be seen that the movements of the lower and upper arm are locked throughout the entire duration of an underarm throwing task. Although potentially useful for showing the extent of locking (i.e. giving an indication of the amount of coordination) between pairs of joints, the position–time plot does not provide a suitable means of depicting the variability in this coordination across trials. This is particularly evident when considering the number of pairs of joints, the number of trials and the number of participants who are typically analysed in an experiment.

The use of variable–variable plots (angle–angle plots, e.g. Vereijken *et al.*, 1992; velocity–position plots or phase–plane portraits, e.g. Schmidt and Lee, 1999, p. 34) provides a solution to this problem of depicting the variability in this coordination across trials. For example, plotting the position of one segment as a function of the position of a second segment for several trials on the same figure enables a visual inspection of the amount of variability and the pattern of coordination to be assessed. An example is shown in Fig. 2, in

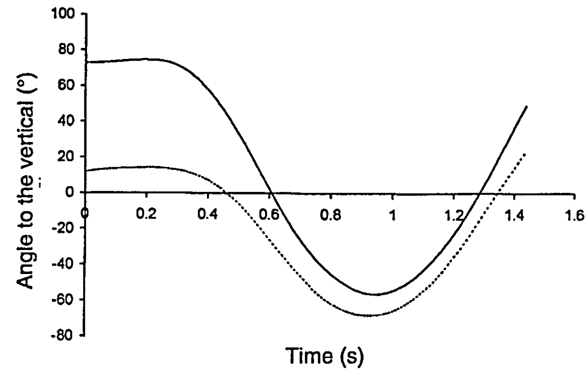


Fig. 1. Angle of the lower (solid line) and upper (dashed line) arm to the vertical as a function of time during an underarm throwing task (data from Al-Abood, 2001). In the anatomical standing position, both angles = 0°. Values are positive and negative when the limbs are anterior and posterior to this position, respectively.

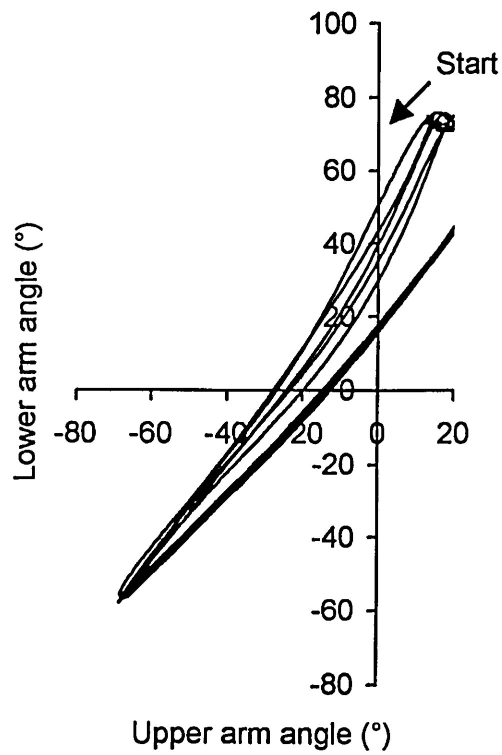


Fig. 2. Intra-limb angle–angle diagram of the lower and upper arm angles to the vertical over five trials for one participant performing an underarm throwing task (data from Al-Abood, 2001). See caption to Fig. 1 for definition of angles.

which the lower arm angle is plotted against the upper arm angle for five trials by the same person performing an underarm throwing task. With this simple movement, we can infer that the coordination strategy between the upper and lower arm is reflected by an essentially positive linear relationship. For each change in the lower arm angle, a similar change occurs in the direction and magnitude of the upper arm angle. In addition, we can see that the size and shape of the curve for each trial is relatively similar. This indicates that the lower and upper arm are coordinated such that they move in a similar direction and with a similar range of motion (i.e. small variability) across the five trials examined.

With a more complex coordinated movement, such as that of the hip and knee in running, a different pattern emerges. As can be seen in Fig. 3, the relationship between the angles of the hip and knee is essentially non-linear. Beginning at heel-strike, the hip angle extends then hyper-extends, while the knee angle tends to extend only slightly (heel-strike to toe-off). This is then followed by a period of hyper-flexion and flexion of the hip and flexion of the knee (toe-off to knee minimum). While the hip continues to flex, the knee begins to extend (knee minimum to hip maximum). During the final phase, the knee continues to extend while the hip extends (hip maximum to heel strike). It can also be seen that the size and shape of the curve for each trial is similar, again indicating that the movements are coordinated in a similar way and with a similar range of motion across the three trials.

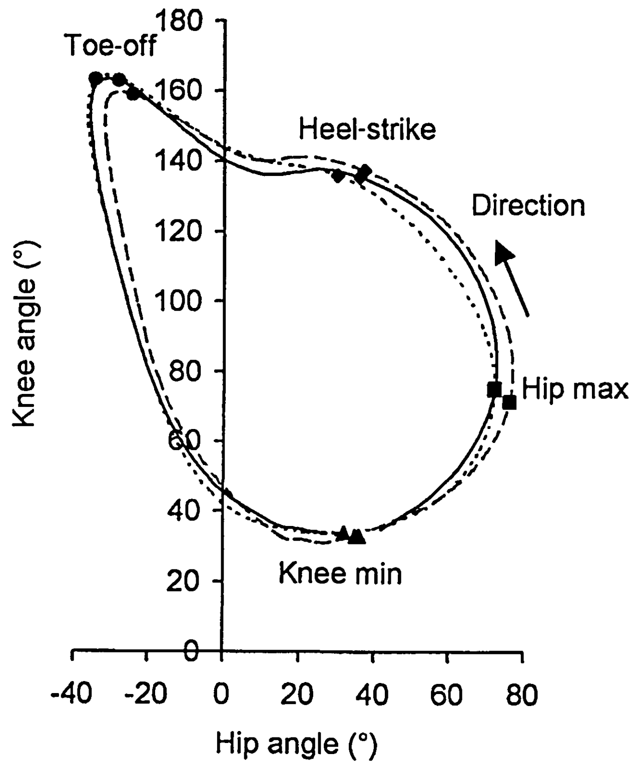


Fig. 3. Intra-limb angle–angle diagram of the hip and knee angles during running over three trials. In the anatomical standing position, the hip is at 0° and the knee is at 180°. Values are positive for hip flexion and negative for hip hyper-extension. The raw data for each of the three trials are contained in Appendix 1.

With qualitative analysis, as in the two examples above, one may discuss such vague concepts as the movements between pairs of joints being 'tightly coupled'. However, it is possible to perform a quantitative analysis on these data. To quantify the variability within a single variable history, several techniques are available as described in Table 6. In addition to assessing the variability of a single variable, intra- and inter-segment coordination can be assessed using a combination of two variables. For example, the ratio of hip and knee angles over a single displacement history could be used as the dependent variable in a coefficient

Table 6. Statistics used in the literature for quantifying variability of repeat trials (from Mullineaux, 2000)

Statistic[a]	Equation[b]	Excel formula[c]
s	$\sqrt{\sum_{i=1}^{n}(\bar{x}-x_i)^2/(n-1)}$	= STDEV(A1:C1)
RMSD	$\sqrt{\sum_{i=1}^{n}(x_C-x_i)^2/n}$	= STDEVP(A1:C1)
95%CI	$1.96s/\sqrt{n}$	= 1.96*STDEV(A1:C1)/COUNT(A1:C1)^0.5
%CV	$100s/\bar{x}$	= 100*STDEV(A1:C1)/AVERAGE(A1:C1)
%RMSD	$100\text{RMSD}/\sqrt{\sum_{i=1}^{n}(x_C)^2/n}$	= 100*STDEVP(A1:C1)/AVERAGE(A1:C1)

[a] Sample standard deviation (s); root mean-square difference (RMSD); 95% confidence intervals (95%CI); percentage coefficient of variation (%CV); percentage RMSD (%RMSD).
[b] Mean ($\bar{x}$); variable (x_i); sample or trial size (n); criterion value (x_C).
[c] Where no criterion exists, the mean value of the data is appropriate and the equations simplify to combinations of: s(STDEV); population standard deviation, σ(STDEVP); $\bar{x}$(AVERAGE); n(COUNT). The formulae provided for Excel 97 (Microsoft Corporation, Redmond, WA, USA) assume three trials with the data contained in cells A1, B1 and C1.

of variation equation. When there are multiple trials, the variability can be quantified at key times (e.g. at the moment of toe-off or heel-strike) using the techniques described in Table 6. Variability over the whole of several trials can be quantified for a single variable or the ratio of two variables across a normalized time interval using the same techniques. However, complex non-linear human movements may require the use of polynomial interpolation for the normalization procedure of the time to equal lengths over trials; an alternative is to analyse trials of equal time intervals (see Whitall and Caldwell, 1992).

All of the statistical techniques described in Table 6 are similar and, subsequently, provide results that are predictably different from each other (Mullineaux, 2000). For certain trial sizes, the magnitude of the variability is fixed for $n \leq 3$ (smallest to largest: RMSD, s, 95%CI), $n = 3$ (smallest to largest: RMSD, 95%CI, s) and $n \geq 3$ (smallest to largest: 95%CI, RMSD, s). For normalized techniques, the root mean-square difference (RMSD) provides a smaller value for variability than percentage coefficients of variation (%CV) for all trial sizes. This predictable nature is useful for comparison of findings to previous literature on variability. Furthermore, as the trial size is included in the denominator in the equation of all these techniques, the quantification of variability decreases predictably for increases in trial size. A table of these conversions for trial sizes ranging from 1 to 10 is provided by Mullineaux (2000); it is a useful resource for enabling variability between different studies to be compared. In general, Mullineaux (2000) recommends that, to quantify variability for small n, use the root mean-square difference and use normalized techniques only when the means are similar. To illustrate the use of the statistical techniques described in Table 4, they have been applied to the data illustrated in Fig. 3 to provide the results shown in Table 7. Hence, as $n = 3$, the root mean-square difference provides the smallest measure of variability. In interpreting the variability, for example, the root mean-square difference for the hip indicates that its movement is more variable at the key time of 'min knee' (4.11°) compared to at 'max hip' (1.89°). However, as the mean scores are not similar, the %RMSD provides contrary results, indicating that the movement in the hip is less variable at 'min knee' (2.57%) than at 'max hip' (13.12%), which emphasizes that normalized techniques should be used with care.

An emerging quantitative analysis method involves the use of cross-correlations. Cross-correlations are based on the assumption of a linear relationship between the dependent variables (e.g. pairs of joints), but do not assume that these variables change in synchrony during the movement. Rather, through the introduction of time lags, ranging from plus or minus one sample less than the number of data points (although Amblard *et al.*, 1994, recommend up to ± 7 to protect against type I errors, as a proportion of the correlation coefficient decreases with an increase in the number of lags), we can find high correlations between two variables in which there is a constant lag between them. For this reason, it has been suggested that cross-correlations are particularly suited to human movement, in which there are often time lags between

Table 7. Quantification of variability in hip and knee angles for running over three trials

Key time[a]	Variable[b]	Mean (°)	Variability: s (°)	RMSD (°)	95%CI (°)	%CV (%)	%RMSD (%)
Heel-strike	Hip : knee	0.25	0.03	0.02	0.03	10.78	8.81
	Hip	34.33	3.79	3.09	4.28	11.03	9.00
	Knee	136.33	0.58	0.47	0.65	0.42	0.35
Toe-off	Hip : knee	−0.18	0.03	0.02	0.03	−16.48	−13.46
	Hip	−28.67	5.03	4.11	5.70	−17.56	−14.34
	Knee	161.67	2.31	1.89	2.61	1.43	1.17
Max. hip	Hip : knee	1.00	0.06	0.05	0.07	6.40	5.22
	Hip	73.33	2.31	1.89	2.61	3.15	2.57
	Knee	73.67	2.31	1.89	2.61	3.13	2.56
Min. knee	Hip : knee	0.96	0.13	0.10	0.14	13.25	10.82
	Hip	31.33	5.03	4.11	5.70	16.06	13.12
	Knee	32.67	1.53	1.25	1.73	4.68	3.82

[a] Raw data for each key time are highlighted in bold in Appendix 1.
[b] No units for hip : knee ratios.

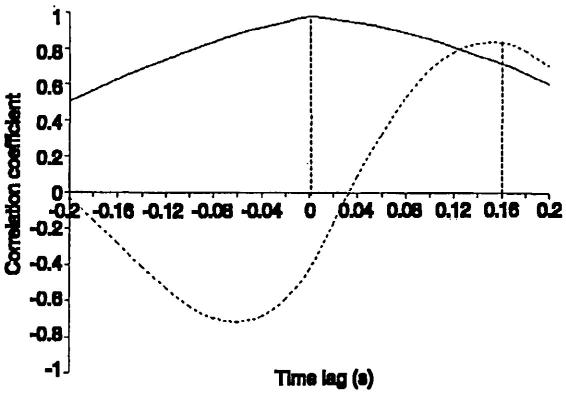


Fig. 4. Cross-correlation function between the lower and upper arm angles (solid line) and the hip and knee angles (dashed line). The data were sampled at 50 Hz and the time lags corresponding to the peak cross-correlations are indicated. See captions to Figs 1 and 3 for definitions of angles. The raw hip and knee angles are trial 3 presented in Appendix 1, and the cross-correlation values are provided in Appendix 2.

coordinated segments (Amblard *et al.*, 1994), to identify the time lag at which the peak correlation occurs. An example of a cross-correlation function computed on the angular displacement data of the lower and upper arm in Fig. 1 is shown in Fig. 4. It can be seen that there are high correlations across the entire data set because of the linear trend between the variables (Fig. 1), with a positive peak of 0.97 at a zero time lag. This indicates that the two segments do in fact move in synchrony and in the same direction. A negative peak would indicate that the variables are inversely linearly related (i.e. as one increases, the other would decrease).

The cross-correlation between the hip and knee angle data (shown in Fig. 3) for one trial is also shown in Fig. 4. In this instance, a negative lag indicates that the knee moves after the hip, whereas a positive lag would indicate that the knee moves before the hip. As there is a high positive peak ($r = 0.83$) at the eighth positive lag (at a sampling frequency of 50 Hz, this equates to +0.16 s), this suggests that the knee is linearly coordinated with the hip when the knee moves before the hip by 0.16 s.

The results from a cross-correlation can be subjected to statistical analysis. A simple analysis has been proposed by Li and Caldwell (1999), where the phase shifting is significant if the peak correlation is greater than the 95% confidence intervals for the correlation at zero lag. However, Grimm (1993) proposed that, before statistically analysing the data obtained from the whole cross-correlation function, we may first need to compute the Z-transform, as we would not expect the cross-correlation coefficients to be normally distributed. We can then test if any of the correlation coefficients across the range of lags is different from zero lag (Amblard *et al.*, 1994) using coefficient data averaged over individuals or trials in a repeated-measures ANOVA. Because the probability of type I errors increases with the number of lags, supporting the recommendation of time lags up to ± 7 (Amblard *et al.*, 1994), in practice it is more common to establish if there are any differences in the correlation coefficients at zero lag for a given pair of variables (see, for example, Vereijken *et al.*, 1992; Whiting and Vereijken, 1993). Alternatively, testing for differences in the cross-correlation coefficients between the peaks or between the time lags for the peaks could be used to assess differences. Although these approaches may be appropriate for examining the effects of an experimental manipulation on the coordination at this specific lag, caution should be taken when suggesting that the movements of joint pairs become less correlated. It could quite simply be the case that the correlation at the specific time lag changes while the actual peak correlation value remains unaffected, although it occurs at a different time lag.

Recently, however, some concern has been expressed about the use of cross-correlations with human movement data because of the assumption of linearity in correlation analysis (Amblard *et al.*, 1994; Sidaway *et al.*, 1995). In a cross-correlation function, the time lag of a peak is an average estimate over the whole of the period analysed and, therefore, relies on the relationship between the two variables being linear throughout the range of data. When a significantly large part of the time interval contains a linear relationship between two variables, it follows that high cross-correlation coefficients will be obtained, but it is not clear how much of a given time interval between two variables needs to be linearly related.

To confirm that the linearity assumption is met, which is necessary to improve the validity of the analysis, the relationship between the two variables when the time lag is included on one variable can be qualitatively assessed by plotting the data against each other in a scatterplot. For example, in Fig. 5, the knee angle data are plotted against the hip angle data without any lag and against the hip angle data with the +0.16 s lag included. It is clear that there is a linear trend between the data when the time lag is included, but still the relationship is evidently more complicated than a simple linear relationship. This results in a lower peak cross-correlation for the hip and knee angles ($r = 0.827$) than for the lower and upper arm angles data illustrated in Fig. 2 ($r = 0.973$). It could, therefore, be interpreted that greater coordination existed for the lower and upper arm angles, simply because these more linearly related data more closely meet the linearity assumption of the cross-correlation. When visually inspecting the hip–knee angle–angle plots for several trials (see Fig. 3), it is

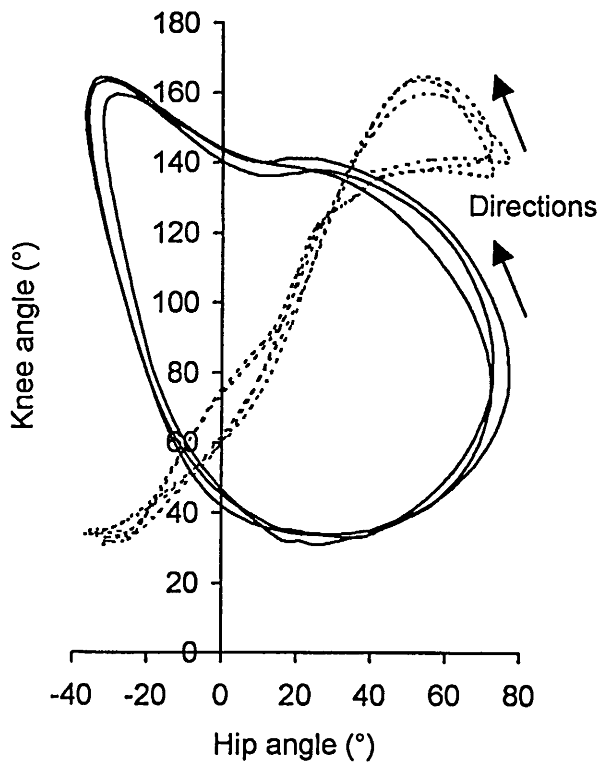


Fig. 5. Intra-limb angle–angle diagram of the hip and knee angles during running over three trials with no lag (solid line) and a lag of 0.16 s in the hip angle (dashed line). See caption to Fig. 3 for definitions of angles. The raw data with no lag is contained in Appendix 1 (trial 3).

apparent, however, that the non-linear relationship that exists between the two variables is consistent. Hence, it would be more accurate to assume that the knee and hip movements are coordinated, but in a non-linear relationship that is performed consistently across trials that cross-correlations are unable to detect. For non-linear data, a transformation could be applied that linearizes the data (e.g. log–log transformations; see Snedecor and Cochran, 1989, for further details on types of transformations) and then cross-correlations could be applied. If the cross-correlations are small, then it is possible that the transformation is still non-linear and a more complex relationship may exist between the variables.

The limitation of cross-correlation analysis is particularly relevant when comparing the changes in coordination tendencies that occur with practice. Sidaway *et al.* (1995) examined the differences in coordination between the angles of the left and right knee on a ski-simulator task with experts and novices. Novices with 3 days of practice exhibited a somewhat negative linear relationship between the left and right knee angles that yielded a cross-correlation coefficient of −0.83. Experts exhibited a non-linear relationship between the left and right knee angles that yielded a cross-correlation coefficient of −0.22. Based on the cross-correlation coefficients alone, it would be tempting to conclude that less coordination existed between the left and right knees in experts. However, observation of the angle–angle plots showed that the non-linear relationship was performed with high consistency across trials and hence the angles were highly coordinated. The proposal by Sidaway *et al.* (1995) to use the normalized root mean-square (NoRMS) error may provide an alternative method to examine the consistency in linear and non-linear patterns of coordination (see equation 4):

$$\text{NoRMS} = 100 \sum_{j=1}^{k} \sqrt{\sum_{i=1}^{n} (\bar{x}_A - x_{Ai})^2 + (\bar{x}_B - x_{Bi})^2 / n_j} \Big/ kR \quad (4)$$

where A and B denote the two variables, R is the resultant excursion as defined in Appendix 1, and the remaining notation is described in Table 6. An example of the use of NoRMS is illustrated in Appendix 1.

In general, the implication is that, in the case of human movement in which the relationship between the two variables is non-linear in a significant portion of the time interval, the use of cross-correlation analysis might need to be extended to involve some form of transformation.

In this section, we have discussed some of the advantages and disadvantages of the qualitative and quantitative methods of analysing intra-segment and inter-segment coordination data. We have shown that the suitability of a particular method will depend on the relationship between the variables analysed (i.e. linear or non-linear) and the type of coordination between the variables being assessed (i.e. continuously over the duration of the movement or discretely at key times during the movement). In using cross-correlations to assess coordination, we recommend that a scatter-plot of the data with the identified time lag is inspected for meeting the linearity assumption and, where this is violated, an appropriate linearity transformation may need to be applied first. To quantify variability, Mullineaux (2000) has recommended the root mean-square difference for small trial sizes and normalized techniques only when the means of different variables are similar.

Reporting research

Many issues have to be addressed if research is to be reported well. A good research report should be intelligible to its reader, leaving that person to concentrate on the research itself and not the way it is presented. Most authorities on good writing style, and most scientific journal editors, would agree that many scientific papers submitted for review are poorly written. The use of incorrect punctuation and syntax, too much jargon and overuse of abbreviations can make research papers difficult to read. Any writer of a scientific report is recommended to read a good text on scientific writing, such as Palmer (1993), O'Connor (1991) or Day (1995). It is neither the intention of this section to enter into a diatribe about poor scientific writing, nor to cover the reporting of empirical biomechanical research, which are dealt with in the various chapters of Bartlett (1997). Instead, in this section, we briefly focus on the reporting of issues that have been highlighted in this review.

The pressure on space in scientific journals does not permit full details to be presented in any paper. Sufficient details should be provided to enable the reader to understand the theoretical underpinning of the research question, use the findings to support future research, replicate the study and interpret the findings. We illustrated earlier examples where data were used to support a theory (e.g. Scott *et al.*, 1997) or to develop a theory (e.g. Blanksby *et al.*, 1998). However, in replicating a study, Table 3 illustrated that the criteria for selecting (11.5%), or methods for treating (19.2%), trials were not clear in a selection of biomechanics papers. In interpreting findings, providing details on the assessment of the assumptions underpinning the statistical tests is necessary, otherwise doubt will be cast on the validity of the findings.

A judicious use of reference to standard procedures and statistical analyses provides one solution to the pressure on space. The writer might well consider that researchers in biomechanics and motor control are normally well-versed in empirical protocols (a strength of the disciplines) but not in the selection and use of the best statistical analysis (a weakness of the disciplines). Rather more emphasis might, therefore, be placed on the latter than the former, for which accurate referencing to standard sources might be adequate. This applies in particular when the researcher is using an analysis tool that is relatively new to the discipline, for example cross-correlations of time series (see section on 'Trend analysis'). Clearly, accurate referencing is very important to the scientific rigour and quality of the final report (e.g. Morrow, 1991; Stull *et al.*, 1991).

Sufficient information needs to be provided on the conduct of the research that enables the reader to assess the validity of the data provided. Providing information on methods to control for experimental errors (see Table 1) and details on the sample characteristics and selection are important. In addition, clarifying the timing, number and sequence of measurements, including details on trial size, is important for other issues such as statistical power (see section on 'Trial sizes').

The choice of statistical test should be based on whether it will provide any valid and meaningful results that can be used to support the interpretation of the data. Descriptive statistics are recommended in most cases, as they provide simple data of practical value. When using statistical significance testing, there are some issues regarding the validity of such tests. Nevertheless, their use will continue and we suggest that when statistical significance tests provide useful information, then the researcher should justify the level of statistical significance (α), the probability (P) for each comparison made and the power of the tests used (see Cohen, 1988, for methods of calculating power). The use of effect size may assist in this issue; hence in biomechanics and motor control research the use of these are recommended either to supplement or supplant significance tests.

When using any statistical test, it is advisable that the justification for the chosen methods of data analysis is fully reported. The writer should give due attention to the appropriateness or otherwise of any statistical tests used, and whether the data satisfied the assumptions of the particular tests. In many scientific papers, the assumptions about the data underpinning parametric statistical tests are often not satisfied or corrected for. Many standard statistical texts now provide simple explanations of the underlying assumptions of various tests and the consequences of their violation (see, for example, Howell, 1997). Many statistical packages provide checks on and methods to treat violations of the assumptions of particular tests, but generally neither the checks nor the adjustments are automatically carried out by these software packages. When using cross-correlations, we recommend that checks on the testing of the linearity assumption are reported.

Variability can be inherent (i.e. inability to exactly replicate a movement) or functional (i.e. used to control the outcome of movements). When interpreting this variability, it is important that the researcher attends to all of the issues that affect the accuracy of the measurements made, and provides estimates of the measurement 'errors' or uncertainties. How these uncertainty estimates were arrived at should be clearly stated and justified. The propagation of these uncertainties in all calculations of important variables or parameter values should be assessed (see, for example, Challis, 1997). Reliability and objectivity checks should also be reported (see section on 'Trial sizes'), together with all calibration procedures and any other attempts to ascertain the validity of any measurements.

Presenting information in the most appropriate format is also crucial to good report writing. Presenting data as either a chart, if a visual inspection is required, or a table, if exact values are important, is appropriate, but duplication should be avoided. There are many examples of incorrectly or poorly presented data. For example, presenting dichotomous relationships by means of line graphs – which imply a continuous relationship – rather than, correctly, bar charts, is quite common. Also, three-dimensional graphical representations of data are now easily accessed through most computer graphics packages, but these can cause confusion rather than keeping the presentation simple, which is recommended (see Day, 1995, for further examples).

Generally, statistical analyses should not be spuriously used to tempt the researcher into 'stretching' the research, especially the discussion of the results and their meaning. Instead, the report should concentrate on the research question that was addressed, the underlying theory, previous related research, the importance of the results of the new study – and how they contribute to existing knowledge – and a frank assessment of the limitations of the study, including the data analyses used.

Conclusions

The research methods and statistical techniques available to the researcher are continually on the increase. The aim of this review has been to highlight some pertinent and contemporary issues associated with planning, conducting, analysing and reporting biomechanics and motor control research. In planning research, the researcher should first define a research question, the answer to which would be expected to contribute to knowledge. Secondly, the workability of the methods should be considered as to whether the results could be obtained in a format suitable to answer the research question using the best methods available. In conducting research, inter- and intra-individual variability can affect the results differently. We recommend including, reporting or discussing some assessment of variability and including more than one trial where possible. In analysing data, effect size statistics can provide informative and simple analyses of results, and may be beneficial in supporting *P*-values that alone can be misleading. In analysing trends, cross-correlations provide a means to test intra- and inter-segment coordination. Reporting research should be in an accurate and comprehensive style, demonstrating the theoretical foundations and including an assessment of the limitations of the study and analyses. Research methods and statistics adopted in biomechanics and motor control are unlikely to change much in the near future, although many of the issues touched upon are developing. Carefully checking or correcting for violations of underpinning assumptions in statistics and reporting the effect of these are useful for improving the quality of future research.

References

Al-Abood, S. (2001). Effects of visual demonstrations on motor skill acquisition: a visual perception perspective. Unpublished doctoral dissertation, Manchester Metropolitan University, Alsager.

Amblard, B., Assaiante, C., Lekhel, H. and Marchand, A.R. (1994). A statistical approach to sensorimotor strategies: conjugate cross-correlations. *Journal of Motor Behavior*, 26, 103–112.

Armstrong, R.L. (1987). The use of statistical significance: a commentary related to Franks and Huck. *Research Quarterly for Exercise and Sport*, 58, 81–82.

Arutyunyan, G.H., Gurfinkel, V.S. and Mirskii, M.L. (1968). Investigation of aiming at a target. *Biophysics*, 13, 536–538.

Atkinson, G. and Nevill, A.M. (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Medicine*, 26, 217–238.

Barrentine, S.W., Matsuo, T., Escamilla, R.F., Fleisig, G.S. and Andrews, J.R. (1998). Kinematic analysis of the wrist and forearm during baseball pitching. *Journal of Applied Biomechanics*, 14, 24–39.

Bartlett, R.M. (ed.) (1997). *Biomechanical Analysis of Movement in Sport and Exercise*. Leeds: British Association of Sport and Exercise Sciences.

Bartlett, R., Müller, E., Lindinger, S., Brunner, F. and Morriss, C. (1996). Three-dimensional evaluation of the kinematic release parameters for javelin throwers of different skill levels. *Journal of Applied Biomechanics*, 12, 58–71.

BASES (1995). *Code of Conduct*. Leeds: British Association of Sport and Exercise Sciences.

Bates, B.T. (1996). Single-subject methodology: an alternative approach. *Medicine and Science in Sports and Exercise*, 28, 631–638.

Bates, B.T., Dufek, J.S. and Davis, H.P. (1992). The effect of trial size on statistical power. *Medicine and Science in Sports and Exercise*, 24, 1059–1068.

Baumgartner, T.A. and Strong, C.H. (1998). *Conducting and Reading Research in Health and Human Performance*. Boston, MA: McGraw-Hill.

Best, B., Bartlett, R.M. and Morriss, C.J. (1993). A three-dimensional analysis of javelin throwing technique. *Journal of Sports Sciences*, 11, 315–328.

BIDS ISI Data Service (1998). Science Citation Index. http://www.bids.ac.uk (accessed 30 June 1999).

Bland, J.M. (1995). *An Introduction to Medical Statistics*. Oxford: Oxford University Press.

Bland, J.M. and Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, i, 307–310.

Blanksby, B.A., Simpson, J.R., Elliott, B.C. and McElroy, K. (1998). Biomechanical factors influencing breaststroke turns by age-group swimmers. *Journal of Applied Biomechanics*, 14, 180–189.

Bootsma, R.J. and Van Wieringen, P.C.W. (1990). Timing an attacking forehand drive in table tennis. *Journal of Experimental Psychology: Human Perception and Performance*, 16, 21–29.

Bootsma, R.J., Houbiers, M.H.J., Whiting, H.T.A. and Van Wieringen, P.C.W. (1991). Acquiring an attacking forehand drive: the effects of static and dynamic environmental conditions. *Research Quarterly for Exercise and Sport*, 62, 276–284.

Borenstein, M. (1997). Hypothesis testing and effect size estimation in clinical trials. *Annals of Allergy, Asthma and Immunology*, 78, 5–11.

BSI (1979). *Precision of Test Methods. I: Guide for the Determination and Reproducibility of a Standard Test Method*, BS 5497, part 1. London: British Standards Institute.

Burden, A.M., Grimshaw, P.N. and Wallace, E.S. (1998). Hip and shoulder rotations during the golf swing of sub-10 handicap players. *Journal of Sports Sciences*, 16, 165–176.

Caldwell, G.E., Li, L., McCole, S.D. and Hagber, J.M. (1998). Pedal and crank kinetics in uphill cycling. *Journal of Applied Biomechanics*, 14, 245–259.

Carver, R.P. (1978). The case against statistical significance testing. *Harvard Educational Review*, 48, 378–399.

Challis, J.H. (1997). Estimation and propagation of experimental errors. In *Biomechanical Analysis of Movement in Sport and Exercise* (edited by R.M. Bartlett), pp. 105–124. Leeds: British Association of Sport and Exercise Sciences.

Chow, S.L. (1998). Prècis of statistical significance: rationale, validity, and utility. *Behavioral and Brain Sciences*, 21, 169–239.

Cohen, J. (1969). *Statistical Power Analysis for the Behavioural Sciences*. San Diego, CA: Academic Press.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioural Sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155–159.

Coolican, H. (1999). *Research Methods and Statistics in Psychology*. London: Hodder & Stoughton.

Day, R.A. (1995). *How to Write and Publish a Scientific Paper*. Cambridge: Cambridge University Press.

Dufek, J.S., Bates, B.T. and Davis, H.P. (1995). The effect of trial size and variability on statistical power. *Medicine and Science in Sports and Exercise*, 27, 288–295.

Goosey, V.L., Campbell, I.G. and Fowler, N.E. (1998). The relationship between three-dimensional wheelchair propulsion techniques and pushing economy. *Journal of Applied Biomechanics*, 14, 412–427.

Gould, S.J. (1981). *The Mismeasure of Man*. London: Penguin.

Grimm, L.G. (1993). *Statistical Applications for the Behavioral Sciences*. Singapore: John Wiley.

Hawking, S. (1996). *The Illustrated a Brief History of Time*. London: Transworld Publishers.

Hopkins, W.G. (1997). A scale of magnitudes for effect statistics. http://www.sportsci.org/resource/stats/effectmag.html (last updated 9 June 1997).

Hopkins, W.G., Hawley, J.A. and Burke, L.M. (1999). Design and analysis of research on sport performance enhancement. *Medicine and Science in Sports and Exercise*, 31, 472–485.

Howell, D.C. (1997). *Statistical Methods for Psychology*. Belmont, CA: Duxbury Press.

Hubbard, A.W. and Seng, C.N. (1954). Visual movements of batters. *Research Quarterly*, 25, 42–57.

James, C.R. and Bates, B.T. (1997). Experimental and statistical design issues in human movement research. *Measurement in Physical Education and Exercise Science*, 1, 55–69.

Kraemer, J.C. and Thiemann, S. (1987). *How Many Subjects? Statistical Power Analysis in Research*. London: Sage.

Kroll, W. (1957). Reliability theory and research decision in selection of a criterion score. *Research Quarterly*, 38, 412–419.

Li, L. and Caldwell, G.E. (1999). Coefficient of cross-correlation and the time domain correspondence. *Journal of Electromyography and Kinesiology*, 9, 385–389.

Ludbrook, J. (1997). Comparing methods of measurement. *Clinical and Experimental Pharmacology and Physiology*, 24, 193–203.

Matthews, R.A.J. (1998a). Flukes and flaws. *Prospect*, November, pp. 20–24.

Matthews, R.A.J. (1998b). *Facts Versus Factions: The Use and Abuse of Subjectivity in Scientific Research*. Cambridge: European Science and Environment Forum.

Michaels, C.F. and Beek, P. (1996). The state of ecological psychology. *Ecological Psychology*, 7, 259–278.

Morriss, C. and Bartlett, R. (1996). Biomechanical factors critical for performance in the men's javelin throw. *Sports Medicine*, 21, 438–446.

Morrow, J.R. (1991). Editor's viewpoint: shocking. *Research Quarterly for Exercise and Sport*, 62, v.

Mullineaux, D.R. (2000). Methods for quantifying the variability in data. In *Proceedings of the 18th International Symposium of Biomechanics in Sport* (edited by Y. Hong and D.P. Johns), pp. 684–688. Hong Kong: The Chinese University of Hong Kong.

Mullineaux, D.R. and Bartlett, R.M. (1997). Research methods and statistics. In *Biomechanical Analysis of Movement in Sport and Exercise* (edited by R.M. Bartlett), pp. 81–104. Leeds: British Association of Sport and Exercise Sciences.

Mullineaux, D.R., Barnes, C.A. and Batterham, A.M. (1999). Assessment of bias in comparing measurements: a reliability example. *Measurement in Physical Education and Exercise Science*, 4, 195–205.

Newell, K.M. and Corcos, D.M. (1993). Issues in variability and motor control. In *Variability and Motor Control* (edited by K.M. Newell and D.M. Corcos), pp. 1–12. Champaign, IL: Human Kinetics.

O'Brien, K.F. and Israel, R.G. (1987). Response to: Why does everyone use the .05 level of significance? *Research Quarterly for Exercise and Sport*, 58, 85–86.

O'Connor, M. (1991). *Writing Successfully in Science*. London: Chapman & Hall.

Ottenbacher, K.J. and Barrett, K.A. (1991). Measures of effect size in the reporting of rehabilitation research. *American Journal of Physical Medicine and Rehabilitation*, 70, S131–S137.

Palmer, R. (1993). *Write in Style*. London: E & FN Spon.

Park, I. and Schutz, R.W. (1999). 'Quick and easy' formulae for approximating statistical power in repeated measures ANOVA. *Measurement in Physical Education and Exercise Science*, 4, 249–270.

Rankin, G. and Stokes, M. (1998). Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. *Clinical Rehabilitation*, 12, 187–199.

Reboussin, D.M. and Morgan, T.M. (1996). Statistical considerations in the use and analysis of single-subject designs. *Medicine and Science in Sports and Exercise*, **28**, 639–644.

Safrit, M.J. and Wood, T.M. (1995). *Introduction to Measurement in Physical Education and Exercise Science.* St. Louis, MO: Mosby.

Salo, A. and Grimshaw, P.N. (1998). An examination of kinematic variability of motion analysis in sprint hurdles. *Journal of Applied Biomechanics*, **14**, 211–222.

Salo, A., Grimshaw, P.N. and Viitasalo, J.T. (1997). Reliability of variables in the kinematic analysis of sprint hurdles. *Medicine and Science in Sports and Exercise*, **29**, 383–389.

Schmidt, R.A. and Lee, D.A. (1999). *Motor Control and Learning: A Behavioral Emphasis.* Champaign, IL: Human Kinetics.

Scott, M.A., Li, F.-X. and Davids, K. (1997). Expertise and the regulation of gait in the approach phase of the long jump. *Journal of Sports Sciences*, **15**, 597–605.

Shultz, B.B. and Sands, W.A. (1995). Understanding measurement concepts and statistical procedures. In *Physiology Assessment of Human Fitness* (edited by P.J. Maud and C. Foster), pp. 257–287. Champaign, IL: Human Kinetics.

Sidaway, B., Heise, G. and Schonfelder-Zohdi, B. (1995). Quantifying the variability of angle–angle plots. *Journal of Human Movement Studies*, **29**, 181–197.

Sim, J. and Reid, N. (1999). Statistical inference by confidence intervals: issues of interpretation and utilization. *Physical Therapy*, **79**, 186–195.

Snedecor, G.W. and Cochran, W.G. (1989). *Statistical Methods.* Ames, IA: Iowa State University Press.

Stull, C.A., Christina, R.W. and Quinn, S.A. (1991). Accuracy of references in Research Quarterly for Exercise and Sport. *Research Quarterly for Exercise and Sport*, **62**, 245–248.

Temprado, J.J., Della-Grasta, M., Farrell, M. and Laurent, M. (1997). A novice–expert comparison of (intra-limb) coordination subserving the volleyball serve. *Human Movement Sciences*, **16**, 653–676.

Thomas, J.R. and Nelson, J.K. (1996). *Research Methods in Physical Activity.* Champaign, IL: Human Kinetics.

Tuckman, B.W. (1978). *Conducting Educational Research.* New York: Harcourt Brace Jovanovich.

Turvey, M.T. (1990). Coordination. *American Psychologist*, **45**, 938–953.

Vereijken, B., Van Emmerik, R.E.A., Whiting, H.T.A. and Newell, K. (1992). Free(z)ing degrees of freedom in skill acquisition. *Journal of Motor Behavior*, **24**, 133–142.

Vincent, W.V. (1999). *Statistics in Kinesiology.* Champaign, IL: Human Kinetics.

Wainer, H. and Thissen, D. (1981). Graphical data analysis. *Annual Review of Psychology*, **32**, 191–241.

Web of Science (2000). Science Citation Index expanded. http://wos.mimas.ac.uk (last updated 12 August 2000).

Whitall, J. and Caldwell, G.E. (1992). Co-ordination of symmetrical and asymmetrical human gait. *Journal of Motor Behavior*, **24**, 339–353.

Whiting, H.T.A. and Vereijken, B. (1993). The acquisition of co-ordination in skill learning. *International Journal of Sport Psychology*, **24**, 343–357.

Young, R.P. and Marteniuk, R.G. (1997). Acquisition of a multi-articular kicking task: jerk analysis demonstrates movements do not become smoother with learning. *Human Movement Science*, **16**, 677–701.

Zhu, W. (1997). Making bootstrap statistical inferences: a tutorial. *Research Quarterly for Exercise and Sport*, **68**, 44–55.

Appendix 1

Hip and knee angles data for running over three trials and calculation of NoRMS (Sidaway *et al.*, 1995) using Excel 97 (Microsoft Corporation, Redmond, WA, USA). Units in degrees unless indicated; NoRMS = AVERAGE(J29:L29)/L32*100 = 3.7%

	A	B	C	D	E	F	G	H	I	J	K	L
1	Time (s)	Trial 1		Trial 2		Trial 3		Mean trace		Resultant		
2		Hip	Knee	Hip	Knee	Hip	Knee	Hip	Knee	Trial 1		Trial 2
3	0.00	40	35	42	35	42	35	41.3	35.0	1.8	0.4	0.4
4	0.02	56	43	55	44	58	44	56.3	43.7	0.6	1.9	2.9
5	0.04	69	55	66	58	68	58	67.7	57.0	5.8	3.8	1.1
6	0.06	**76**	**71**	**72**	**75**	**72**	**75**	73.3	73.7	14.2	3.6	3.6
7	0.08	76	90	72	93	69	93	72.3	92.0	17.4	1.1	12.1
8	0.10	69	108	65	110	60	109	64.7	109.0	19.8	1.1	21.8
9	0.12	59	122	56	122	50	122	55.0	122.0	16.0	1.0	25.0
10	0.14	48	131	46	130	40	131	44.7	130.7	11.2	2.2	21.9
11	0.16	**37**	**137**	**36**	**136**	**30**	**136**	34.3	136.3	7.6	2.9	18.9
12	0.18	27	141	26	138	20	138	24.3	139.0	11.1	3.8	19.8
13	0.20	18	141	18	137	10	140	15.3	139.3	9.9	12.6	28.9
14	0.22	10	140	10	136	−2	145	6.0	140.3	16.1	34.8	85.8
15	0.24	1	143	−2	142	−15	154	−5.3	146.3	51.2	29.9	152.2
16	0.26	−12	151	−16	153	−26	163	−18.0	155.7	57.8	11.1	117.8
17	0.28	**−24**	**159**	**−28**	**163**	**−34**	**163**	−28.7	161.7	28.9	2.2	30.2
18	0.30	−31	157	−35	160	−37	150	−34.3	155.7	12.9	19.2	39.2
19	0.32	−31	144	−35	143	−33	127	−33.0	138.0	40.0	29.0	121.0
20	0.34	−28	122	−30	118	−26	101	−28.0	113.7	69.4	22.8	164.4
21	0.36	−23	99	−24	93	−19	78	−22.0	90.0	82.0	13.0	153.0
22	0.38	−18	79	−18	73	−12	59	−16.0	70.3	79.1	11.1	144.4
23	0.40	−10	61	−10	58	−3	45	−7.7	54.7	45.6	16.6	115.2
24	0.42	1	46	0	46	8	37	3.0	43.0	13.0	18.0	61.0
25	0.44	15	33	14	36	21	34	16.7	34.3	4.6	9.9	18.9
26	0.46	21	32	20	35	25	34	22.0	33.7	3.8	5.8	9.1
27	0.48	**26**	**31**	25	34	**32**	**34**	27.7	33.0	6.8	8.1	19.8
28	0.50	35	33	**36**	**33**	40	35	37.0	33.7	4.4	1.4	10.8
29	√Mean									4.9	3.2	7.3
30	Max	76	159	72	163	72	163	73.3	161.7	82.0	34.8	164.4
31	Min	−31	31	−35	33	−37	34	−34.3	33.0	0.6	0.4	0.4
32												140.5

Mean trace is the average of the three trials' hip or knee angle coordinates at each time instance (e.g. H3 = AVERAGE(B3,D3,F3); I3 = AVERAGE(C3,E3,G3)).

Resultant is the absolute distance between each trial's hip and knee coordinates and the mean trace (e.g. J3 = (B3 − $H3)^2 + (C3 − $I3)^2; K3 = (D3 − $H3)^2 + (E3 − $I3)^2; L3 = (F3 − $H3)^2 + (G3 − $I3)^2).

Rows 29 to 31 provide, respectively, the square root of the mean (√Mean), maximum (Max) and minimum (Min) scores of the data located in rows 3–28 for each column presented.

L32 = SQRT((H17 − H27)^2 + (I17 − I27)^2). This is *R*, the 'resultant excursion of the mean angle–angle curve over the entire cycle' (Sidaway *et al.*, 1995, p. 188), estimated using the hip and knee data corresponding to the maximum and minimum values at either the hip or knee, depending on which had the greatest range. The data for the maximum and minimum values for the hip were used, as the range for the hip was larger (161.8 to 33.0 = 128.8) than for the knee (73.4 to −34.3 = 107.7), hence L32 = SQRT((161.8 − 33.0)^2 + (−29.0 − 27.7)^2).

Bold numbers indicate key times: maximum hip (0.06 s); heel-strike (0.16 s); toe-off (0.28 s); minimum knee (0.48 s for trials 1 and 3 and 0.50 s for trial 2).

Appendix 2

Cross-correlation data for Fig. 4

Time lag	Lower/upper arm (95%CI)	Hip/knee (95%CI)*
−0.20	0.504 (0.247)	−0.035 (0.490)
−0.18	0.567 (0.245)	−0.154 (0.476)
−0.16	0.627 (0.243)	−0.283 (0.463)
−0.14	0.684 (0.241)	−0.412 (0.449)
−0.12	0.738 (0.239)	−0.530 (0.439)
−0.10	0.788 (0.237)	−0.628 (0.427)
−0.08	0.835 (0.235)	−0.695 (0.417)
−0.06	0.877 (0.235)	−0.718 (0.410)
−0.04	0.914 (0.233)	−0.685 (0.400)
−0.02	0.946 (0.231)	−0.588 (0.392)
0	0.973 (0.229)	−0.426 (0.384)
0.02	0.959 (0.231)	−0.168 (0.392)
0.04	0.939 (0.233)	0.087 (0.400)
0.06	0.913 (0.235)	0.322 (0.410)
0.08	0.882 (0.235)	0.527 (0.417)
0.10	0.846 (0.237)	0.679 (0.427)
0.12	0.805 (0.239)	0.778 (0.439)
0.14	0.759 (0.241)	0.825 (0.449)
0.16	0.710 (0.243)	0.827 (0.463)
0.18	0.657 (0.245)	0.783 (0.476)
0.20	0.602 (0.247)	0.695 (0.490)

* Calculated for the Trial 3 data presented in Appendix 1.

Appendix 15

Nevill, A.M., Atkinson, G. and Mullineaux, D.R. (2001). Editorial: new horizons in research methods. *Journal of Sports Sciences*, **19**, 737-8.

Journal of Sports Sciences, 2001, 19, 737–738

Editorial

New horizons in research methods

This special issue was conceived in the knowledge that the range and variety of methods used to conduct sport science research are continuously developing. As such, there is a need to identify and report regularly contemporary trends, as well as good practice, in research methods adopted in Biomechanics, Physiology, Psychology and Sports Performance. With this aim in mind, the *Journal of Sports Sciences* invited leading sport scientists in the UK, who specialize in research methods, to write articles identifying the key issues, themes and trends associated with research within their particular fields. Consequently, this issue contains four papers to inform readers of the most important and relevant issues in research methods. However, readers are strongly advised not to read only the paper most closely associated with their own specialist area of research, but to read all four papers in this special issue, since some of the methods and topics discussed are common.

The paper by Mullineaux and co-workers highlights the pertinent and contemporary issues associated with the planning, conducting, analysing and reporting of research in biomechanics and motor control. The importance of planning, by defining a research question expected to contribute to knowledge, is emphasized. The authors go on to discuss the importance of reporting inter- and intra-individual variability, together with the advice to include the results from more than one trial wherever possible. In describing the results of biomechanics and motor control research, the authors recommend reporting effect size statistics as well as *P*-values, which, when reported alone, can be misleading. When analysing trends, the authors recommend using cross-correlations as a means to test intra- and inter-segment coordination. The authors also recommend the careful checking and, if necessary, correcting of violations of the underpinning assumptions in statistical analyses, a theme echoed in a recent editorial (Nevill, 2000); advice likely to improve and advance the future quality of research in biomechanics and motor control.

The paper on research methods in physiology of exercise and kinanthropometry by Winter and co-workers emphasizes the importance of well-designed studies and appropriately analysed results. The authors acknowledge that, with the recent advances in personal computing and the availability of statistical software, increased opportunities to investigate data with increasing sophistication are now available. At the same time, the ease with which such analyses can be performed can mask underlying philosophical and epistemological shortcomings. Winter *et al.* examine in detail the use of four techniques that are especially relevant to physiological studies: bivariate correlation and linear and non-linear regression; multiple regression; repeated-measures analysis of variance; and multi-level modelling. The authors stress the importance of adhering to underlying statistical assumptions and ways to accommodate violations when these assumptions are not met, which is an important and re-occurring theme that appears in all four papers in this issue.

The wide-ranging paper on aspects of research methods in sport and exercise psychology by Biddle and co-workers is organized around the major themes of quantitative and qualitative research. The authors highlight areas that can be problematic or controversial (e.g. stepwise statistical procedures), underused (e.g. discriminant analysis), increasingly used (e.g. meta-analysis, structural equation modelling, qualitative content analysis) and emergent (e.g. realist tales of writing). Perspectives range from the technical and speculative to the controversial and critical. Biddle *et al.* deliberately avoid providing a 'cookbook' approach to research methods, but provide enough material to help researchers appreciate the diversity of potential methods and to adopt a more critical perspective in their own research consumption and production. The section dealing with qualitative research methods is particularly welcome, an emergent research method that the *Journal of Sports Sciences* has adopted with the recent publication of the paper by Roberts *et al.* (2001).

Adopting a similar approach to that of Biddle *et al.*, the paper on research methods in sports performance by Atkinson and Nevill addresses issues relevant to the design of research in this area, rather than providing an authoritative 'cookbook' of methods. The authors communicate some possible solutions to the general problem of how to maintain both internal and external validity in applied sport performance research. The authors argue that some sport performance research has been overly concerned with physiological predictors of sport performance, at the expense of not providing a valid and reliable description of the exact nature of

Journal of Sports Sciences ISSN 0264-0414 print/ISSN 1466-447X online
http://www.tandf.co.uk/journals

the sport performance task in question. Adopting competitive sport performances as dependent variables, the authors illustrate how the influence of certain explanatory factors can be identified using linear or logistic regression. They stress the importance of checking the assumptions of any statistical tests, in particular those associated with repeated-measures analysis of variance. Indeed, when these assumptions are found to be questionable, the authors recommend a little-used and simpler technique known as 'analysis of summary statistics'. Other experimental designs are considered, including the use of matched-pairs subject designs and their value when designing experiments such as training studies (intervention and control). Finally, the authors promote the use of confidence intervals to help researchers make statements about the probability of the population difference in performance exceeding the value designated as being worthwhile or not.

The four articles contained in this 'Research Methods' issue were not designed to provide a comprehensive list of all methods used in sport and exercise science research. The main objective is to provide the readership of the *Journal of Sports Sciences* with an insight into the way research methods are evolving and developing in our main disciplines. Hopefully, the topics discussed in this issue will at least provide our readers with some 'food for thought' and possibly provoke some researchers into novel and, as yet, uncharted research paradigms!

References

Nevill, A. (2000). Just how confident are you when publishing the results of your research? *Journal of Sports Sciences*, **18**, 569–570.

Roberts, J., Jones, R., Harwood, C., Mitchell, S. and Rothberg, S. (2001). Human perceptions of sports equipment under playing conditions. *Journal of Sports Sciences*, **19**, 485–497.

ALAN M. NEVILL
University of Wolverhampton
GREG ATKINSON
Liverpool John Moores University
DAVID R. MULLINEAUX
Sheffield Hallam University