

Selected issues in the design and analysis of sport performance research

GREG ATKINSON^{1*} and ALAN M. NEVILL²

¹Research Institute for Sport and Exercise Sciences, Liverpool John Moores University, 15–21 Webster Street, Liverpool L3 2ET and ²School of Performing Arts and Leisure, University of Wolverhampton, Walsall Campus, Gorway Road, Walsall WS1 3BD, UK

Accepted 11 April 2001

The aim of this review is to discuss some issues in the design and statistical analysis of sport performance research, rather than to supply an authoritative ‘cookbook’ of methods. In general, we try to communicate some possible solutions to the conundrum of how to maintain both internal and external validity, as well as optimize statistical power, in applied sport performance research. We start by arguing that some sport performance research has been overly concerned with physiological predictors of performance, at the expense of not providing a valid and reliable description of the exact nature of the task in question. We show how the influence of certain factors on competitive performances can be described using linear or logistic regression. We discuss the choice of analysis for factorial repeated-measures designs, which is complicated by the assumption of ‘sphericity’ in a univariate general linear model, and the relatively low statistical power of the multivariate approach when used with small samples. We consider a little-used and simpler technique known as ‘analysis of summary statistics’. In multi-group pre- and post-test designs, a useful technique can be to pair-match individuals on their performance scores in a counterbalanced fashion before the intervention or control has been introduced. Finally, we outline how confidence intervals can help in making statements about the probability of the population difference in performance exceeding the value designated as being worthwhile or not.

Keywords: external validity, predictors of performance, statistical power, violation of assumptions.

Introduction

One way of defining research is whether it answers ‘basic’ or ‘applied’ questions. Basic research is designed to corroborate or discount theories of the underlying mechanisms of a particular phenomenon. Basic researchers may ask binary-type questions, such as ‘Does variable x explain variable y , when all other variables are controlled?’ Such questions are usually part of the process involved in modelling physiological or psychological mechanisms. For example, there is considerable debate on whether endurance performance is governed by oxygen use by muscle or by other central and peripheral mechanisms (Noakes, 1997). Theory-driven research questions like these can be addressed by classical hypothetico-deductive methods, the null hypothesis testing procedure and a sound

experimental design, as Chow (1996) has discussed at length. In principle, these procedures should allow the researcher to be reasonably certain that, if all variables other than x have been controlled in an experiment, and the observed changes in y cannot be attributed to chance influences, then x must be the cause of y .

Applied researchers, on the other hand, may wish to investigate factors affecting variables in a ‘real-world’ setting. A more relevant question to the researcher working in an applied field may be: ‘Does variable x , whatever the mechanism of its action, make a worthwhile difference to variable y in the real world?’ Such a question is relevant to research on whether proposed ergogenic aids improve sport performance (the term ‘sport performance’ is delimited in the present review to competitive sport performance, especially competitions involving national standard athletes). For example, caffeine is thought to improve endurance performance by at least three mechanisms (Graham *et al.*, 1994). An applied researcher might be more interested in whether an observed effect of caffeine is worthwhile to different

* Address all correspondence to Greg Atkinson, School for Health, University of Durham, Ebsworth Building, Stockton Campus, Stockton TS17 6BH, UK.

types of athletes in the real world rather than whether the underlying mechanisms are fully elucidated. We stress that there is no difference in quality or intellectual rigour between basic and applied investigations, merely a difference in the type of research question.

A relationship exists between the basic–applied research continuum and the internal–external validity continuum (Thomas and Nelson, 1996). For example, a study on sport performance needs to be externally valid – that is, it needs to involve sport competitions or simulate as closely as possible what happens in real events with athletes sampled from the population of interest. This optimization of external validity can impact negatively on internal validity in that the researcher may have less control over extraneous variables in a real-world setting. Conversely, to establish underlying causes of phenomena with basic research, the research setting needs to be constrained to maximize internal validity. These constraints inevitably reduce external validity (Thomas and Nelson, 1996). For example, researchers interested in the mechanisms of caffeine as an ergogenic aid might need to control the intensity of exercise (e.g. at 70% of maximal oxygen uptake), since exercise intensity itself may influence the explanatory variables of interest (e.g. mobilization of plasma free fatty acids). Such control reduces the external validity of the study, since most popular sports involve the competitors producing as much work as possible within a finite distance or time, rather than exercising at a constant prescribed submaximal pace.

We maintain that the above characteristics of sport performance research, especially the possible paradoxical relationship between internal and external validity, have important implications. Primarily, researchers should decide *a priori* where their research question lies on the basic–applied research continuum. Is the main aim of a study to establish whether an effect is large enough to be worthwhile to specific sport competitions, or is it to determine the underlying mechanisms of an effect that is already known to be present (or is suspected to be present)? Although such advice appears obvious, our first discussion point is whether sport performance researchers have been happy to concentrate on applied research questions and, if not, what have been the implications?

Detection of worthwhile effect or mechanism for the effect?

We have observed a general reluctance among researchers to answer in full the necessary applied research questions in studies on sport performance. Some sports scientists appear to believe in the need to describe the underlying mechanisms for an effect on

performance as well as to detect whether an externally valid and worthwhile effect is present. Such a belief may have led researchers with research questions that were originally applied in nature to concentrate on dependent variables such as physiological responses to prescribed intensities of exercise (e.g. maximal oxygen consumption, lactate minimum, onset of blood lactate accumulation, heart rate, etc.). Maximal oxygen consumption ($\dot{V}O_{2max}$), for example, can shed light on various physiological processes that are at play during exercise and is an indicator of general cardiovascular fitness (Franklin, 1999). Nevertheless, we are unaware of any sport event in which the outcome measure of performance is the point of exhaustion after exercise intensity has been increased every few minutes from an initially low work rate, as is the case in a $\dot{V}O_{2max}$ test. Moreover, some researchers have suggested that $\dot{V}O_{2max}$ is a poor predictor of performance among homogeneous samples of elite athletes (Noakes, 1998) and is relatively insensitive to detect obvious variations over time in the performances of elite athletes (Koutedakis, 1995; Jones, 1998).

Extrapolation of percentage changes in a study variable to sport performance

Inappropriate selection of dependent variables means that the true magnitude of effect of treatments or interventions on sport competitions may not be fully elucidated. For example, we were unable to locate any research that used an externally valid cycling test such as a time-trial (Atkinson *et al.*, 1999) or an intermittent protocol for bicycle road racing (Schabort *et al.*, 1998a) in the examination of the ergogenic effects of exogenous erythropoietin. This is despite the widely held view among cyclists that erythropoietin improves ‘performance’ by 5–10%. Ekblom and Berglund (1991) and Birkeland *et al.* (2000) found that erythropoietin increased haematocrit, haemoglobin and $\dot{V}O_{2max}$ by 6–11%. Nevertheless, the question remains: how do these changes in physiological variables translate to real cycling performance and competitive cyclists of international standard?

In answering the above question, a researcher might be tempted to extrapolate a certain percentage change in a physiological variable or power output to an equal percentage change in the real outcome measure of athletic performance (e.g. performance time). In an excellent discussion of their results, Birkeland *et al.* (2000) stressed that the percentage improvement in sport performance owing to erythropoietin doping could be smaller than those obtained for the physiological variables. Nevertheless, other researchers have not been so cautious. Gledhill and Warburton (2000, p. 424) stated that, ‘if the haemoglobin of an endurance

athlete falls from $155 \text{ g}\cdot\text{l}^{-1}$ to $140 \text{ g}\cdot\text{l}^{-1}$, it may be accompanied by a 5% decrease in $\dot{V}\text{O}_{2\text{max}}$ and a *parallel* impairment of endurance performance' (emphasis added). First, the quantitative link between increases in physiological variables and external work should be taken into account; that is, does a 5% increase in $\dot{V}\text{O}_{2\text{max}}$ extrapolate to a 5% increase in maximal power output? Secondly, as Hopkins *et al.* (1999) discussed, a 5% change in power output may equate to a much smaller percentage change in performance time, depending on the sport and the exact nature of the relationship that exists between speed and power. In cycling, the relationship between power and speed is non-linear (Martin *et al.*, 1998). At high power outputs, smaller changes in cycling speed result from any given change in power output. For example, if one extrapolates a change in power from 400 to 420 W (5% change) to a change in cycling speed, and ultimately a change in real performance in an externally valid 50 km time-trial, the improvement in time can be predicted to be less than 2%. We stress that we are not contesting the ergogenic effects of manipulating haematological variables here, merely the accuracy of statements regarding the quantitative impact on specific sport competitions.

Range of measurements and prediction of performance

If a researcher does choose a physiological predictor of performance as the dependent variable in a study, particular attention should be paid to the variability of scores for the population on which the test was originally validated. This issue has been important to the debate about the predictive value of maximum oxygen uptake, for example (Bassett and Howley, 1999). It is conventional to 'calibrate' actual performance with some predictor variable using correlation and regression techniques. The relationship is modelled and the adequacy of the model is examined by observing the magnitude of the correlation coefficient (r), coefficient of determination (r^2) and standard error of the estimate. We stress that high values of r and r^2 , and a low standard error of the estimate – which all suggest good predictive value of the model – can easily be obtained by choosing a sample that is heterogeneous in performance (Atkinson and Nevill, 1998). Some validation studies have, for example, pooled males and females, as well as old and young, into one population. It may be that the researchers in such studies erroneously concluded that some variables predict sport performance well with elite athletes, when, in reality, the predictive test holds just enough sensitivity to discriminate already obvious performance differences between elite athletes and club athletes, or veterans and young athletes. When such tests are applied to more homogeneous samples of elite

athletes, as illustrated in Fig. 1, they could be found to be poor discriminators of differences in performance.

Some variables (e.g. the heart rate responses to exercise) have been promoted as indicators of performance variables, such as power output and energy expenditure, based on observation of good relationships between the variables over a wide range of enforced exercise intensities in an incremental test (e.g. 100–600 W in the case of cycling). Researchers may be disappointed to find that such variables can, in fact, be poor predictors of within-competition work rate when applied to the narrow range of exercise intensities found during sport competitions (Atkinson and Brunskill, 2000). Some validation studies have also assessed whether two methods of taking a physiological measurement (e.g. $\dot{V}\text{O}_2$) correlate over a wide range of exercise intensities, such as those found in an incremental test to exhaustion. A high correlation is, again, almost guaranteed in such circumstances, but this observation does not answer the most pertinent question: do the methods agree on a given value of $\dot{V}\text{O}_2$ measured at a particular time during exercise?

In summary, it is important that researchers think carefully about the most important aim of their study, ideally before they select dependent variables and before they extrapolate their results to real sport performance. There is no reason to believe that a study is 'limited' if the aim is to establish whether performance is affected by a certain variable and no dependent variable is used other than sport performance itself (or a simulation). Moreover, with more externally valid outcome measures, researchers can devote their attention to ascertaining which component of a performance outcome has been improved in a study. For example, power outputs or times can be examined *within* a simulated race to ascertain whether it is a fast start, a fast finish or a general increase in speed over the race as a whole that has been mediated by an ergogenic aid (Atkinson and Brunskill, 2000; Atkinson *et al.*, 2001). Similarly, in field games involving intermittent activity, a researcher who uses a simulation of all the activities involved (e.g. Drust *et al.*, 1998) could determine whether improvements in low-intensity (walking, cruising) or high-intensity (running, sprinting) components explain a particular improvement in general work rate.

We have argued that more research should involve sport-specific dependent variables. This research may be performed by experimenting on athletes in real competition, by describing the influence of various measurable factors on an athlete's or a team's performance or by simulating sport performance in a valid and reliable way, so that it can be examined under various experimental treatments and conditions in a more controlled environment. Each of these methods has its advantages and disadvantages, which will be discussed next.

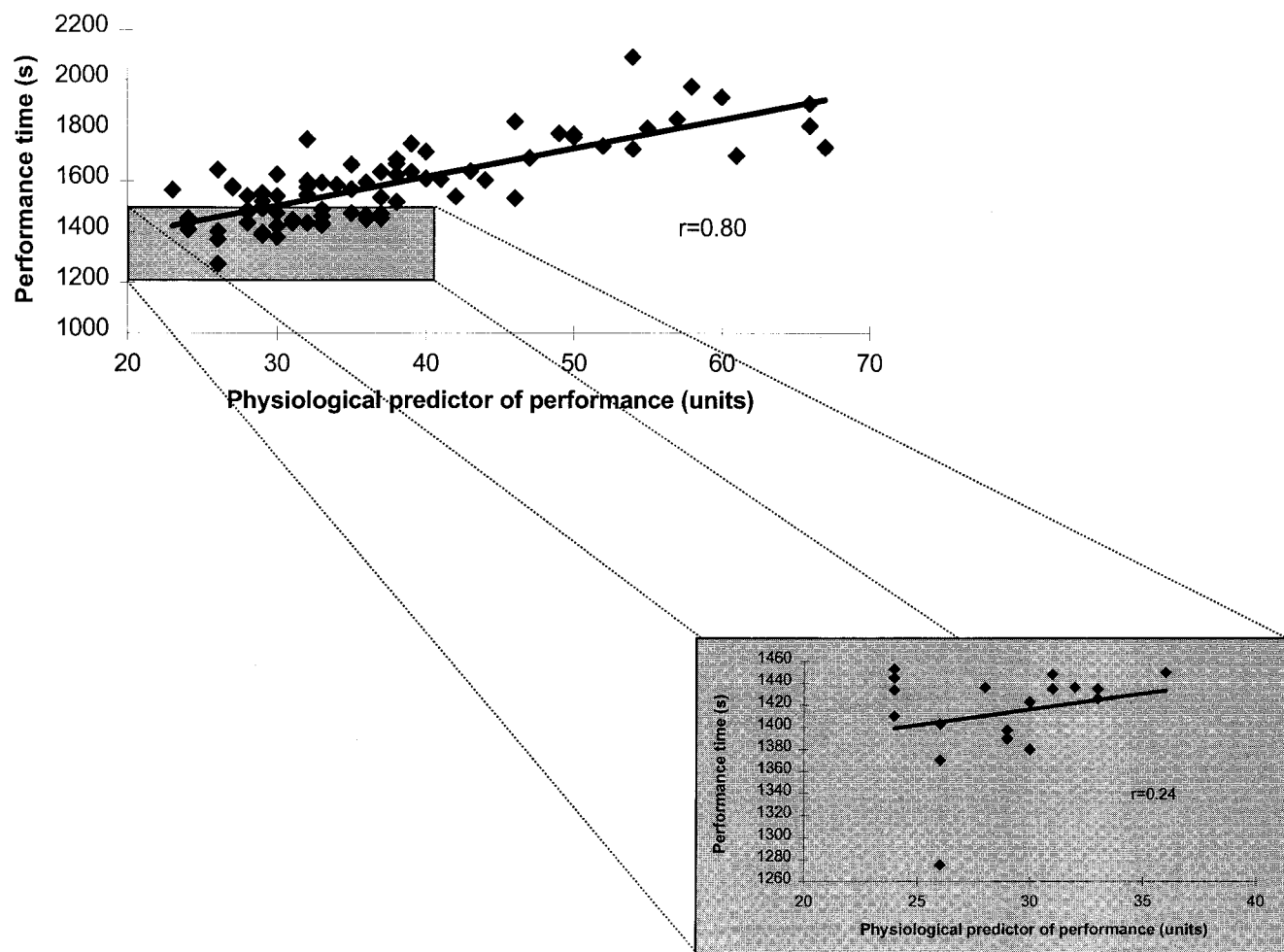


Fig. 1. The importance of considering the population of interest when predicting sport performance (simulated data). When concentrating on a more homogeneous subset (e.g. elite athletes), the relationship between predictor and outcome variables may not be so good. The predictor variable may only be useful for confirming already obvious differences between top performers and recreational athletes.

Can we experiment on athletes in real competition?

In theory, it is possible to use competitive performance as the dependent variable in a true experimental design to determine the influence of certain interventions, such as training methods or diets. This would involve repeated observations of competitive performances sandwiched between different blindly administered treatment interventions, or the random allocation of athletes to treatment groups before an event. Despite the pleas from scientists such as Youngstedt and O'Connor (1999) for more of this research on, for example, transmeridian travel and athletic performance, the practical and ethical problems are severe, especially if high-performance athletes are the intended study population. The problems with such research are discussed in more depth by Hopkins *et al.* (1999).

Quasi-experimental designs are possible (e.g. Partick and Hrycaiko, 1998) in which all elite athletes in the study are exposed to treatment and control periods sandwiched between multiple observations of performance over time. The use of quasi-experimental designs is a good example of how an increase in external validity (i.e. the use of elite athletes) leads to a decrease in internal validity (i.e. the absence of a control group because of ethical problems with restricting a particular treatment to elite athletes).

Minimal worthwhile effect for sport performance

One issue in the analysis of sport performance is the *hypothetical* change in performance that would be meaningful to the athlete, since this would govern the conclusions made in any study of sport performance. That is, it would help answer the prospective question of

whether the worthwhile effect can be detected with a workable sample size or, conversely, the retrospective question of whether any observed effect on performance from a study is worthwhile to real competition.

Hopkins *et al.* (1999) studied the within-athlete variability of performances in elite 100 m running races. In a simulation, they calculated that the minimum change in 100 m performance that would be meaningful (operationally defined as one that resulted in a change in race position that would increase a country's medal tally at an Olympic Games) could be as low as 0.3%. Hopkins *et al.* (1999) discussed the full implications of this worthwhile effect for research. Suffice it to say that this change is small, which makes it even more important to ensure that a type II error is not made in sport performance research (see section on 'Experimental studies on performance enhancement').

When describing worthwhile differences in performance, one point worth noting is that the within-athlete variability calculated from repeated sport competitions may include systematic errors (those of known cause) because of between-event differences in pre-competition travel influences, environmental changes (e.g. temperature, wind speed, humidity), health status of the athlete (e.g. presence of upper respiratory tract infections) or menstrual cycle status. Taking an extreme example, those female cyclists who competed in both the 1997 and 1998 British 16.1 km time-trial championships, which were held in different wind conditions and on different courses, showed mean recorded times of 1504 ± 101 s and 1532 ± 104 s, respectively (mean \pm s; paired *t*-test, $P < 0.001$). Such influences would be controlled in a laboratory-based study and so would not contribute to the overall test variability. Therefore, it may be that a laboratory-based simulation of sport performance shows less variability than actual sport performance. This would help in minimizing the chance of type II error in a laboratory-based study (see section on 'Descriptive research'). Nevertheless, as Hopkins *et al.* (1999) discussed, researchers have to be extremely wary of external validity factors (whether the observed effect in the more controlled environment would occur in the real sport setting). For example, optimal pacing strategies during cycling time-trials may be different between the controlled conditions of the laboratory and the variable external conditions (hills and wind) in real cycling events (Swain, 1997; Atkinson and Brunskill, 2000). Hopkins *et al.* (1999) suggested measuring as many of these systematic factors as possible and including them as covariates in the analysis.

To summarize, intervention studies on athletes in competition are difficult to administer. Researchers should be aware of the minimal important change in performance that can be estimated from the likelihood that an athlete moves up and down competitive places

or ranks (Hopkins *et al.*, 1999). The issue of worthwhile differences in performance is obviously more complicated for team sports, but could be resolved by more applied research on the relative influences of individual factors that contribute to overall team performance. One way of investigating such factors is through descriptive research.

Descriptive research

Descriptive research on sport performance involves the recording of competitive outcomes of athletes or teams, together with details of interest that are thought to influence these outcomes. The choice of predictor variables may be made in a planned fashion before the events take place or a researcher might retrospectively group the performances or decide on predictor variables in an *ex post facto* (after the event) fashion. An example of the latter type of descriptive research is provided by Atkinson *et al.* (1994). They catalogued the performances of cyclists who raced at varying times of day. The cyclists were allocated *ex post facto* into groups on the basis of age. Time of day of competition was also examined as a repeated-measures factor. The cyclists were found to perform better in evening than in morning races. Most match analysis research also fits into this category, in that performance outcomes are recorded and then analysed *ex post facto* for the effect of some operationally defined variable (e.g. part of the pitch from which a goal is scored, home advantage, environmental factors). Because this type of research is correlational, care is needed in both the research design and the statistical analysis before making any conclusions about hypothetical causes of performance outcome.

Research design

Researchers need to control for the number of observations within levels of factors in descriptive research. For example, Sears and Harris (1999) reported that the ratio of home to away wins was 3 to 1 in the football Premier League for the first third of the 1999–2000 season, whereas it was 4 to 1 for the whole of the previous season. The apparent decrease in home advantage for the first third of the 1999–2000 season could be due to a bias of the better teams playing relatively more games away from home than the inferior teams. Over a whole season, this bias would not be apparent, since all teams would have played equal numbers of home and away games.

Similarly, the observation that most of the track and field athletics records have been set in the afternoon or evening should be treated with caution, since the

finals of events are conventionally scheduled at these times, and athletes would be likely to 'save' their best performances for the final (Atkinson and Reilly, 1996). As well as care in controlling the frequency of observations, researchers must try to control for any intervening variables in a study. For example, environmental factors (e.g. fluctuations in air density) could have influenced the results of Atkinson *et al.* (1994), who found that cycling performances are better in the evening. Such control cannot be done as part of the research design, since the study is 'in the field', but it can be attempted as part of the statistical analysis of the data by modelling the relative influences of factors on performance.

Statistical analysis

It is worth mentioning at this point that the most appropriate analysis for some descriptive research on performance outcomes might involve 'finite population' statistics. It is possible that a researcher has obtained all, or a good proportion of (>5%), the possible observations from the 'population' of interest. For example, it is not logical to think of all the performance outcomes from a particular soccer World Cup competition to be from an infinite 'population' of all World Cup games that have ever been and will be played. In this example, it could be considered that the researcher has obtained all observations from the finite 'population' of games in the competition. The analysis of such information is beyond the scope of this review. Interested readers are directed to Zar (1999) and Hamburg (1970) for the finite population correction formulae. If relevant to a particular study, such corrections of statistical tests for sampling from a finite population may be important, since the precision of the estimation of population parameters is improved by making the appropriate corrections.

Regression is the most widely used data analysis technique available to help researchers identify factors that are associated with optimal sport performance. The particular type of regression analysis used will depend on the sport performance dependent variable being investigated. If the dependent variable is continuous, unbounded and measured on the interval or ratio scale (e.g. distance thrown, run time), then ordinary linear or multiple linear regression methods are appropriate. On the other hand, if the dependent variable is categorical (e.g. win *vs* loss, presence *vs* absence of an injury), then logistic regression or discriminant function analysis are more appropriate. For more information on the methods described in this section, we refer the reader to Dillon and Goldstein (1984), Kleinbaun (1994) and Agresti (1990). Discriminant function analysis is described by Biddle *et al.* (this issue).

Multiple linear regression. Multiple linear regression allows the researcher to identify which independent (or predictor) variables are associated with a dependent variable. More precisely, multiple linear regression enables the investigator to predict values of a dependent variable from values of a collection of other predictor variables, assuming the predictor variables are either linearly, or in various linear combinations, associated with the dependent variable. However, when predicting sport performance, there may be certain instances when the dependent variable can be predicted more than adequately using a subset of the predictor variables – that is, a 'reduced model'. This reduced model is often referred to as the 'parsimonious solution' to the regression analysis. There are several useful methods for choosing good reduced models. These fall into two categories: best subset selection methods and stepwise regression methods.

Suppose we wish to identify the mood states or factors that are best associated with successful cross-country running performance. Cockerill *et al.* (1991) used multiple linear regression to identify which of Morgan's (1985) Profile of Mood States (POMS) were best able to predict the cross-country race times of 81 runners competing in the 1990 British Students' Cross-country Championships. When all six POMS factors were entered into Minitab's 'BREG' multiple regression routine as possible predictors of run time, the best subset of mood factors was found to be:

$$\text{Time} = 62.6 - 0.266T + 0.246D - 0.317A \quad (1)$$

where Time = race time in minutes, T = Tension, D = Depression and A = Anger.

The 'BREG' command in Minitab performs a 'best subsets' multiple-regression analysis, using the maximum coefficient of determination (R^2) as the criterion, by first examining all one-predictor regression models and then selecting the two models giving the largest R^2 . Next, the analysis examines all two-predictor models, selects the two models with the largest R^2 , and then displays criterion information on these two models. This process continues until the model contains all the available predictors. For each model, the Minitab output provides information based on four criteria: the coefficient of determination, R^2 ; the adjusted coefficient of determination, $\text{adj } R^2$; Mallows's criterion, C_p ; and the standard deviation of errors about the regression line, s . (For further information about the BREG procedure, see Hocking, 1976; Goodnight, 1979.) The model chosen, and shown in equation (1), provided the maximum adjusted R^2 of 0.337 ($F_{3,77} = 4.31$, $P < 0.01$) and the smallest Mallows's C_p value of 2.8. Evaluating all possible combinations of subset models is the most

comprehensive and thorough way to proceed in variable selection. Unfortunately, the computational demands of evaluating every possible 'subset' model can be prohibitive, especially when more than 15 predictor variables are available. An alternative method of identifying a reduced subset model is to use stepwise regression.

Stepwise regression methods either remove or add variables for the purpose of identifying a reduced model. The three commonly used procedures are: standard stepwise regression (adds and removes variables), forward selection (adds variables) and backwards elimination (removes variables). The best of these methods, backward elimination, begins with a full or saturated model and the least important variables can then be eliminated sequentially (based on the size of the t -statistic for dropping the variable from the model). When stepwise regression using backward elimination was applied to the cross-country running results of Cockerill *et al.* (1991), the same solution as that chosen from the Minitab output (equation 1) was obtained. Note that when both standard and forward stepwise regression methods were used to predict the athletes' run times, only the factor 'Tension' was selected. The preferred solution (see equation 1) suggests that a linear combination of three mood states, working in combination, provides the researcher with valuable insight into the complex relationships between moods that are likely to result in successful cross-country running performance. Clearly, this example emphasizes the importance of exploring the data using several of different regression methods to provide a 'parsimonious solution' to the question of which mood states or factors, or combination of mood states and factors, are best associated with successful cross-country running performance.

Logistic regression. If the dependent variable is dichotomous (e.g. win *vs* loss, presence *vs* absence of an injury), the assumptions necessary for hypothesis testing in linear or multiple linear regression analysis are likely to be violated. A more appropriate multivariate technique for such categorical data is logistic regression, which will estimate the probability that either one or the other categorical event will occur, based on a range of predictor or independent variables.

We introduce logistic regression with the aid of Pollard and Reep's (1997) study on the effectiveness of playing strategies in soccer. Another example of logistic regression can be found in Lee and Garraway (2000), who identified the influence of environmental factors on rugby union injuries. Using team possessions as the units of observation, Pollard and Reep (1997) recognized that adopting 'goals scored' as their categorical [0,1] dependent variable would result in too few

'goals scored' compared with the many team possessions that failed to result in a goal. Indeed, out of nearly 6000 team possessions recorded, only 47 goals were scored. Clearly, if 'goals scored' were to be used as the categorical dependent variable, over 99% of team possessions would be classified as failed possessions and provide little or no information about likely effective strategies leading to goals. Hence, an alternative logistic regression analysis was carried out on the 489 team possessions that resulted in a shot, leading to the 47 goals scored. Pollard and Reep argued that various factors (e.g. location of the shot) were likely to influence the probability of scoring. Hence, using whether a goal was scored or not as the dependent variable, binary logistic regression was used to identify which factors were likely to influence the probability of scoring a goal. The predictor or independent variables were:

- the distance ('Dist') in yards from the goal;
- the angle ('Angle') in radians to the nearest goal-post (Angle = 0 if the shot was directly in front of goal);
- a measure of how many touches ('Touch') the player had before shooting (one touch, Touch = 0; more than one touch, Touch = 1);
- a measure of how close ('Close') the nearest defender was when taking the shot (less than a yard, Close = 0; more than one yard, Close = 1);
- whether the possession originated in either open ('Open') play or set play (open play, Open = 0; set play, Open = 1).

Logistic regression is able to incorporate a mixture of categorical [0,1] indicator variables as well as continuous explanatory variables when predicting the probabilities associated with a binary dependent variable. Logistic regression estimates the probability (p) of an event occurring as follows:

$$p = \exp(y) / [1 + \exp(y)] \quad (2)$$

where y is a linear combination of the predictor variables, known as the 'logit' model, and where $\exp(y) = e^y$.

The 'logit' model for the 410 kicked shot possessions was found to be:

$$y = 1.245 - 0.219 \text{ Dist} - 1.578 \text{ Angle} + 0.947 \text{ Close} - 1.069 \text{ Open}$$

from which the probability (p) of scoring a kicked goal can be calculated using equation (2). Note that the 'Touch' [0,1] indicator variable was not significant and, therefore, was not included in the logit model. Pollard

and Reep (1997) provided an example of a shot from 16 yards, directly in front of goal, with an opponent within 1 yard (Close = 0) and from a team possession originating as a set play (Open = 1). The value of y was -3.328 and hence the probability of scoring was estimated as $p = 0.0346$.

The 'logit' model can also be used to help understand and interpret the relationship between the predictor or explanatory variables and the dependent variable (scoring a goal). For example, using the beta weight for 'Dist', we can calculate that, for every yard nearer goal, the odds [$\exp(0.219) = 1.24$] of scoring increases by 24% (obtained by taking 1 from the odds ratio = 1.24 and describing the difference as a percentage).

Experimental studies on performance enhancement

The final way of studying the influences of certain factors on sport performance is to examine analogues of performance in a more controlled setting. In this section, we delimit our discussion to the more common scenario of a researcher investigating whether a certain factor *improves* performance; that is, performance enhancement research. In descriptive research, the number of study cases may well be large, since performances can be retrospectively catalogued over several seasons or years. Large sample sizes for experiments in the laboratory are obviously more difficult to obtain, particularly if the population of interest is top-class performers. Together with the researcher choosing an appropriate dependent variable relevant to performance (discussed earlier), and in light of the probable presence of a limited sample size, an experiment also needs to be designed for optimal statistical power, especially in light of the small meaningful effect sizes that may characterize sport performance research.

Statistical power is the probability of rejecting the null hypothesis when it is false and should be rejected (Altman, 1991). If the small sample size and sub-optimal design of a study lead to inadequate statistical power for a particular null hypothesis, a type II error may have occurred. In other words, it could be erroneously concluded from a study that no effect is present when, in fact, there is an effect on performance. The factors that influence statistical power are given in Table 1. It has been shown that the size of the worthwhile effect could be small in sports such as 100 m sprinting (Hopkins *et al.*, 1999). The smaller the worthwhile effect, the smaller the power for a given sample size and within-individual variability. In the absence of other research similar to that of Hopkins *et al.* (1999), researchers trying to estimate statistical power or sample size before

an experiment might like to consider a general effect on sport performance if it mediates a change of 1% in a variable measured on a ratio scale. We stress that, ideally, a 'sport-by-sport' estimation of worthwhile magnitude of effect should be considered before any study on sport performance, since worthwhile effects could be smaller than 1% for some sports (Hopkins *et al.*, 1999).

The statistical power component of effect size cannot be controlled by the researcher. However, some of the factors cited in Table 1 can be optimized for statistical power by careful consideration of the research question, research design and analysis. Before these are considered, we offer a word of warning concerning statistical power. Many statistical software packages (e.g. SPSS) provide *post-hoc* power calculations as part of the hypothesis tests that are performed on a set of data (e.g. repeated-measures analysis of variance). It is extremely important that the researcher does not use these power calculations from their *own* data without considering whether the differences that were observed are close to the effect size that is considered worthwhile. It is always good practice to quote statistical power if one has not rejected the null hypothesis. Nevertheless, it is important not to overlook the fact that the observed effect may have been far from being meaningful to sport performance. We stress that statistical power calculations are related to the *minimum worthwhile effect* that is decided *a priori*, and not necessarily the observed effect from a particular study.

One- or two-tailed hypotheses in performance enhancement research?

Researchers seldom rationalize the choice of a one-tailed or two-tailed analysis. One-tailed analyses are selected when the hypothesis of interest is directional (e.g. an *increase* in sport performance is hypothesized), whereas two-tailed tests are chosen when the hypothesis of interest involves a *change*, irrespective of direction. We argue that a one-tailed test might be used when the researcher is only interested in enhancement of the performance outcome and when that performance outcome is directly measurable. We stress that it is erroneous and unethical to use a one-tailed test just so that a change in performance is shown to be significant, when there is no justification for a one-tailed test. Nevertheless, it might be equally erroneous to use a two-tailed test when only an *improvement* in performance is of interest, if the intervention has an effect at all. Such a decision is important, since one-tailed test statistics offer a gain in statistical power over the corresponding two-tailed test (Rice and Gaines, 1994).

In this discussion of one-tailed versus two-tailed tests, we wish to stress that it is incorrect to think of the null

Table 1. Factors affecting statistical power in an experiment on sport performance

-
1. *Sample size:* The greater the sample size, the greater the power of the test, so the chance of making a type II error is decreased
 2. *Effect size:* Also known as the ‘worthwhile’ effect or ‘substantive significance’, this is the magnitude of differences or changes that is considered to be of practical significance (e.g. worth spending time and money on to improve athletic performance). This factor can be expressed in units of measurement (e.g. 10 s), a percentage change (e.g. 5%) or a ‘standardized effect’ (Altman, 1991), where the effect size is expressed relative to the sample standard deviation or standard deviation of differences. The larger the effect size, the greater the power of the test, all other factors being equal. Researchers are encouraged to operationally define an effect size from, for example, worthwhile differences in sport (Hopkins *et al.*, 1999) rather than use general descriptors of effect size (e.g. ‘moderate’)
 3. *Error variance:* This is the variance between individuals when group differences or correlations are considered, and the variance within individuals (residual error) when changes over time are examined. Measurement error statistics from reliability studies can be used to estimate the population within-individual variance. Ultimately, for power calculations, these are converted to the terms that are entered into equations for the hypothesis tests (i.e. the standard deviation of differences in the case of a paired *t*-test; the mean-square error term for repeated-measures analysis of variance). The smaller the error variance, the greater the statistical power. Error variance can be reduced substantially in human measurements by performing enough trials to familiarize the participants with the equipment and protocols
 4. *Alpha level:* This is the probability of making a type I error (rejecting the null hypothesis when it is in fact true). A common alpha level that researchers use to denote ‘significance’ is 5% (0.05). By lowering alpha, a researcher increases Beta (type II error rate) and decreases the power of the test (lowers the chance of correctly accepting a null hypothesis)
-

hypothesis as representing zero change or difference in all research circumstances. For example, if it is hypothesized that a particular treatment leads to an increase in performance compared with a placebo, but either no change or, paradoxically, a decrease in performance is observed, then the null hypothesis is not rejected (the null hypothesis would be written as H_0 : treatment mean \leq placebo mean; Zar, 1999). Conversely, a two-tailed test should be used when the researcher believes it is just as important to examine whether performance is hindered as it is improved (the notation for the null hypothesis would be written, in this case, as H_0 : treatment mean = placebo mean). If one thinks logically about the applied nature of performance enhancement research, the latter non-directional scenario might not be as appropriate in some circumstances. This is because, irrespective of whether a treatment does not work or actually hinders performance, the same logical conclusion should be reached by the performance enhancement researcher after accepting the null hypothesis of $H_0 \leq 0$; that is, the conclusion would be that sport performers should not use the treatment, as it will either be a waste of time or hinder performance. In other words, detection of a paradoxical outcome to an enhancement in performance has no practical significance for the sport performer.

We stress the difference between some types of performance enhancement research and health or clinical studies in this discussion of directional and non-directional hypotheses. Bland (1995, p. 137) stated that ‘the position (of choice of test) depends on the field in which the testing is actually done’ and the presence

or absence of ‘complicated relationships amongst variables’. We agree with this statement. In health research, the use of one-tailed significance tests has been discouraged (Altman, 1991) on the basis that an unexpected paradoxical finding could be important. Although this view has been challenged (Peace, 1988), we agree that, in health research, there are usually many dependent variables (symptoms) of interest that make up a construct of ‘health’ and these should all be monitored in response to some intervention. If only one of these dependent variables shows an unexpected paradoxical response, this may be important enough for the researcher to exert caution in any conclusion. For example, a researcher investigating an intervention designed to improve cardiovascular ‘health’ might have good reason to suspect that cholesterol should decrease and not increase, but little may be known about how the intervention affects blood pressure; therefore, two-tailed tests are appropriate. When the performance outcome is singular and directly measurable (e.g. time, distance), a paradoxical response, no matter how unexpected, should still result in the non-use of the particular intervention; therefore, one-tailed tests might be appropriate. We stress that not all performance outcomes are single dependent variables. Soccer ‘performance’, for example, could be considered to be a construct comprising many factors. Although a particular intervention may improve one component of soccer performance (e.g. sprinting), there is no guarantee that another component (e.g. endurance) will not be detrimentally affected; therefore, two-tailed testing is more appropriate. The one-tailed versus two-tailed debate is interesting and has been covered before in other subject

areas. Readers are directed to Rice and Gaines (1994) and Peace (1988) for persuasive arguments supporting the use of one-tailed tests. As a final point, Peace (1988) believed that the equivalent debate for using confidence intervals to interpret study results is whether to adopt 95% or 90% limits. The confidence interval approach to analysis is covered on pp. 824–825.

The matched-pairs design

Research designs that involve correlated data (e.g. repeated measures) are more powerful than those involving separate unrelated groups. Sometimes it is difficult to use repeated measures or crossover designs, since a treatment might have long residual effects on performance. A research design that is worth considering by sport and exercise scientists in such cases is the pre-test matched-pairs approach. This design involves not just matching a treatment group and a control group for any intervening variables such as age or body mass. The important point is that the treatment and control groups are matched according to their initial pre-treatment performance score in a counterbalanced fashion. After familiarization and measurement of all the participants' performances in a pre-test, they are ranked according to their performance score and allocated, in a specific sequence, to the treatment or control group. This sequence is not as simple as one might think, as the researcher needs to equalize mean performance between the groups. Therefore, one does not allocate participant 1 to control, participant 2 to treatment, participant 3 to control, participant 4 to treatment, and so on, as this would introduce a bias for the mean pre-test score of the control group to be different to that of the treatment groups. The sequence must be participant 1 to control, participant 2 to treatment, participant 3 to treatment, participant 4 to control, and so on. Vincent (1999) referred to this sequence as the 'ABBA assignment procedure' where 'A' and 'B' represent two study groups. With this procedure, the initial difference in mean performance of the two groups would be small. We would be less likely to find that the pre-tests are significantly different from each other with this design compared with the random allocation of participants. We stress that bias in the matched-pairs design should not be any greater than that of random allocation, since it is unlikely that the participants ranked 1, 4, 5, 8, 9 respond any differently to a treatment than those ranked 2, 3, 6, 7, 10.

One can analyse matched-pairs designs in several ways. The simplest analysis is to compare the post-intervention performance scores between the experimental groups. First, if there is an effect of the treatment on performance, this would show up as a significant difference on the post-tests. Secondly, the test of differ-

ences between post-tests would be a paired analysis, since the participants were pair-matched initially making the data correlated. This increases the power of the test in comparison to a two-sample *t*-test, for example, since it factors out the between-individuals differences (i.e. reduces the error variance). A multi-factorial model (group \times test) can also be used to analyse the data and, although more difficult to calculate and interpret, this may be the more powerful approach. The interaction term of this analysis would be of greatest interest in a treatment-control group study. A similar analysis would be to test whether the 'delta changes' (the difference between pre- and post-intervention tests) are different between the treatment and control groups. The complication in these latter two analyses, compared with a comparison of the post-test scores, is that there is no guarantee that the delta changes are correlated between the groups. We advise that the correlation between the difference data of the groups is examined before adopting a paired (within-individual) analysis or an unpaired (between-individuals) analysis, since it is this characteristic which governs the statistical power of the hypothesis test (Zar, 1999).

The pre-test matched-pairs approach is efficient for investigating the effects of treatments that have long washout times, such as creatine. In a paired analysis on the same participants (i.e. a repeated-measures crossover design), the researcher would have to wait for the residual effects of the treatment to dissipate, which could take months. It is also the most preferred design for training studies for which it is not possible to adopt a repeated-measures crossover approach because of time constraints and the fact that external validity would be affected if a sample completed a training phase first followed by a control phase. One disadvantage of the matched-pairs design is that only one variable may be able to be analysed with a paired statistical test. Any related physiological responses to the performance test, for example, might not be matched in a pairwise fashion between the groups. The researcher could check the assumption of related samples by examining the associated correlation coefficient before performing a paired analysis on dependent variables other than the performance scores. Another disadvantage associated with this design is that the matching process would be compromised by one participant dropping out of the study between pre-test and post-test; another participant would have to be omitted from the analysis to retain the pairs, which might lead to sample bias. If the overriding aim is to assess a performance test and the sample is likely to remain intact, the matched-pairs analysis can be a powerful design for training studies or interventions with unworkably long washout times and small sample sizes.

The number of levels in a design

This issue is complicated and depends on whether the experiment has 'between-individuals' factors or not. Given a particular sample size and effect size, and that the participants are to be grouped according to operationally defined between-individuals variables (e.g. age), the statistical power will be reduced the more groups are formed. For example, all other things being equal, an analysis involving two groups of 15 participants would have more power than one involving three groups of 10 (Zar, 1999). In light of the small worthwhile effects and limited sample size (discussed below) for studies on sport performance, we advise that the researcher keeps such studies as simple as possible by concentrating on the most important research questions. Again, there are no prizes for having the most levels for a factor within a between-individuals analysis (investigating the maximal number of different categories of a variable). Such a strategy may merely increase the possibility of a type II error.

Matters are different when repeated measurements are taken within each level of a particular factor of interest. As discussed by Mullineaux *et al.* in this issue, the presence of replicates is desirable, in that the average of the data is closer to the 'true value' of the outcome. Often, time is a factor of interest in an experiment (e.g. pre-exercise, in-exercise and post-exercise measurements). In these designs, multiple levels of a factor could be analysed with either multivariate or univariate repeated-measures analysis of variance (Schutz and Gessaroli, 1987).

The most common approach to repeated-measures analysis is the univariate method, since the small sample sizes that characterize sport performance experiments mean that the multivariate method would have low statistical power. Moreover, with low sample sizes, the calculations for multivariate analyses of variance might not be possible (Maxwell and Delaney, 1990). As a general rule, Maxwell and Delaney (1990) proposed that the multivariate approach should probably not be used if the sample size is less than $(a + 10)$, a being the number of levels for repeated measures. If the sample size is larger than $(a + 10)$, then the multivariate method is preferred (Stevens, 1992). It may be difficult to obtain more than 13 top-class performers (the suggested sample size for a study involving three treatments given to the same participants) in an experiment, so the multivariate method would not be the most appropriate choice.

It is important that the assumption of sphericity is examined as part of a univariate repeated-measures analysis of variance (ANOVA). The exploration and correction of the complicated issue of sphericity in ANOVA is discussed eloquently by Maxwell and

Delaney (1990) and summarized by Field (2000) and Atkinson (in press). In brief, the sphericity assumption pertains to the population variances of the difference scores between all possible pairs of repeated measures being equal. In other words, the population variance of the test1–test2 difference scores should be similar to the population variances of the test2–test3 and test1–test3 differences (Maxwell and Delaney, 1990; Field, 2000). Suffice it to say, we have found violation of the assumption to be so common with measurements on humans, that the power of the test is reduced after correction. An alternative strategy that may be useful for examining serial measurements is an 'analysis of summary statistics'.

Analysis of summary statistics

Analysis of summary statistics (Mathews *et al.*, 1990) is an alternative to a repeated-measures ANOVA when comparing changes in performance over several instants between different treatment groups or conditions. Using ANOVA to examine the hypothesis that the changes over time are different between treatments, one would examine the treatment-by-time interaction. A significant interaction would mean that the changes over time (e.g. the gradients) are different. An analysis of summary statistics involves the description of meaningful aspects of the data for each participant and then examining differences with a simpler hypothesis test. For example, the time of peak performance, the value of the peak itself and the mean performance over time can be summarized for each participant into a single statistic and then compared between two different treatments with a paired *t*-test. The assumption of compound symmetry is not relevant to the *t*-test, as it is used on only two levels of a factor. Mathews *et al.* (1990) maintained that an analysis of summary statistics has the advantage that it directs the researcher to test specific hypotheses about the data, rather than the 'data dredging' that sometimes occurs with multifactorial ANOVA, followed by *post-hoc* multiple comparisons. For a very detailed discussion on the use of summary statistics in repeated-measures designs, readers are directed to the review by Senn *et al.* (2000).

Within-individual variability

This factor probably has the most influence on statistical power and is conventionally examined as part of a test–retest reliability study. We consider within-individual error to be 'measurement error' – that is, any error, biological or instrumental, of unknown or unexplained origin. Atkinson and Nevill (1998) stressed the importance of providing a measure of within-individual error (absolute reliability) in a reliability

study and not just relying on a test–retest correlation coefficient to indicate whether a measurement tool is reliable enough to be used in sport performance research. Test–retest correlation is highly sensitive to between-individuals variability, and even a correlation above 0.9 may not mean the performance test is reliable enough to be used in studies with a workable sample size (Atkinson and Nevill, 1998). Conversely, a test–retest correlation of 0.5 may be indicative of adequate reliability if the analytical goal (effect size) for future research is large (Hofstra *et al.*, 1997). Another ‘rule of thumb’ that should be treated with caution is a coefficient of variation being designated as acceptable if it is less than 10%. As detailed in Table 2, we maintain that an aim of a reliability study is not just to describe the measurement error in an appropriate way, but to extrapolate what the measurement error means for sport performance research.

There are several statistics that can be used to represent within-individual error, including the standard error of measurement, the coefficient of variation and the limits of agreement (absolute and ratio). The last of these statistics, when applied to measurement error, is also known as the coefficient of repeatability (Bland and Altman, 1999). Assuming errors are normally distributed, the standard error of measurement or ‘within-individual standard deviation’ represents, for the ‘average’ score or person, the range above and below the observed score for which there is a 68% probability of the hypothetical true score falling (Harvill, 1991). This calculation is only an approximation, since the true score can never be known (Harvill, 1991). The coefficient of variation is a similar statistic to the standard error of measurement, but should only be used with variables measured on a ratio scale and when there is evidence that the error increases as the magnitude of the measured variable increases (Nevill and Atkinson, 1997).

It would be expected that the test–retest differences purely due to measurement error would be no greater than the limits of agreement for 95% of individuals in a population (Bland and Altman, 1986). For example, a limits of agreement of ± 5 s for a cycling time-trial would mean that one can be reasonably certain ($P = 0.95$) that an individual scoring 200 s would score between 195 and 205 s with another observation, and that the difference would be purely due to measurement error. The sport performance researcher can then judge whether this test–retest error is acceptable (i.e. makes little difference regarding an athlete’s ability). The limits of agreement can also be applied to judgements on whether changes in performance in individual athletes are ‘real’ or merely due to measurement error (Eliasziv *et al.*, 1994), or whether replicate measurements should be obtained to improve precision (ISO, 1994).

It is erroneous for a researcher (e.g. Lucia *et al.*, 1999) to accept the measurement tool as being reliable on the basis of 95% of the test–retest differences being within the limits of agreement for the participants involved in the reliability study itself, since these participants were used to calculate the limits of agreement in the first place (Atkinson and Nevill, 2001). The confidence intervals of the measurement error statistic should be quoted in a reliability study, and the adequate precision of these confidence intervals should be reflected in the design of the study (i.e. a sample size of at least 40 participants) before any data are collected (Critchley *et al.*, 1999).

We promote the inclusion of limits of agreement, together with the other statistics, in reliability studies, since it involves an informative way of looking at measurement error with a Bland-Altman plot (Bland and Altman, 1999). It also allows the examination of the nature of the relationship between error and the magnitude of the measured variable and it is based on a clear analytical goal – the predicted difference between a test and a retest that would be due purely to measurement error for most individuals (95%) in a population. Atkinson and Nevill (1998) and Nevill and Atkinson (2001) discuss the use of limits of agreement with worked examples.

The designs and analyses for multiple measurements in single-case research are related to the issue of measurement error for assessment of individual changes. This type of research is quasi-experimental at best and so will not be covered here. Nevertheless, single-case designs and analyses could be useful to researchers offering sport science support to athletes. Interested readers are directed to the reviews of Aeschleman (1991) and Barlow and Hersen (1984), as well as an extremely useful website covering designs and statistical tools (Jones, 2000).

As Atkinson and Nevill (1998) noted, it is not the measurement error statistic that is so important in studies involving *samples* of individuals, but how the statistics are applied to analytical goals for future research in which the measurement tool of interest is used. In this respect, it is important that the reliability researcher does not stop at describing the measurement error, but extrapolates the measurement error to questions regarding whether the measurement tool would detect typical effects with workable sample sizes. Atkinson *et al.* (1999) showed a worked example of such a ‘sensitivity analysis’ calculated from limits of agreement. The nomogram that was used by Atkinson *et al.* (1999) is reproduced in Fig. 2. We stress that such an analysis can be performed with other measurement error statistics, such as the coefficient of variation (Schabort *et al.*, 1998b) or correlation (Hofstra *et al.*, 1997).

Table 2. Checklist for reliability researchers with guidance notes

1. *Randomly sample at least 20 – but preferably 50 – individuals from the population of interest.* The consideration of measurement error is a parameter estimation problem, not a hypothesis testing problem. Therefore, the sample size in a reliability study should be large enough for a precise estimate of the chosen measurement error parameter (intraclass correlation, standard error of measurement, coefficient of variation, coefficient of repeatability) for the population of interest
2. *Calculate confidence limits for the measurement error statistics.* In light of the above nature of measurement error as a parameter estimation problem, it is also important to cite confidence limits for the measurement error statistics. Bland and Altman (1999) provide details relevant to the coefficient of repeatability. Hopkins (2000) shows how to calculate confidence limits for the coefficient of variation and Morrow and Jackson (1993) provide details for the intraclass correlation. Interestingly, few researchers have considered confidence limits for the standard error of measurement, possibly because this statistic has erroneously been referred to as a confidence interval itself in the past. The standard error of measurement, like the coefficient of repeatability, is, in reality, a reference interval or 'normal' range for the difference between the observation and a hypothetical true score
3. *Make the test protocol and recovery time between tests compatible with common research uses of the measurement tool.* This means that all components of measurement error (biological and mechanical) are considered initially, as would be the case when the measurement tool is used in future research. Although exploration of different sources of error (experimenter, manufacturer, etc.) with, for example, generalizability theory is important, it is more important to assess whether the measurement error for the simplest research scenario (same experimenter, same equipment) is acceptable for future use. If the measurement tool is not deemed reliable under the simplest of scenarios, the addition of variance from different measurement tools and observers will only make a bad situation worse
4. *Establish whether any systematic error exists between test and retests.* This can be done with a paired *t*-test or analysis of variance, although care should be taken to examine the error in relation to practical significance as well as statistical significance. This is because the results of the above hypothesis tests are compromised by the presence of large random error (Atkinson and Nevill, 1998). Calculating the confidence interval for the mean differences between repeated tests may help in this respect. If systematic error is present, the reliability researcher should first focus on the number of familiarization tests that are required for learning influences to dissipate
5. *Determine the relationship between measurement error and magnitude of measured value.* Bland and Altman (1999) and Atkinson and Nevill (1998) provide details of how this is done. If the measurement error does increase in proportion to the size of the measured value, the data should be logarithmically transformed and a ratio statistic should be used to describe the measurement error (e.g. ratio of coefficient of reliability and coefficient of variation). If the error is 'homoscedastic', measurement error can be described in the particular units of measurement by calculating the coefficient of reliability or standard error of measurement. Complicated relationships between magnitude of error and measured value can be analysed using non-parametric methods according to Bland and Altman (1999) or by calculating measurement error statistics for separate sub-samples within a population (Lord, 1984)
6. *Examine whether random error changes between separate test and retests.* After stages 4 and 5, and if there are still multiple retests left in the reliability analysis, as the results of the present study indicate, the researcher should explore whether random error (described by the coefficient of reliability, standard error of measurement or coefficient of variation) is reduced as more tests are administered. If so, the researcher should communicate this so that future users know exactly how many familiarization sessions are required for total error variance to be minimized for their research
7. *Perform a 'sensitivity analysis' using the described measurement error.* This involves predicting whether the error is small enough for the measurement tool to detect typical 'analytical goals' (e.g. a 5% change due to an intervention of some sort) with a workable sample size in a future experiment. Before this stage is performed, measurement error has merely been explored and described and the question of whether the measurement tool is reliable has not been investigated. The use of arbitrary 'rules of thumb' such as accepting adequate reliability on the basis of an intraclass correlation being above 0.9 or a coefficient of variation being below 10% is discouraged, since no relation is made between error and real uses of the measurement tool with such generalizations. An example of a sensitivity analysis with the coefficient of reliability is provided by Atkinson *et al.* (1999). Hopkins (2000) and Charter (1997) also provide calculations relevant to the coefficient of variation and a correlation coefficient, respectively
8. *State how the measurement error would impact on individual measurements.* An example should be provided to show the impact of error on individual measurements. A person scoring the average value in a population could be used for the example, although if heteroscedasticity is present, the error for a person scoring a large value should also be discussed. It is important to note that the coefficient of reliability and standard error of measurement are based on two different definitions of measurement error and two different probability levels. The coefficient of reliability represents the expected difference between two repeated measurements due to measurement error with probability of 0.95. This statistic is that adopted by the International Standards Organization (ISO, 1994) and is useful when it is difficult to take multiple measurements on participants to conceptualize a 'true score'. The standard error of measurement represents the expected difference between a measurement and a hypothetical true score with a probability of 0.68. If it is possible to make hundreds of repeated measurements, then the average of these would be an estimate of the true score. Some authors have suggested multiplying the standard error of measurement (SEM) by 1.96, giving a probability of 0.95 that the true score is within $\pm 1.96\text{SEM}$ of the measurement

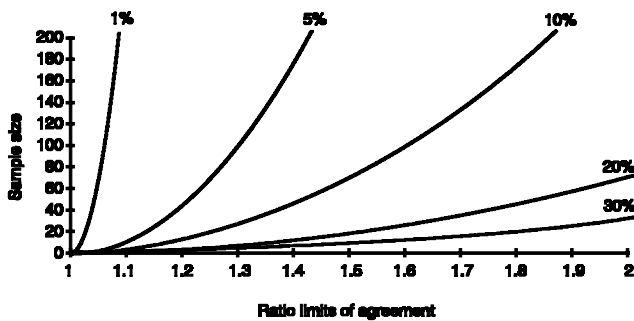


Fig. 2. A nomogram that can be used to estimate the effects of measurement error (limits of agreement) on whether 'analytical goals' are attainable or not in sport science research. The statistical power is 0.9. The different lines represent different worthwhile changes caused by some intervention of 1%, 5%, 10%, 20% and 30%. The measurement error statistic is the ratio limits of agreement. We stress that other measurement error statistics can be used in statistical power calculations.

One question relevant to the consideration of within-individual error is, how many trials are needed in a reliability study? The answer to this depends first on how many trials it takes for any systematic error to become negligible. If there are learning effects on a test, the reliability researcher should communicate to other researchers how many trials are necessary for familiarization. Analysis of variance approaches to describing random measurement error should not be performed until this investigation for systematic error is carried out.

It is possible that random error as well as systematic error is reduced with more trials. If an ANOVA is used to assess measurement error from, say, four trials in a reliability study, the error could be overestimated if it is smallest between the third and fourth trials. It might be more informative for the reliability researcher to conclude in such circumstances something like, 'two familiarization trials markedly reduce both systematic and random error. Following this familiarization period, the measurement error has been found to be x units. Based on statistical power calculations, this error variance is acceptable for workable sample sizes in future studies examining typical analytical goals'. It is for this reason that we advise in Table 2 the exploration of random error between separate pairs of test-retests (trial 1 *vs* trial 2, trial 2 *vs* trial 3, etc.) in a multiple-trial reliability study and the exploration of what the measurement error means to future research. We stress that this characteristic of reduced random error with more retests may show up as a violation of the sphericity assumption in repeated-measures ANOVA.

Sample size

This is probably the most well-known component of statistical power. In short, the larger the sample size, the greater the statistical power for a given meaningful effect and error variance. It is extremely useful for researchers to do a sample size estimation before conducting research, as it may indicate the most efficient workload for the researchers. Scientists do not want to waste time and money testing too many athletes. Conversely, it would be wasteful and, as the growing number of statisticians on ethics committees illustrates, unethical to test humans without any chance of detecting worthwhile effects. This conundrum has led some researchers to promote sequential 'sample size on the fly' methods (Hopkins, 1999), in which the researcher recruits more and more participants until a worthwhile effect is detected. This, of course, has many implications for statistical sampling of a population, in that, as time passes, there may be biases due to recruiting individuals who for some reason are less representative of the target population. As Chow (1996) noted, the failure to find a hypothesized worthwhile effect is just as much a reason to scrutinize the research design as it is the statistical power. For example, a type II error may have resulted from too few familiarization sessions to minimize test variability, rather than too few participants.

Confidence intervals

Recently, there have been several publications on how confidence intervals can help interpret study results (for detailed discussions, see Guyatt *et al.*, 1995; Curran-Everett *et al.*, 1998). These authors stressed that the null hypothesis test procedure tells one nothing about whether an effect is worthwhile, merely whether an observed effect is unusual. Although it is important not to ignore research design and the logic behind an hypothesis (Chow, 1996), it is apparent that confidence intervals do complement the null hypothesis test procedure in that they help estimate the magnitude of the population effect. Confidence intervals can be calculated with the aid of most statistical packages. Of greatest interest in an experiment on performance enhancement is the confidence interval of the *differences* between treatments, not the confidence interval for each treatment mean. To illustrate, briefly, how confidence intervals can complement the null hypothesis test procedure, we consider several examples.

Null hypothesis not rejected but possible worthwhile effect

A researcher finds that the mean (a point estimate) improvement in 10 km running performance following

an intervention is 5 s. This improvement was found not to be statistically significant ($P > 0.10$). The researcher calculates the 95% confidence interval (an interval estimate) to be -2 to 12 s. This interval means that there is a 95% probability that any sample difference will be between -2 and 12 units. A useful, but not entirely statistically accurate way of thinking about confidence intervals, is as the possible range of values within which the population difference is thought to lie. Suppose the researcher decides that a 6 s difference in performance justifies putting the athletes through the intervention. Instead of concluding that there is no effect, the conclusion should be that the real effect could be slightly worse (-2 s) or definitely worthwhile (12 s). Until more participants are tested or the error variance is reduced through a better experimental design (which will give the researcher more confidence in the estimation of the population difference), the researcher cannot really conclude whether the observed effect of 5 s is worthwhile or not.

Null hypothesis rejected but no worthwhile effect

Suppose the same researcher found a mean difference of 3 s and a confidence interval of 1–5 s. Here, an effect is present (null hypothesis rejected), since the lower bound of the confidence interval is above zero. However, the upper bound of the confidence interval is below the difference (6 s) designated as justifying putting the athletes through the intervention. Therefore, the researcher should not only state that an effect was found, but that the true population effect is unlikely to make a worthwhile difference to the athlete, given all that the athlete went through (the intervention might have been particularly arduous and costly). The danger here is that, without confidence intervals, the researcher may have regarded statistical significance itself as indicating an important effect.

These concepts are discussed in more detail by Curran-Everett *et al.* (1998) and Guyatt *et al.* (1995). A related issue that may be of interest to readers is clinical significance versus statistical significance (Curran-Everett *et al.*, 1998).

Overview

It is apparent that sport performance researchers should take great care in matching the particular aims of a study with the correct choice of dependent variable. Unlike clinical researchers, who need to predict the effects of interventions on the construct of 'health' by examining specific symptoms of disease (e.g. blood lipids for risk of heart disease), performance researchers may sometimes forget that they can measure final outcomes

(performance) directly, rather than relying solely on 'symptoms' (predictors) of good performance. This advantage brings with it many important considerations, including the external validity of the sample and test, the delimitation of a worthwhile performance enhancement, the choice of descriptive or intervention research, and adequate research design and analysis. We hope the topics covered in this paper are of interest and useful to sport performance researchers.

References

- Aeschleman, S.R. (1991). Single-subject research designs: some misconceptions. *Rehabilitation Psychology*, **36**, 43–49.
- Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley.
- Altman, D.G. (1991). *Practical Statistics for Medical Research*. London: Chapman & Hall.
- Atkinson, G. (in press). Analysis of repeated measurements in physical therapy research. *Physical Therapy*.
- Atkinson, G. and Brunskill, A. (2000). Effects of pacing strategy on cycling performance in a time trial with simulated head- and tail-winds. *Ergonomics*, **43**, 1449–1460.
- 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.
- Atkinson, G. and Nevill, A.M. (2001). Response to Lucia *et al.* *Medicine and Science in Sports and Exercise*, **33**, 852–853.
- Atkinson, G. and Reilly, T. (1996). Circadian variation in sports performance. *Sports Medicine*, **21**, 292–312.
- Atkinson, G., Coldwells, A., Reilly, T. and Waterhouse, J. (1994). The influence of age on diurnal variations in competitive cycling performances. *Journal of Sports Sciences*, **12**, 127–128.
- Atkinson, G., Nevill, A.M. and Edward, B. (1999). What is an acceptable amount of measurement error? The application of meaningful 'analytical goals' to the reliability analysis of sports science measurements made on a ratio scale. *Journal of Sports Sciences*, **17**, 18.
- Atkinson, G., Wilson, D. and Eubank, M. (2001). Effects of music on pacing strategy during a cycling time trial. *Medicine and Science in Sports and Exercise*, **33**, S158.
- Barlow, D.H. and Hersen, M. (1984). *Single-case Experimental Designs: Strategies for Studying Behaviour Change*, 2nd edn. New York: Pergamon Press.
- Bassett, D.R. and Howley, E.T. (1999). Limiting factors for maximum oxygen uptake and determinants of endurance performance. *Medicine and Science in Sports and Exercise*, **32**, 70–84.
- Birkeland, K.I., Stray-Gundersen, J. and Hemmersbach, P. (2000). Effect of rhEPO administration on serum levels of sTfR and cycling performance. *Medicine and Science in Sports and Exercise*, **32**, 1238–1243.
- 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*, **1**, 307–310.
- Bland, J.M. and Altman, D.G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, **8**, 135–160.
- Charter, R.A. (1997). Effect of measurement error on tests of statistical significance. *Journal of Clinical and Experimental Neuropsychology*, **19**, 458–462.
- Chow, S.L. (1996). *Statistical Significance: Rationale, Validity and Utility*. London: Sage.
- Cockerill, I.M., Nevill, A.M. and Lyons, N. (1991). Modelling mood states in athletic performance. *Journal of Sports Sciences*, **9**, 205–212.
- Critchley, L.A.H. and Critchley, A.J.H. (1999). A meta-analysis of studies using bias and precision statistics to compare cardiac output measurement techniques. *Journal of Clinical Monitoring and Computing*, **15**, 85–91.
- Curran-Everett, D., Taylor, S. and Kafadar, K. (1998). Fundamental concepts in statistics: elucidation and illustration. *Journal of Applied Physiology*, **85**, 775–786.
- Dillon, W.R. and Goldstein, M. (1984). *Multivariate Analysis: Methods and Applications*. New York: John Wiley.
- Drust, B., Reilly, T. and Rienzi, E. (1998). Analysis of work rate in soccer. *Sports, Exercise and Injury*, **4**, 151–155.
- Eklblom, B. and Berglund, B. (1991). Effect of erythropoietin administration on maximal aerobic power. *Scandinavian Journal of Medicine and Science in Sports*, **1**, 88–93.
- Eliaszewicz, M., Young, S.L., Woodbury, M.G. and Fryday-Field, K. (1994). Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. *Physical Therapy*, **74**, 777–788.
- Field, A. (2000). *Discovering Statistics Using SPSS for Windows*. London: Sage.
- Franklin, B.A. (1999). Cardiovascular responses to exercise and training. In *Exercise and Sports Science* (edited by W.E. Garrett and D.T. Kirkendall), pp. 107–116. London: Lippincott Williams & Wilkins.
- Gledhill, N. and Warburton, D. (2000). Haemoglobin, blood volume and endurance. In *The Encyclopaedia of Sports Medicine II: Endurance in Sport* (edited by R.J. Shephard and P.O. Åstrand), pp. 423–437. Oxford: Blackwell.
- Goodnight, J.H. (1979). A tutorial on the sweep operator. *American Statistician*, **33**, 149–158.
- Graham, T.E., Rush, J.W.E. and Van Soeren, M.H. (1994). Caffeine and exercise: metabolism and performance. *Canadian Journal of Applied Physiology*, **19**, 111–138.
- Guyatt, G., Jaeschke, R., Heddle, N., Cook, D., Shannon, H. and Walter, S. (1995). Interpreting study results: confidence intervals. *Canadian Medical Association Journal*, **152**, 169–173.
- Hamburg, M. (1970). *Statistical Analysis for Decision Making*. London: Harcourt.
- Harvill, L.M. (1991). An NCME instructional module on standard error of measurement. *Educational Measurement: Issues and Practice*, **10**, 33–41.
- Hocking, R.R. (1976). A biometrics invited paper: the analysis and selection of variables in linear regression. *Biometrics*, **32**, 1–49.
- Hofstra, W.B., Sont, J.K., Serk, P.J., Neijens, H.J., Kuethe, M.C. and Duiverman E.J. (1997). Sample size estimation in studies monitoring exercise-induced bronchoconstriction in asthmatic children. *Thorax*, **52**, 739–741.
- Hopkins, W. (1999). A new view of statistics. Online at: <http://www.sportsci.org/resource/stats/index.html> (accessed June 2000).
- Hopkins, W. (2000). Measures of reliability in sports medicine and science. *Sports Medicine*, **30**, 1–15.
- 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.
- ISO (1994). *Accuracy (Trueness and Precision) of Measurement Methods and Results. Use in Practice of Accuracy Values*. International Standards Publication ISO 5725-4. Geneva: ISO.
- Jones, A.M. (1998). A five year physiological case study of an Olympic runner. *British Journal of Sports Medicine*, **32**, 39–43.
- Jones, P. (2000). Single-case research and statistical analysis in school psychology and counseling. Online at: <http://www.unlv.edu/Colleges/Education/EP/scsaguid.htm#guide0> (accessed June 2000).
- Kleinbaum, D.G. (1994). *Logistic Regression: A Self-learning Text*. New York: Springer.
- Koutedakis, Y. (1995). Seasonal variation in fitness parameters in competitive athletes. *Sports Medicine*, **19**, 373–392.
- Lee, A.J. and Garraway, W.M. (2000). The influence of environmental factors on rugby football injuries. *Journal of Sports Sciences*, **18**, 91–95.
- Lord, F.M. (1984). Standard errors of measurement at different ability levels. *Journal of Educational Measurement*, **21**, 239–243.
- Lucia, A., Sanchez, O., Carvajal, A. and Chicharro, J.L. (1999). Analysis of aerobic–anaerobic transition in elite cyclists during incremental exercise with the use of electromyography. *British Journal of Sports Medicine*, **33**, 178–185.
- Martin, J.C., Milliken, D.L., Cobb, J.E., McFadden, K.L. and Coggan, A.R. (1998). Validation of a mathematical model for road cycling power. *Journal of Applied Biomechanics*, **14**, 276–291.
- Mathews, J.N.S., Altman, D.G., Campbell, M.J. and Royston, P. (1990). Analysis of serial measurements in medical research. *British Medical Journal*, **300**, 230–235.
- Maxwell, S.E. and Delaney, H.D. (1990). *Designing Experiments and Analysing Data*. Belmont, CA: Wadsworth.
- Morgan, W.B. (1985). Selected psychological factors limiting performance: a mental health model. In *Limits of Human Performance* (edited by D.H. Clarke and H.M. Eckert), pp. 70–80. Champaign, IL: Human Kinetics.
- Morrow, J.R. and Jackson, A.W. (1993). How ‘significant’ is your reliability? *Research Quarterly for Exercise and Sport*, **64**, 352–355.
- 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**, 314–318.
- Nevill, A.M. and Atkinson, G. (2001). Statistical methods in kinanthropometry and exercise physiology. In *Kinanthro-*

- pometry and Exercise Physiology Laboratory Manual*, 2nd edn. London: E & FN Spon.
- Noakes, T.D. (1997). Challenging beliefs: *Ex Africa semper aliquid novi*. *Medicine and Science in Sports and Exercise*, **29**, 571–590.
- Noakes, T.D. (1998). Maximal oxygen uptake: ‘classical’ versus ‘contemporary’ viewpoints: a rebuttal. *Medicine and Science in Sports and Exercise*, **30**, 1381–1398.
- Partick, T.D. and Hrycaiko, D.W. (1998). Effects of a mental training package on endurance performance. *Sport Psychologist*, **12**, 283–299.
- Peace, K.E. (1988). Some thoughts on one-tailed tests. *Biometrics*, **44**, 911–912.
- Pollard, R. and Reep, C. (1997). Measuring effectiveness of playing strategies at soccer. *The Statistician*, **46**, 541–550.
- Rice, W.R. and Gaines, S.D. (1994). Heads I win, tails you lose – testing directional alternative hypotheses in ecological and evolutionary research. *Trends in Ecology and Evolution*, **9**, 235–237.
- Schabert, E.J., Hopkins, W.G., Hawley, J.A., Mujika, I. and Noakes, T.D. (1998a). A new reliable laboratory test of endurance performance for road cyclists. *Medicine and Science in Sports and Exercise*, **30**, 1744–1750.
- Schabert, E.J., Hopkins, W.G. and Hawley, J.A. (1998b). Reproducibility of self-paced treadmill performance of trained endurance runners. *International Journal of Sports Medicine*, **19**, 48–51.
- Schutz, R.W. and Gessaroli, M.E. (1987). The analysis of repeated measures designs involving multiple dependent variables. *Research Quarterly for Exercise and Sport*, **58**, 132–149.
- Sears, B. and Harris, N. (1999). And statistics. Why Leeds are the away-day kings. *The Independent*, 18 September, p. 31.
- Senn, S., Stevens, L. and Chaturvedi, N. (2000). Repeated measures in clinical trials: simple strategies for analysis using summary measures. *Statistics in Medicine*, **19**, 861–877.
- Stevens, J.P. (1992). *Applied Multivariate Statistics for the Social Sciences*, 2nd edn. Hillsdale, NJ: Erlbaum.
- Swain, D.P. (1997). A model for optimizing cycling performance by varying power on hills and in wind. *Medicine and Science in Sports and Exercise*, **29**, 1104–1108.
- Thomas, J.R. and Nelson, J.K. (1996). *Research Methods in Physical Activity*. Champaign, IL: Human Kinetics.
- Vincent, J. (1999). *Statistics in Kinesiology*. Champaign, IL: Human Kinetics.
- Youngstedt, S.D. and O’Connor, P.J. (1999). The influence of air travel on athletic performance. *Sports Medicine*, **28**, 197–207.
- Zar, J.H. (1999). *Biostatistical Analysis*. London: Prentice-Hall.