Research Methods for Sports Performance Analysis

Peter O'Donoghue

CHAPTER 7

MEASUREMENT ISSUES IN PERFORMANCE ANALYSIS

INTRODUCTION

In Chapter 8, analysis of quantitative data using statistical procedures will be covered. In a scientific investigation, the correct use of the most appropriate descriptive and inferential statistical techniques allows data to be summarised effectively and conclusions to be drawn. However, in any scientific investigation, the analysis is meaningless if the raw data that have been gathered are invalid or inaccurate. In many commercial, industrial and legal scenarios when management information or other evidence is being considered, there are often important questions to be addressed about the sources of evidence. Are the data up to date? What are the definitions of the variables being used? Are the data accurately measured? Consider the example of the university league tables published by the national press in the UK. These have columns such as teaching quality, entry standard of students, research quality, student to staff ratio, library resources and IT resources. When these league tables are published they are often analysed by the senior management of universities. However, these senior officers will not blindly accept the position in the league table quoted for their university. The initial response is usually to question the data reported in the league tables. What is meant by teaching quality? How is it measured? What raw data are used? Are the data valid, reliable and up-to-date? What is meant by student to staff ratio? Are part-time staff included? Are part-time students included? In performance analysis of sport, there are similar measurement issues on which the acceptance of the whole study depends. These measurement issues include validity, objectivity and reliability. Therefore, this chapter precedes

the Chapter 8 on quantitative data analysis in recognition that the quality of measurement is of critical importance to a scientific study.

VALIDITY

The word 'validity' can refer to the validity of the whole study or it can refer to the validity of individual variables. When referring to the validity of a whole study, there is ecological validity, catalytic validity, internal validity and external validity, which have been described in Chapter 2. The validity of a variable used in a research study depends on its relevance and its reliability (Morrow Jr *et al.*, 2005: 82). The relevance of a variable is the degree to which the variable represents an important concept being measured. For example, is aerobic endurance an important concept to the area being studied and is VO_2 max a valid measure of aerobic endurance? Is flexibility an important concept to the research area and is the sit and reach test a valid measure of flexibility?

The reliability of a variable in performance analysis is the consistency with which the measurement procedure for the variable can be used by independent operators to measure the same performances. A variable that is not measured reliably cannot be valid, no matter how relevant the variable is to understanding sports performance.

Norm referenced validity

Morrow Jr *et al.* (2005: 80–125) and Thomas and Nelson (1996: 214–19) classified two broad types of validity: norm referenced validity and domain referenced validity. Norm referenced validity exists where a measured variable can be used to compare a player performance to norms for the relevant population of players. There are four categories of norm referenced validity.

1. *Logical validity or face validity* – is where the variable is valid by definition. This is often the case with performance variables such as 10km running time. There are many outcome indicators in sports performance that have logical validity as they are the score-related variables that the performers seek to maximise, minimise or optimise. The time required to complete a running, cycling, walking or swimming event is a logically valid performance indicator that performers seek to minimise. The distance that a field athlete jumps or throws an object is a logically valid performance indicator that the performer seeks to maximise. The angle of release of a javelin throw is a valid indicator of javelin throwing performance that must be an optimal angle that maximises the distance the javelin is thrown.

2. *Content validity* – is the extent to which the variable (or set of variables) covers different components of the concept of interest. Does a questionnaire

about worry cover all of the areas of worry? Does a test for referees cover all of the situations they will face in a game? In performance analysis investigations, the dependent variables of interest are often a set of performance indicators chosen to cover the broad aspect of sports performance that is of interest to the study. This broad aspect could be strategy, technique, technical effectiveness, work-rate or decision making. In analysing technique, there are many biomechanical indicators of technique including joint and angular displacements, velocities and accelerations as well as kinetic variables. The chosen biomechanical indicators will together have content validity if they cover all relevant details of the technique. A performance profile of technical effectiveness in a team game has content validity if it is composed of technical effectiveness variables for the key skills of that sport.

3. *Criterion validity* – is where the variable is validated against some gold standard measurement that has been accepted as a measure of the concept of interest. The reasons why the gold standard measurement itself cannot be used include the possibility that the gold standard is a very time-consuming measure to apply or involves the use of very expensive equipment or consumable resources. Thomas *et al.* (2005: 194–6) described two main contexts of criterion validity: concurrent and predictive. In concurrent validity, the measurement is correlated against some criteria administered to the same participants within the same study (concurrently). One example of this was the estimation of distance covered by soccer players used in a study by Martin *et al.* (1996). Pre-match speed measures were made and used as velocity multipliers, with the time recorded for different locomotive movements. The product of time and velocity gave an estimate of the distance covered. A more detailed and time-consuming time-motion analysis system was used to analyse the same video recordings of player performances with player locations being entered on an image of the playing surface. The estimates based on velocity multipliers were compared with those derived from entering the path travelled by the players. Predictive validity involves correlating the variable against some gold standard variable and determining a predictive model for the gold standard variable in terms of the variable being validated. Cross-validation is a type of predictive validity where a predictive model is determined based on a subset of the sample of participants and then tested using the remainder of the sample. The test of the predictive model involves making the gold standard measurement for each subject in the remainder of the sample and comparing the actual value with the predicted value using the model based on the variable being validated against it.

4. *Construct validity* – is the validity of some construct used to represent a property that is not directly observable. Construct validity is particularly important in sport and exercise psychology where areas such as anxiety, mood and confidence are measured using questionnaire instruments that compute overall scores for these areas as well as sub-dimensions of them. In performance analysis of sport, the best examples of where construct validity

may be needed are in the evaluation of psychological aspects of performance and in the evaluation of decision making, tactics and strategy. The strategy devised before a match and the moment-to-moment tactical decisions that are made during competition cannot be seen but may be inferred from observable behaviour. Correlation techniques (similar to those used in concurrent validity) can be used to compare the constructs with counts of behaviours one would associate with the construct. For example, if the profile of mood states were used to retrospectively gauge anger during a competition, it might be validated by examining its correlation with behaviours and body language use that would be associated with anger. The degree to which a construct distinguishes between different groups it would be expected to distinguish between (the known group difference method) is also used to evaluate construct validity. In performance analysis, valid outcome indicators would be expected to be different for winning and losing players. Measures of tactics and strategy would be expected to distinguish between athletes and teams who expert opinion would classify as adopting different tactics and strategy.

Criterion referenced validity

In addition to norm referenced validity, there is criterion referenced validity where the measure should accurately indicate whether the necessary level of proficiency has been reached. Decision accuracy is a common type of criterion referenced validity in performance analysis. An example of decision accuracy is the scoring of amateur boxing using the computerised system introduced by IABA (International Amateur Boxing Association) after the 1988 Olympic Games. The system is operated by five judges who use a red button and a blue button to record the punches that are deemed to satisfy the criteria for scoring punches by the boxer in red and the boxer in blue respectively. Where three or more judges press a button of the same colour within a second, a point is awarded to the boxer wearing that colour. The score that is output by this system typically under-estimates the actual number of scoring punches made by each boxer (Coalter *et al.*, 1999). However, as long as the score indicates a win to the boxer who made the most scoring punches, then the system has decision accuracy validity.

Processes of determining valid performance indicators

The dependent variables used in performance analysis investigations are often referred to as 'performance indicators', with some being referred to as 'key performance indicators'. Some students mistakenly refer to the raw performance data that is collected as performance indicators. When a point in a game of tennis is observed, it can be classified as (for example) an ace, a double fault, a serve winner, a return winner, a net point or a baseline rally (O'Donoghue and Ingram, 2001). However, the nominal variable

'point type' used to classify each point is not a performance indicator. The total number of aces served is not a performance indicator because some matches contain more points than others and will have a higher number of aces simply because of the increased number of service points. The percentage of service points where a player serves an ace is a possible performance indicator.

A performance indicator must represent some relevant and important aspect of sports performance in order to be valid. Identifying the valid performance indicators to use in a research project depends on a number of factors that are explained in this chapter. The validity of the performance indicators can be determined through expert coach opinion, review of coaching and performance analysis literature related to the sport of interest, relation to key outcome indicators or discrimination between performers of different levels. In undergraduate performance analysis research projects, there is not sufficient time to quantitatively investigate the validity of performance indicators, unless the whole purpose of the dissertation is to evaluate their validity. Therefore, undergraduate research projects typically select and justify the performance indicators used based on surveying coaches and performance analysis literature or by undertaking preliminary qualitative research to elicit performance indicators from expert coach opinion using a focus group or individual interview. When using performance analysis literature, students will often find that there are no standard performance indicators used in previous published research. For example, when one considers elite tennis strategy, Hughes and Clarke (1995) and O'Donoghue and Ingram (2001) used different variables. Hughes and Clarke (1995) used a combination of rally times, player positioning and shot placement as indicators of strategy. O'Donoghue and Ingram (2001) used rally times and the percentage of points where players attacked the net to characterise strategy. The student should consider which variables are most important to their research, the feasibility of possible methods for collecting the raw data required and the reliability of possible systems and methods that could be used to record the necessary data.

When using coaching literature, whether coaching science research sources or more practical texts and professional coaching resources, the student should consider the aspects of the sport being covered. In non-scientific sources, definitions may be vague and broad areas of technique, tactics, decision making or physical aspects may be written about without identifying any operationalised variables. Therefore, students should use such coaching literature to first identify broad areas of importance within the scope of their research question before considering how these areas can be represented by observable actions that can be counted, timed or assessed. If assessing the effectiveness or quality of an action, it is necessary to consider the number of 'grades' to be used and criteria to be associated with each. Morrow Jr *et al.* (2005: 138–41) provided examples of guidelines that can be used when setting grades.

Another way of determining performance indicators is to elicit important areas of performance from expert coaches. This can be done during exploratory interviews with individual coaches or using a focus group. The process of turning the identified areas into variables to be analysed within the investigation is similar to when the areas are identified using non-scientific literature. An example of this was an early study of rugby World Cup performance (McCorry *et al.*, 1996) where a rugby expert was initially interviewed about areas of the game that were important to concentrate on when describing the performances in international rugby. This interview was interspersed with periods of watching a video recording of a rugby match, allowing the expert to explain and identify behaviours that were of the greatest importance.

Quantitative methods have also been used to establish the validity of variables used in performance analysis. The process of establishing validity in this way often amounts to gathering the volume of data that a student would be expected to do during an undergraduate research project. Therefore, establishing validity in this way is rarely used as part of an undergraduate student project that has a wider purpose of describing the chosen area of sports performance. It is possible that an undergraduate research project could have the sole purpose of validating a set of performance indicators. At Master's and PhD level, such a validation study could be one of a series of studies that make up the overall research (Choi, 2008). There are different ways in which quantitative methods can be used to establish the validity of performance variables. These include neural networks (Choi *et al.*, 2006b), multiple regression (Choi *et al.*, 2006b), correlation analysis (O'Donoghue, 2002), binary logistic regression, discriminant function analysis and principal components analysis (O'Donoghue, 2008a).

Multiple regression techniques identify the relative contribution of each process indicator in predicting the chosen outcome indicator (Choi *et al.*, 2006b). Choi *et al.* (2006b) found multiple regression to be a more successful predictor of outcome indicators in elite tennis than artificial neural networks. Artificial neural network techniques are also more complex, difficult to use and describe in methods sections of research reports.

Known group difference is a way of establishing the validity of process indicators that can be done using inferential statistical tests. If candidate process indicators are claimed to distinguish between winning and losing performers within matches, statistical tests can be used to confirm or refute this. Similarly, successful and unsuccessful performers can be identified based on finishing position within tournaments and process indicators can be compared between them.

Some valid process indicators are not expected to have an association with match outcome. For example, in tennis there are players who adopt a net strategy in all parts of the World rankings. Similarly, there are players who adopt a baseline strategy in all parts of the World rankings. It is important in practice to understand whether an opponent plays using a net or a baseline strategy. The fact that the percentage of points where a player

attacks the net may not be associated with the percentage of points won in a match does not mean that this process indicator is invalid. Similarly, there will be soccer teams that adopt a slow build-up style of play at all levels of the sport and there will be soccer teams that adopt a more direct style of play at all levels of the sport. It is important in practice for soccer squads to have an understanding of the style of play of their opponents even though process indicators representing playing style may not be associated with match outcome.

Statistical techniques for establishing criterion validity and techniques for establishing known group difference often produce sets of process indicators that are not entirely independent (Choi *et al.*, 2006b). Therefore, a more efficient analysis of the given sport can be undertaken if a more concise set of independent process indicators can be identified. Principal components analysis is a data reduction technique that allows a smaller set of principal components to be identified that are uncorrelated variables representing different dimensions in the data. O'Donoghue (2008a) proposed a way in which principal components analysis could be used to determine a set of performance indicators in tennis that represented independent aspects of performance in the sport.

The set of chosen performance indicators should be concise enough to support effective communication but should also have content validity covering all relevant aspects of the area of performance of interest. The performance indicators chosen dictate the action variables that will be used during data gathering. However, an increased number of performance indicators does not necessarily mean that there will be an increased volume of data entry. Consider the POWER system (O'Donoghue *et al.*, 2005a) where operators use two function keys to record when each period of 'work' and 'rest' commences. Originally, this system reported the frequency, mean duration and percentage observation time for 'work' and 'rest'. The enhanced system described by O'Donoghue *et al.* (2005a) included outputs for the frequency of 'work' periods of seven different duration ranges, the frequency of 'rest' periods of eight different duration ranges and 72 frequency variables for each combination of 'work' period duration and following 'rest' period duration. These additional outputs did not require any additional data entry activity by the operators.

DO PRECISE OPERATIONAL DEFINITIONS GUARANTEE GOOD RELIABILITY?

It is essential for system operators and the eventual consumers of the information generated by performance analysis that there is a shared understanding of the variables used. Therefore, there is a view that these variables

should be defined with a level of precision that makes their meaning unambiguous (Williams, 2009). The operational definitions used in performance analysis do not have the detail seen in legal contracts, mobile phone contracts or gym membership contracts, and readers are encouraged to examine such contracts to gain an appreciation of how difficult it is to achieve legally binding precision. When it is essential to understand the requirements for software-intensive systems, rigorous formal specification techniques are used based on set theory, mapping functions and temporal logic (O'Donoghue and Murphy, 1996). It may not be wise for operational definitions in performance analysis systems to achieve such levels of unambiguous detail. Performance analysis is often a real-time observational task where there is not sufficient time to consider the level of detail one would see in a legal contract, which is inspected carefully over a much longer timescale before determining if it has been conformed to or breached. Furthermore, such a level of detail would give students and other researchers a serious problem in describing their methods within the required word limits.

O'Donoghue (2007b) provided evidence from two research studies that challenges the traditional view that operational definitions are essential in performance analysis research. One study was an investigation of international rugby union performance (Armitage, 2006) and the other was an analysis of defensive styles used in international netball (Williams and O'Donoghue, 2006). Armitage (2006) produced detailed definitions of named states and transition events in rugby union. The terms defined included 'possession gained during set play', 'possession gained during loose play', 'gaining territory by going over', 'gaining territory by going through' and 'gaining territory by going around' the opposition. The operational definitions came to eight double-sided pages of Armitage's (2006) thesis, with terms introduced during the operational definitions, such as 'gain line', also being defined. Armitage and his supervisor (the author of this book) participated in an inter-operator reliability study. First, the operational definitions were read, discussed and agreed before the final of the 2003 Rugby World Cup was analysed independently by the two observers. The inter-operator reliability study revealed serious limitations in reliability despite the efforts made prior to the inter-operator agreement study. The two operators discussed the reasons why they disagreed so much and what each of them was counting for each type of event. It was concluded that it would have been useful to discuss the operational definitions while viewing example video sequences so the operational definitions would then be considered in terms of the analysis task; an observational task. This would have facilitated learning by both operators of the types of observed behaviours that would be counted for each class of event and for what reasons. One problem with using video observation in this way is that it creates an issue for the replicability of the study as readers of the final report will only see the operational defini-

tions without seeing the example video sequences of the defined behaviours.

The second example used by O'Donoghue (2007b) to challenge the traditional view that operational definitions are essential in performance analysis research was a completely opposite situation to Armitage's (2006) study. Williams and O'Donoghue (2006) did not use any operational definitions of a key independent variable and yet achieved a 100 per cent inter-operator agreement for that independent variable. The independent variable was the style of the defence used during opposition possessions in netball. The definitions provided for the four different styles of defence were as follows:

1. Zone defence – where all players concerned in the area of play analysed are marking the space.
2. Man-to-man defence – where all players concerned in the area of play analysed are marking a player
3. Part man-to-man/part zone defence – where some players concerned in the area of play analysed are marking the space and some players concerned are marking a player.
4. Other – where the defence could not be classified in to one of the man-to-man or zonal strategies.

These definitions fail to precisely define the point at which a pattern of seven netball players playing against an opposing pattern of seven netball players changes from one type of defence to another. This is an example of a valid aspect of sports performance that cannot be described precisely or practically in words. One reason for the high level of agreement obtained was that both operators were experienced netball players, coaches and qualified umpires who were very familiar with the terms used and the types of defensive play that counted for each.

Ultimately, any performance analysis technique that involves human operation will involve an element of subjective classification of behaviour. Where fully automated systems such as Hawkeye (Hawkeye Innovations, Winchester, UK) are used in performance analysis research, it is possible to use precise operational definitions for variables. There are examples of situations in sport that are easier for human operators to classify than others. For example, a tennis ball landing in court or out of court is usually straightforward to observe. Furthermore, the observer may decide to use the outcome of the point decided by the line judges and umpire officiating the match. Other aspects of behaviour such as quality of technique, type of technique and locomotive movement classification involve subjective classification by a human observer. Therefore, the timings and frequencies recorded during match observation are often quantitative counts of subjective judgements.

OBJECTIVITY AND RELIABILITY

Reliability of measurement is important in all disciplines of sport and exercise science. In performance analysis of sport, reliability of measurement is a common feature of work in both notational analysis and biomechanics (Bartlett, 2001). Objectivity and reliability are related concepts in performance analysis of sport. A performance indicator is objective if its value is independent of the opinion of individual operators.

The first step in achieving objectivity is to use an objective measurement procedure that does not rely on the subjective opinion of an individual observer. As has already been mentioned, this is often not possible in performance analysis of sport where human observers form part of the measurement system. Objectivity requires the use of definitions for all action variables that will be used in the calculation of a performance indicator. For example, if we have a performance indicator that is the percentage of points where a tennis player goes to the net, we must define what counts as a point and we must define what counts as a net point. O'Donoghue and Ingram (2001) excluded lets (points that had to be replayed) from their study of Grand Slam singles tennis and defined a net point as a point where a player crossed the service line (the line at the back of the service boxes) and there were still one or more shots to be played in the point.

A reliability test seeks to evaluate the consistency with which a method can be used. In other disciplines, such as exercise physiology, test-retest reliability tests are done using the same participants who perform a test more than once. If consistent results are produced for repeated application of the test then the test can be deemed to be reliable. Sports performance variables are different to anthropometric and fitness test variables in that sports performance does not take place under controlled conditions. This allows test-retest studies to be used to investigate the reliability of anthropometric measurement methods and fitness tests. Performance in many sports is unstable and performance indicator values are influenced by many factors including:

- venue;
- players available for selection;
- quality of the opposition;
- importance of the match;
- weather conditions;
- tactical decisions made by coaches and players.

The main factor responsible for variability in sports performance is the quality of the opposition (McGarry and Franks, 1994). One study of variability in sports performance found that David Beckham's percentage match time spent performing high intensity activity in 11 different English FA

Premier League soccer matches had a greater variability than that observed between 115 different English FA Premier League midfielders (O'Donoghue, 2004). Therefore, test-retest reliability studies are not used to evaluate the reliability of action variables, performance indicators or methods of gathering such data. Test-retest studies can be used in investigations of stability of performance, but this is a different issue to the reliability with which a performance can be analysed. Indeed, to be able to study match-to-match variability in player performance, it is essential that the method use to analyse match performances is reliable.

Reliability studies in performance analysis of sport use independent observations of the same performance. This can be done live during a competition or through post-match analysis of the competition. There are two types of reliability study that have been used in performance analysis research: intra-operator agreement studies and inter-operator agreement studies. Intra-operator agreement studies are often used in undergraduate student research projects where it is not feasible to train another operator to use the system developed for the project. However, intra-operator agreement studies are of limited value as they fail to demonstrate the objectivity of the system. A good level of intra-operator agreement in a time-motion analysis system will simply indicate that the particular operator can consistently classify locomotive movement into the defined movement classes. The intra-operator agreement study does not answer the question of whether anybody else using the system to analyse the same performance would obtain similar results. Therefore, inter-operator agreement studies are preferable. Intra-operator agreement studies can be useful for the student during pilot testing of the system.

Inter-operator agreement studies allow the objectivity of systems to be established. They demonstrate that systems can be used consistently, recording data that are independent of individual operator perceptions. Although inter-operator agreement studies are more difficult to organise than intra-operator agreement studies, it is recommended that performance analysis project students use inter-operator agreement tests. One way of doing this is for a pair of students to operate both students' systems during the inter-operator agreement tests. This requires the systems to be described and action variables to be specified to a high enough standard for the systems to be used by operators other than the students who developed the systems. End-user training will have to take place before the inter-operator agreement test can take place.

There is a question as to what data should be included in reliability studies. One point of view is that each performance indicator to be used when presenting results and conclusions must have its level of reliability described. There is another point of view that the observation process should be demonstrated to be reliable and then any performance indicators derived using the method can be deemed reliable through a process of inductive reasoning. This author would not wish to outlaw the second approach as

there are sound reasons for using it. The *Journal of Sports Science and Medicine* publishes research in various areas of sports science, including sports performance. The word limit for original research articles is 3,800 including 300 words for each table or figure that is used. This prohibits a complete certification of reliability of every performance indicator to be used in many studies. Students of undergraduate research projects face similar pressures. There may be an overall word limit of around 10,000 words, which can allow students to devote more words to the description of the reliability study in the methods chapter, but this may be at the expense of using words in the discussion for which there are a lot of marks allocated.

The next two sections of this chapter cover the different types of reliability statistic that can be used in performance analysis. These can be used to evaluate intra-operator agreement or inter-operator agreement. The reliability statistic to be used depends on the scale of measurement of the action variable or performance indicator of interest. The four scales of measurement described in statistics textbooks (Vincent, 2005: 6–7) are nominal scale, ordinal scale, interval scale and ratio scale. The first two scales are categorical scales of measurement as variables use values from a finite set of named values. The difference between nominal scale variables and ordinal scale variables is that the values of a nominal scale variable are just different values (for example male and female), whereas ordinal scale variables use values that have a defined order (for example 'never', 'seldom', 'sometimes', 'often' and 'always'). In performance analysis, we further subdivide nominal scales into two different types of nominal scale. The first is where any error involving two values is equally serious. The second type of nominal scale is where there are some pairs of values that are 'neighbouring' and others that are not. An example of a nominal variable with neighbouring values is area of the playing surface. This is not an ordinal variable, but there will be some pairs of areas (or cells) of the playing surface that border each other and others that do not. Where an action occurs on the border of two cells, one operator may record one of the cells and the other operator might record the other cell. This is a less serious error than when an error involves two cells that are not even bordering each other. Computerised systems such as SPSS merge interval and ratio scales into a single numerical scale (called 'scale' in SPSS). This has little impact on the use of descriptive and inferential statistics for the main purpose of a performance analysis project. However, interval and ratio scale variables do need to be considered differently for the purposes of reliability assessment, as some reliability statistics that can be used with ratio scale variables may not be valid for use with interval scale measures. Therefore, there are five scales of measure that we consider in performance analysis research:

1. nominal scale with no neighbouring values;
2. nominal scale with neighbouring values;

3. ordinal scale;
4. interval scale;
5. ratio scale.

RELIABILITY STATISTICS FOR CATEGORICAL VARIABLES

Reliability of nominal scale variables

The reliability of nominal scale variables will be illustrated using an example from a study of elite tennis strategy (Brown and O'Donoghue, 2008b); the variable of interest is point type. The data used in the inter-operator agreement study were from 1,356 points played in 12 different Grand Slam singles matches with at least one men's match and at least one women's match from each of the four Grand Slam tournaments. Hughes *et al.* (2004) proposed the use of a percentage error calculation to evaluate the reliability of nominal variables that computes a single percentage error value. However, in Table 7.1, the current author has also shown the percentage errors for the frequency of each nominal value. The method used by Hughes *et al.* (2004) is shown in equation 7.1 and expresses the sum of the absolute difference in frequencies as a percentage of the sum of the mean frequencies recorded by the two operators (f1 and f2):

$$\text{per cent error} = \Sigma \mid f1 - f2 \mid / (\Sigma\ (f1 + f2)/2) \qquad (7.1)$$

The total percentage error in this example is 4.9 per cent, which is within the typical stated level of error of 5 or 10 per cent. However, as Table 7.2 shows, this total percentage error conceals the fact that there are two values of point type with percentage errors greater than 10 per cent. One of the problems with using percentage error as a reliability statistic is that it may conceal errors that cancel each other out because only the totals are used. For example, Table 7.2 reveals that there were 690 points that both operators agreed were baseline rallies. There were 12 points that the first operator classified as points where the server went to the net and the second operator classified as baseline rallies. When inspecting the total frequency of each point type recorded by the two operators, we see that these 12 errors were cancelled out by 12 of the 33 points that the second operator classified as points where the server went to the net and that the first operator classified as baseline rallies.

The kappa statistic (Cohen, 1960) is a measure of reliability for nominal variables that can be used when we have a chronologically ordered sequence of values for each operator. Where we have accumulated totals through a

Table 7.1 Percentage error of point type in tennis in an inter-operator agreement test (data from Brown and O'Donoghue, 2008b)

Point Type	*Operator 1*	*Operator 2*	*Absolute Difference*	*Mean Value*	*Percentage Error*
Ace	79	80	1	79.5	1.3
Double Fault	42	41	1	41.5	2.4
Serve Winner	237	243	6	240	2.5
Return Winner	19	19	0	19	0.0
Server to Net	169	151	18	160	11.3
Receiver to Net	92	78	14	85	16.5
Baseline Rally	718	744	26	731	3.6
Total	1356	1356	66	1356	

Table 7.2 Inter-operator reliability table for point type (data from Brown and O'Donoghue, 2008b)

First Operator	*Second Operator*							
	Ace	*DF*	*S.W*	*R.W*	*S.Net*	*R.Net*	*BL*	*Total*
Ace	78						1	79
DF	1	38		1			2	42
S.W	1		232	1	1		2	237
R.W		1		17			1	19
S.Net			2		131	3	33	169
R.Net					7	70	15	92
BL		2	9		12	5	690	718
Total	80	41	243	19	151	78	744	1356

Key:

DF	Double Fault
S.W	Serve Winner
R.W	Return Winner
S.Net	Server attacks net first
R.Net	Receiver attacks net first
BL	Baseline rally

tallying system, the cross-tabulation shown in Table 7.2 cannot be produced. Kappa determines the proportion of points that are agreed by the two operators excluding the proportion that one would expect to be agreed by chance. This is a very important consideration where one value (such as baseline rally) is dominant. If the first operator realised the most common point type was baseline rally and guessed this for every point, the agreement with the second operator would still be greater than 50 per cent of points

(744/1,356). In our example, there were 1,256 of the 1,356 points where the operators agreed on point type (92.6 per cent). Therefore, the proportion of points where the operators agreed on point type, $p_0 = 0.926$. However, this includes many points where the operators could expect to agree by guessing. For example, the fraction of points that the first operator classifies as aces is 79/1,356. Therefore, we could expect this fraction of the 80 points where the second observer classified the point as an ace to be agreed by guessing. Applying this logic to each of the seven point types allows the number of points expected to be agreed by guessing to be determined as shown in Table 7.3.

This gives a proportion that could be expected to be agreed by guessing, $p_C = 466.7/1{,}356 = 0.344$. If we consider p_0 to be itself divided by one, the proportion of points agreed can be adjusted by subtracting the expected proportion agreed by guessing, p_C, from the top and bottom of this division to give kappa, κ. Cohen's (1960) original kappa statistic is shown in equation 7.2. The equations 7.3 and 7.4 define the proportion of agreements, p_0, and the expected proportion of agreements by chance, p_C, respectively. In these equations, V is the number of different nominal values for the performance indicator of interest and N is the total number of cases recorded by each observer. A is a two-dimensional table of V x V values representing the number of occasions the observers recorded each value, with $A_{i,j}$ representing the number of occasions that the first observer recorded the ith nominal value and the second observer recorded the jth nominal value.

$$\kappa = \frac{p_0 - p_C}{1 - p_C} \quad (7.2)$$

$$p_0 = \frac{\sum_{k=1}^{V} A_{k,k}}{N} \quad (7.3)$$

$$p_C = \frac{\sum_{k=1}^{V} \left(\frac{\left[\sum_{i=1}^{V} A_{i,k} \right] \left[\sum_{j=1}^{V} A_{k,j} \right]}{N} \right)}{N} \quad (7.4)$$

Table 7.3 Frequency distribution of point type in tennis in an inter-operator agreement test (data from Brown and O'Donoghue, 2008b)

Point Type	*Calculation*	*Expected agreements by guessing*
Ace	80 x 79/1356	4.7
Double Fault	41 x 42/1356	1.3
Serve Winner	243 x 237/1356	42.5
Return Winner	19 x 19/1356	0.3
Server to Net	151 x 169/1356	18.8
Receiver to Net	78 x 92/1356	5.3
Baseline Rally	744 x 718/1356	393.9
Total		466.7

Table 7.4 Interpretation of kappa values (Altman, 1991: 404)

Kappa	*Strength of agreement*
$\kappa \geq 0.8$	Very good
$0.6 \leq \kappa < 0.8$	Good
$0.4 \leq \kappa < 0.6$	Moderate
$0.2 \leq \kappa < 0.4$	Fair
$\kappa < 0.2$	Poor

Therefore, in this example of point type, the kappa value for inter-operator agreement of point type was (0.926 - 0.344)/(1 - 0.344) = 0.888. This is a very good strength of agreement according to the Altman's (1991: 404) evaluation scheme for kappa shown in Table 7.4.

Reliability of nominal scale variables with neighbouring values

Very often in performance analysis, the area of the playing surface in which an event occurs is recorded. Hughes and Franks (2004c) described examples of systems where the area of the playing surface where events are performed is recorded for field hockey and squash. Hughes and Franks (2004d) provided further examples of systems where the playing surface is divided into cells for tennis, basketball, soccer and netball. Figure 7.1 illustrates how the netball shooting circle is decomposed for the purposes of analysis (Robinson and O'Donoghue, 2007).

Where a player shoots from an area on the border between areas 2 and 5, one observer might record area 2 while the other might record area 5. Kappa treats this disagreement as a total disagreement the same way as it

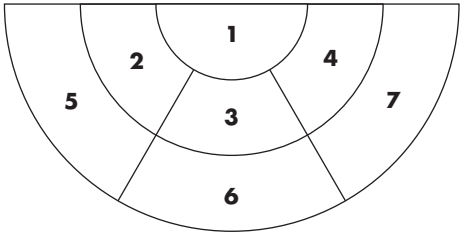

Figure 7.1 Areas of shooting circle of a netball court

would treat a disagreement between areas 2 and 7, which do not border each other. Where the areas recorded by two observers border each other, the observers should receive some credit. This can be done using Cohen's (1968) weighted kappa statistic, which introduced the two-dimensional array of V x V weights. This array is referred to as W. Equation 7.2 is still used to determine κ using p_0 and p_C, but p_0 and p_C are calculated differently using equations 7.5 and 7.6 respectively.

$$p_0 = \frac{\sum_{i=1}^{V}\sum_{j=1}^{V} A_{i,j}W_{i,j}}{N} \qquad (7.5)$$

$$p_0 = \frac{\sum_{i=1}^{V}\sum_{j=1}^{V}\left(\frac{\left[\sum_{i=1}^{V} A_{i,k}\right]\left[\sum_{j=1}^{V} A_{k,j}\right]}{N}W_{i,j}\right)}{N} \qquad (7.6)$$

In Robinson and O'Donoghue's (2007) example, the weighted version of kappa actually resulted in a slightly reduced value for kappa with the weights that they used. The calculation of the weighted version of kappa is illustrated in the following section, where it also applies.

Reliability of ordinal scale variables

Ordinal variables have all of the properties of nominal variables, but additionally values on an ordinal scale are ordered. An example of an ordinal variable is aggressiveness with which a shot is played in tennis (Boffin, 2004). Table 7.5 shows that Boffin used seven different levels of aggressiveness and used the numbers 1 to 7 to rate them. This variable is clearly

Table 7.5 An ordinal scale for rating aggressiveness of tennis shots (adapted from Boffin, 2004)

Rating	*Guidelines*
1	Totally defensive. Shot has no depth, spin or power. Barely clears net. Hit out of position on stretch, only aim to get ball in play somehow.
2	Defensive lob. High over net with no pace or spin, but element of control allowing player time to get back into central court position, hit on stretch.
3	Shot hit with limited pace, may have spin, landing about the service line, often hit from behind the baseline, allowing opponent to move forward and be aggressive.
4	Basic neutral rally shot. Landing centrally in court with no attempt to manoeuvre opponent. Lands midway between service line and baseline, mid-pace, will have spin, hit from behind the baseline.
5	Shot played from about the baseline with the aim of moving opponent around. Hit into space away from opponent, either drop shot or ball aimed into open court.
6	Hit with increased pace. Player aiming for shots with lower margin for error, either very deep towards the baseline, very close to the sidelines (literally an inch or two from the lines) or aggressive short angles pushing opponent fully out of court. Usually hit whilst moving forward.
7	Totally aggressive shot. Looking directly for outright winner, low over net, away from opponent or flat out at opponent. Hit as a result of opponents defensive shot (1, 2 or 3) therefore player usually mid-court when hitting shot.

ordinal with increasing levels of aggressiveness from ratings 1 to 7. However, just because the interval between the numbers used is 1, does not mean that this is an interval scale measure. The numbers 1 to 7 merely represent the different values of the scale.

Table 7.6 shows the cross-tabulation of 60 shots rated by two independent operators of Boffin's (2004) system. There are 44 agreements out of 60 shots analysed (p_0 = 0.733) with the proportion where we could expect agreement by guessing to be p_C = 0.164. Therefore kappa value is 0.681, indicating a good strength of inter-operator agreement.

The unweighted version of kappa (Cohen, 1960) is a particularly harsh measure of reliability for this type of variable because minor disagreements (like ratings of 6 and 7 being made by the operators) are treated the same as serious disagreements (like ratings of 1 and 7 being made by the operators). In ordinal scales, every adjacent pair of values is a pairing of neighbouring values. This should be recognised by giving some credit where minor disagreements have occurred between the two operators. Therefore, weights such as those shown in Table 7.7 should be used within Cohen's (1968) weighted kappa statistic for evaluating the reliability of ordinal variables.

Table 7.6 Inter-operator reliability of rating of aggressiveness of tennis shot (fictitious data using Boffin's, (2004) method)

Operator 1 rating	*Operator 2 rating*							
	1	*2*	*3*	*4*	*5*	*6*	*7*	*Total*
1	2	1						3
2		3		1				4
3			8	3				11
4				*7*	2	1		10
5					8	4		12
6					1	*7*	2	10
7						1	9	10
Total	2	4	8	11	11	13	11	60

This increases p_0 to 51.0/60 = 0.850, p_C to 18.58/60 = 0.310, leading to an increased kappa value of 0.783, which still represents a good strength of agreement between the operators.

Reliability statistics for ratio scale variables

Correlation coefficients are measures of relative reliability that reflect how well values maintain their ranks during reliability studies. This can either be during a test-retest reliability study in physiology, for example, or during independent observations of the same performances in a performance analysis

Table 7.7 Weightings to be used when applying the weighted kappa statistic to inter-operator reliability of a shot aggressiveness in tennis

Operator 1	*Operator 2*						
	1	*2*	*3*	*4*	*5*	*6*	*7*
1	1.0	0.5					
2	0.5	1.0	0.5				
3		0.5	1.0	0.5			
4			0.5	1.0	0.5		
5				0.5	1.0	0.5	
6					0.5	1.0	0.5
7						0.5	1.0

reliability study. Relative reliability studies should not be used on their own because they do not describe error values. Measures of absolute reliability are concerned with the amount of error expressed in the units of measurement or as a percentage of the value recorded. Measures of absolute reliability that are used with numerical scale variables are mean absolute error, root mean square error, percentage error, standard error of measurement, 95 per cent limits of agreement and 95 per cent ratio limits of agreement.

There are two numerical scales of measurement of interest to performance analysis: the interval scale and the ratio scale. Interval scales of measurement have all of the properties of ordinal scales except that in addition to the values being ordered, there is a fixed interval between values. Therefore, subtraction is meaningful when using interval scales whereas only comparison (greater than and less than comparison) can be used with ordinal scales. Displacements, velocities, accelerations, angular velocities and angular accelerations are all examples of interval scale variables that can be used in performance analysis. Some of these can have negative values as well as positive values, meaning that division is not always meaningful. Ratio scale variables possess all of the properties of interval scale variables except that they also use zero to represent a total absence of value. Therefore, division and ratios of values are meaningful. Distance, time, height and mass are measured on ratio scales, with 6m (or s or kg) being twice as long (or heavy) as 3m (or s or kg). Ratios and division such as this cannot be done validly with variables measured on interval scales that are not also ratio scales. Ratio scale variables are more common in performance analysis research then interval scale variables and, therefore, the author has chosen to describe reliability statistics that can be used with ratio scale variables before describing reliability statistics for interval scale variables.

In this section, 100m split times during an 800m track athletics event are used as an example of reliability evaluation for ratio scale variables. The data are the split times recorded during Brown's (2005) study of the 2004 Olympic women's 800m. The final of the women's 800m was analysed by two independent operators using Brown's manual system. Once elapsed times are recorded, split times can be determined for each athlete during each 100m section of the race. Figures 7.2 and 7.3 plot the data recorded by the two operators against each other for elapsed times and split times respectively. There are a total of 64 value pairs as there were eight split times for each of the eight athletes participating in the race. If Pearson's r is used as a measure of relative reliability (how well the values maintain their rankings within the set of 64 values), the use of elapsed times gives an artificially high value of $r > 0.999$. Basically, the elapsed times include eight 100m timings, eight 200m timings, and so on, up to eight 800m timings. Essentially, there are eight different variables, whose values form clusters within the scatter plot shown in Figure 7.2.

Figure 7.3 plots the split times recorded by the two operators against each other. There are a total of 64 value pairs as there were eight split times

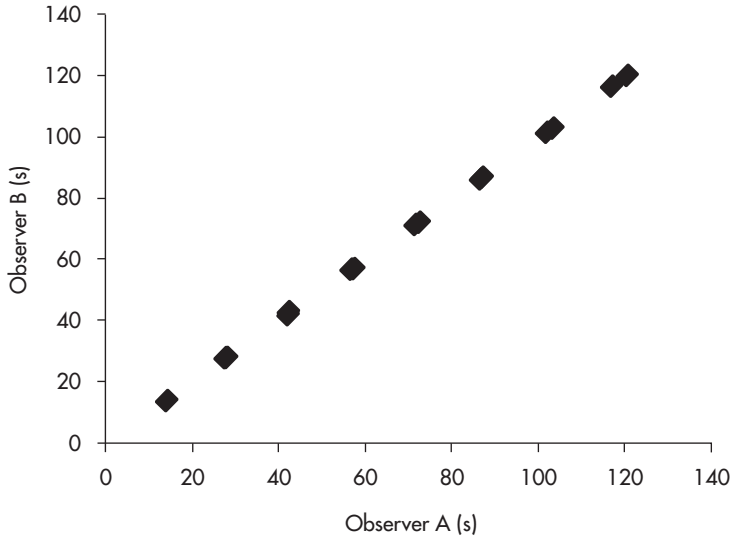

Figure 7.2 Relative reliability of elapsed time in 800m running (using data from Brown 2005)

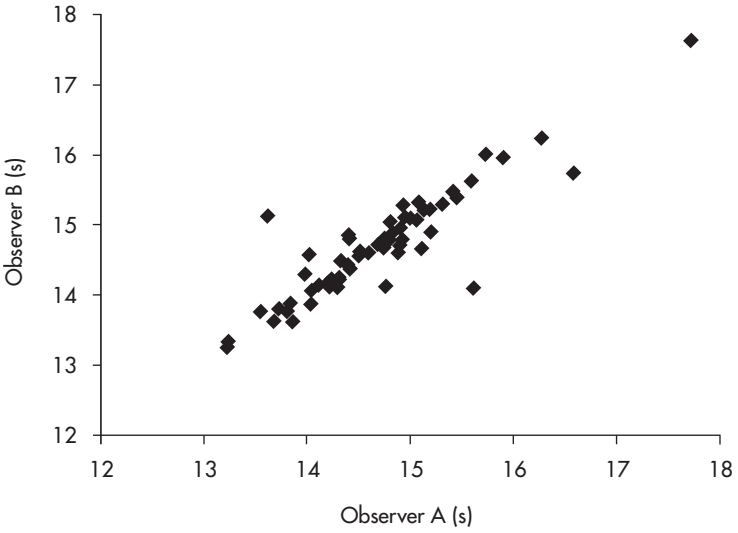

Figure 7.3 Relative reliability of split time in 800m running (using data from Brown 2005)

for each of the eight athletes participating in the race. Pearson's r is used as a measure of relative reliability (how well the values maintain their rankings within the set of 64 values) of the split times (r = 0.897).

The data from the reliability study are used to illustrate change in the mean, standard error of measurement (SEM), 95 per cent limits of agreement, percentage error as well as mean absolute error. Percentage error is determined by expressing the absolute error between recorded values as a percentage of the mean of the two values. The mean is used because there may be no reason to expect one observer's value to be more accurate than the other's. Hughes *et al.* (2004) used equation 7.7 to determine total percentage error where multiple frequency variables were recorded. In equation 7.7, |x| represents the magnitude of x; that is the sign is removed. Note that in equation 7.7, A_i and B_i are *i*th values recorded by two independent operators *A* and *B* out of a total on *N* values recorded by each. The use of this equation can be criticised because it does not actually use any percentage errors in the calculation of total percentage error. A single total of all absolute errors is expressed as a percentage of a single total of mean values between the two observers. Therefore, a high frequency variable such as passes will dominate the divisor of this equation and conceal large percentage errors for other frequency variables such as tackles, crosses and shots. The percentage error should be determined for these frequency variables individually.

$$Total\ \%\ Error = 100\ \frac{\sum_{i=1}^{N} |A_i - B_i|}{\sum_{i=1}^{N} (A_i + B_i)/2} \quad (7.7)$$

Percentage error values can be combined where they are percentage error values from the same variable. For example, 100m split time during an 800m race could be combined to determine a mean percentage error, as shown in equation 7.8.

$$Mean\ \%\ Error_1 = 100\left[\sum_{i=1}^{N} \frac{|A_i + B_i|}{(A_i + B_i)/2}\right] /N \quad (7.8)$$

One criticism of mean percentage error 7.8 is that it divides the absolute difference by the mean of the two values and can thus determine low values for percentage error. The problem with this is that the full range of values from 0s to the mean is not the meaningful range of values that would be

observed for 100m split times; an athlete will not run any 100m section of an 800m race faster than 10s and will certainly not cover 100m in 0s. An alternative is to specify a theoretical maximum value, *Max*, of say 18s, as well as a theoretical minimum, *Min*, and express the inter-operator disagreement as a percentage of the range of meaningful values from *Min* to *Max*, as in equation 7.9.

$$Mean\ \%\ Error_3 = 100\left[\sum_{i=1}^{N} \frac{|A_i + B_i|}{Max - Min}\right] /N \quad (7.9)$$

Equations 7.8 and 7.9 are used in Table 7.8 to determine percentage error yielding 64 percentage values in each case. The first column of Table 7.9 identifies the athlete and point in the race at which the time was recorded. The next two columns are the elapsed times up to each point in the race by each athlete according to observers *A* and *B* respectively. The split times (columns 4 and 5) can be determined by subtracting the elapsed time for the previous 100m point in the race from any given 100m point from 200m to 800m. This can easily be accomplished in Microsoft Excel. Split times are used in the reliability study because, a) using elapsed times would conceal errors, b) split times are all 100m times whereas elapsed times use totally different ranges of values at the different 100m points in the race, and c) the analytical goals of the study usually involve split times rather than elapsed times.

The final two columns of Table 7.8 report the two different types of percentage error described in the current chapter. Column 9 is the unadjusted percentage error value used in equation 7.8, which expresses the absolute error as a percentage of the mean of the values recorded by the two observers (column 8 shows the mean values). The final column of the table shows the percentage error produced by equation 7.9, which does not use the mean of the two operators' values. The error is expressed as a percentage of the meaningful range of values; in this case the meaningful range of values is 8s, the difference between a theoretical minimum of 10s and a theoretical maximum of 18s.

We can see that the mean of the percentage errors is 1.26 per cent and 2.32 per cent, but we may wish to summarise the percentage errors in different ways. For example, we may wish to report the maximum percentage errors, which are 10.44 per cent and 19.00 per cent. Alternatively, we may wish to use the 95th percentile of the percentage errors to give a percentage error that we can state that 95 per cent of percentage errors are less than. This can be done with the Percentile function in Microsoft Excel and in the current example the results would be 4.34 per cent and 7.83 per cent.

In computing mean absolute error, the difference between the two operators' recorded split times is first computed (column 6 of Table 7.8) and then

Table 7.8 Calculation of mean percentage error for split time (data from Brown, 2005)

Athlete: Distance	*Elapsed Time A (s)*	*Elapsed Time B (s)*	*Split Time A (s)*	*Split Time B (s)*	*Difference (s)*	*Absolute Difference (s)*	*Mean Value (s)*	*%Error (equ 7.8)*	*%Error (equ 7.9)*
1:100m	13.81	13.76	13.81	13.76	-0.05	0.05	13.785	0.36	0.63
1:200m	27.54	27.56	13.73	13.80	0.07	0.07	13.765	0.51	0.88
1:300m	41.75	41.74	14.21	14.18	-0.03	0.03	14.195	0.21	0.38
1:400m	56.43	56.45	14.68	14.71	0.03	0.03	14.695	0.20	0.38
1:500m	71.84	71.91	15.41	15.46	0.05	0.05	15.435	0.32	0.63
1:600m	86.97	87.10	15.13	15.19	0.06	0.06	15.160	0.40	0.75
1:700m	103.24	103.33	16.27	16.23	-0.04	0.04	16.250	0.25	0.50
1:800m	120.95	120.95	17.71	17.62	-0.09	0.09	17.665	0.51	1.13
				There are a total of 64 rows (8 athletes x 8 splits)					
2:100m	14.42	14.37	14.42	14.37	-0.05	0.05	14.395	0.35	0.63
2:200m	28.46	28.42	14.04	14.05	0.01	0.01	14.045	0.07	0.13
8:600m	87.31	87.89	15.59	15.62	0.03	0.03	15.605	0.19	0.37
8:700m	103.89	103.62	16.58	15.73	-0.85	0.85	16.155	5.26	10.63
8:800m	119.62	119.62	15.73	16.00	0.27	0.27	15.865	1.70	3.37
Mean			14.69	14.69	0.00	0.19	14.690	1.26	2.32
SD			0.78	0.74	0.35	0.29	0.74	1.98	3.64

the magnitude of this is determined with the ABS function in Microsoft Excel (column 7). The mean of the 64 absolute errors is the mean absolute error, which is 0.19s. The use of the magnitude of the error guards against errors cancelling each other out in the calculation of the reliability statistic.

The absolute errors may be reported in the units of measurement (s in this case) without using percentages. This may be appropriate when relating measurement error to the analytical goals of the study. Mean absolute error, the 95th percentile for absolute errors or maximum absolute error could be reported. There may be particular splits that are more reliable than others due to camera positioning or track markings. For example, Brown and O'Donoghue (2007) found that the split times at 300m, 700m, 1,100m and 1,500m points of the 1,500m where the athletes passed the finish line on each lap were more accurate than the 100m, 500m, 900m and 1,300m split times where the athletes were passing the 200m start. Therefore, reliability statistics reported separately for different points in the race could be a more informative way of presenting reliability results. An example of this for the 800m is shown in Table 7.9.

The first seven columns of Table 7.8 can also be used in the calculation of 95 per cent limits of agreement, change of the mean and standard error of measurement. One of the disadvantages of using mean absolute error or a derivative of it is that an overall error statistic is reported without providing any information about the different components of error. Atkinson and Nevill (1998) described how total error is the sum of systematic bias and random error. Systematic bias is concerned with any tendency for one operator to record a higher value than another. For example, in an inter-operator agreement study, one operator may classify a movement as being performed at a high intensity more often than another operator. If we are aware of such a tendency, then this can be addressed when data are gathered by the operators. This could be done by using the difference between the observer means and adding half of this value to any value recorded by the observer who tends to record lower values and subtract half of this value from any value recorded by the observer who tends to record higher values.

Random error is the additional error that may be due to individual perceptual errors, data entry errors or operator fatigue. Bland and Altman's (1986) 95 per cent limits of agreement are computed using the individual errors (column 6 of Table 7.8). There is very little systematic bias here with the mean difference in the split times recorded by the two operators being 0.00s. As we can see in column 6 of Table 7.8, not all errors are the same as the systematic bias value. This is because random error also has an influence on measurement. Random error is computed using the standard deviation of the differences between the two operators' recorded split times. This can be done using the STDEV function in Microsoft Excel, giving 0.346s. The standard deviation only represents 68 per cent of the spread of error values if the errors are normally distributed. Therefore, the standard deviation has

to be multiplied by $z_{1-0.05/2} = 1.96$ to give a spread of 95 per cent of the error values. In this example, the random error is ±0.678s. The 95 per cent limits of agreement are expressed as the systematic bias ± the random error (0.000 ± 0.678s). The Bland Altman plot (Figure 7.4) can be a useful way of illustrating the types of errors associated with different values. However, decisions about the level of reliability can be made from the systematic bias ± the random error. One thing that researchers must remember is that the random error is not a mean error or a typical error value, but the error value that only 5 per cent of errors are expected to exceed.

Hopkins (2000) recommended using change of the mean and standard error of measurement (SEM) to represent systematic bias and random error respectively. The change of the mean is simply the difference between the mean values recorded by the two operators (14.69s - 14.69s = 0.0s). This will always give the same result as the systematic bias in Bland and Altman's (1986) 95 per cent limits of agreement. The measure of random error recommended by Hopkins (2000) is the SEM (sometimes referred to as the 'typical error'). This is computed by dividing the standard deviation of the error values by √2, which covers 52 per cent of the errors if they are normally distributed. In Brown's (2005) data of 100m split times during the 800m, the SEM is 0.247s.

Table 7.9 summarises the absolute reliability of the split times recorded using different reliability statistics. The most reliable split time was 500m to 600m according to mean absolute error and the two percentage error statistics. These three measures do not distinguish between the systematic bias and random error components of total error. The change in the mean and the systematic bias are equivalent. Knowledge of systematic bias allows measurements to be adjusted for known tendencies of particular observers,

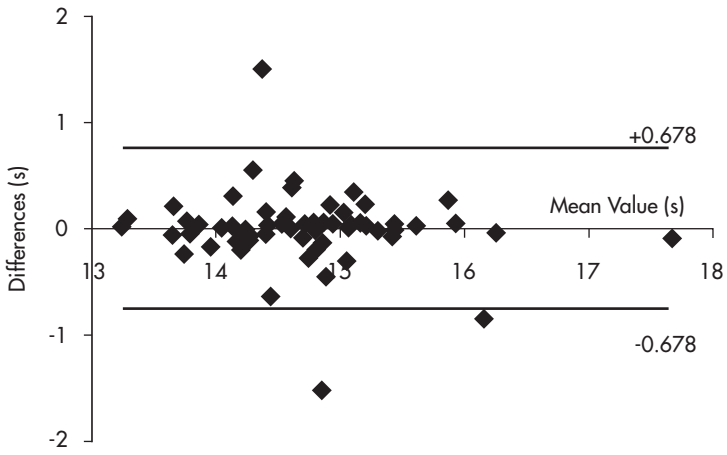

Figure 7.4 Bland Altman Plot (using data from Brown 2005)

Table 7.9 Summary of absolute reliability statistics (s)

Race section	*Mean Absolute error*	*%Error (equ 7.8)*	*%Error (equ 7.9)*	*Change in the mean*	*SEM*	*Systematic Bias*	*Random Error*
0m-100m	0.13	0.93	1.64	0.04	0.16	0.04	0.44
100m-200m	0.11	0.78	1.33	0.00	0.10	0.00	0.28
200m-300m	0.29	2.04	3.66	0.22	0.38	0.22	1.06
300m-400m	0.30	2.06	3.80	-0.27	0.38	-0.27	1.04
400m-500m	0.15	0.99	1.88	0.13	0.09	0.13	0.26
500m-600m	0.10	0.67	1.25	0.06	0.12	0.06	0.33
600m-700m	0.28	1.85	3.52	-0.27	0.23	-0.27	0.63
700m-800m	0.12	0.76	1.47	0.09	0.10	0.09	0.27
All	0.19	1.26	2.32	0.00	0.24	0.00	0.68

meaning that only the random error remains. SEM and random error are both derived from the standard deviation of the inter-operator errors.

The use of the 95 per cent limits of agreement is based on the assumption of homoscedasticity. This means that there should be no association between the magnitude of the errors (column 7 of Table 7.8) and the mean values (column 8). In this case the data is homoscedastic (r = 0.040). Heteroscedasticity is where there is a positive association between the magnitude of the errors and the mean values recorded by the two operators. That is, higher values are associated with higher errors than lower values are; this is sometimes termed a 'shotgun' effect. Where such a 'shotgun' effect occurs, the systematic bias and random error are better expressed using ratios (x/÷) than using differences (+/-) because larger values have larger errors. Readers who are interested in the detail of how 95 per cent ratio limits of agreement are calculated are directed to the worked example by Atkinson and Nevill (1998).

Reliability statistics for interval scale variables

There are different types of interval scale variables in performance analysis that include negative values. Examples from analysis of technique include displacement, velocity, acceleration, angular velocity and angular acceleration. While automatic data capture systems such as Coda (CodaMotion, Rothley, Leicestershire, UK) and Vicon (Vicon, Los Angeles, CA) are used in many kinematic analysis studies, there are other systems with analysis tools that involve human operator input. These systems include Dartfish (Dartfish, Fribourg, Switzerland), siliconcoach (Siliconcoach, Dunedin, New Zealand) and Quintic biomechanics (Quintic Consultancy Ltd, Coventry, UK). Where human operator input is involved in the calculation of angles, distances or any velocity and acceleration derivatives of these, it is essential to demonstrate the reliability of the methods used.

Many of the reliability statistics that can be used with ratio scale variables can also be used with interval scale variables. Change of the mean, standard error of measurement, mean absolute error, root mean square error and 95 per cent limits of agreement can all be used with both interval and ratio scale variables. These techniques are based on differences in values recorded by different operators. Subtraction of values is valid on any interval or ratio scale.

Percentage error, as defined by Hughes *et al.* (2004), is one technique that cannot be used with interval scale variables that include negative values. During the calculation of percentage error, the absolute difference between two operators' recorded values is divided by the mean of the two values. This could involve a division by zero error or a division by a negative number leading to a negative percentage error being computed. Even when total percentage error is used, the sum of the absolute errors is divided by a sum of mean values which may be positive, negative or zero. The adjusted per-

centage error (equation 7.8) can be used to evaluate the reliability of interval scale variables because the absolute errors are always divided by a positive value representing the range of meaningful values. The 95 per cent ratio limits of agreement should not be used with interval scale variables. This reliability statistic involves dividing one operator's recorded value by the corresponding value recorded by the other operator. Such divisions can only be validly applied to ratio scale data.

SUMMARY

Notational analysis has often been portrayed as a purely objective analysis technique. However, this chapter has explained that when a human operator is part of the system used to measure performance variables, there will always be an element of subjective judgement when classifying behaviour. It is, therefore, essential to provide guidance on system use, train system users and determine the level of reliability of the system. Different variables are measured on different measurement scales requiring the use of different reliability statistics.