An investigation of inter-operator reliability tests for real-time analysis system.

Hyongjun Choi, Peter O'Donoghue and Mike Hughes
Centre for Performance Analysis, University of Wales Institute, Cardiff.

Abstract

Sports performance data has been measured and collected by a variety of methods including different measurement tools or systems in the field of performance analysis of sport. Lapse-time analysis (LTA) has often been used rather than real-time analysis (RTA) because of the difficulties of data gathering, presentation of the results, non-feasibility of feedback during match time and the selection of valid performance indicators (PIs). RTA data has often been used to enhance performances and, therefore, it is necessary to investigate the reliability tests used on such data.

Pearson's r, Chi-square, % error and Kappa tests statistics were used to evaluate the reliability between independent observers. A peer review process of modelling different levels of reliability with synthetic data demonstrated that kappa was the only one of the four statistics to have construct validity for the purpose of reliability assessment. With the Cybersport system for basketball, kappa values for the system as a whole of 0.8 or above are interpreted as good while values of 0.6 to 0.8 are acceptable. When using kappa for post hoc reliability analysis of individual event types, values of 0.5 or above can be considered as acceptable.

Keywords: reliability, real-time data gathering, real-time analysis, construct validity.

1. Introduction.

Performance analysis techniques have been used to enhance performances in sport in recent years. In particular, the data gathered by methods such as hand-notation and computerised-notation systems has to be tested for reliability in a systematic manner. Objective information has been used in the field of performance analysis of sport. The

methods of gathering performance data have been developed and improved efficiently and accurately (Hughes and Franks, 2004). When analysing sports performance data, it is essential to understand the level of inter- and intra-operator agreement. Therefore reliability tests have often been issued in the field of performance analysis of sport. Hughes et al. (2004) have researched previous papers in order to identify the type of reliability tests used in the field of performance analysis finding that 70 % of the papers have not used reliability tests and 5 different tests such as correlation, method error (%), Chi-square, t-test and Cronbach's alpha have been used in the past. Additionally, the method error (%) could be used in the performance analysis with different meanings (Hughes et al., 2004). Recently, Bloomfield et al. (2006) have investigated inter-operator reliability of a computerised work-rate analysis system using kappa. Kappa has also been used to test the reliability of the CAPTAIN (McLaughlin and O'Donoghue, 2001) and POWER (O'Donoghue et al., 2005) time-motion analysis systems. An algorithm has been developed to use kappa with continuous time data rather than with discrete events (O'Donoghue, 2005).

As the number of issues relating to reliability in the field of performance analysis of sport increases, the rationale for the methods used to determine the reliability of performance analysis systems such as real-time and lapse-time systems (Choi et al., 2006) have been a concern. A further concern is that the values of reliability statistics required to demonstrate an acceptable level of reliability are uncertain. While Altman (1991) provides threshold values for fair, moderate, good and very good strength of agreement, these may not be suitable for performance analysis of sport. Therefore, the purpose of the current paper is to compare reliability statistics and the values they produce. One aim of this study is to determine the inter-operator reliability of real-time data capture. A second aim was to use synthetic data to deliberately determine reliability statistics for good, acceptable and poor levels of reliability.

2. Methods.

This project compared observations of basketball events made by 4 operators (T1, T2, T3 and T4) using kappa, chi square, Pearson's r and %error. The observers were chosen from research students of the Sports Recording and Analysis Centre in South Korea who had no previous experience of the computerised data gathering system for basketball, but with a basic knowledge of basketball. The equipment used in the project included a TV screen, VHS video-player and 4 laptop computers which executed

the CyberSports for basketball software. Additionally, 10 minutes of a game in the Korean professional basketball league was randomly selected and recorded onto VHS video tape. There were 2 training sessions and 1 experimental session using the CyberSports for basketball software where different video footage was shown. The training sessions were designed for the observers to understand how to tag the data and how to correct the data during the observations. All 4 operators underwent the same training programme involving system demonstrations as well as hands on experience operating the system while observing video recordings of basketball competition. Five day breaks between the training sessions were required in order to reduce memory factors that might distort reliability statistics. Additionally, the observers could discuss any questions or problems during training with the trainer. For the experimental session, however, no communication was allowed in order not to influence to the observations. The following performance indicators (PIs) were observed during the 10 minute basketball performance.

- 3 point attempt
- 3 point made
- Assist
- Defensive Rebound
- Field Goal attempt
- Field Goal made
- Free throw attempt
- Free throw made
- Offensive Rebound
- Personal fouls
- Steals
- Turnover

Although the post hoc chi square calculations were made using 2 x 2 cross tabulations, it was decided not to use Fisher's Exact test. Fisher's Exact test gives higher P values making it harder to achieve a significant difference and therefore giving greater confidence in any significant differences found. When used for reliability, higher P values indicate greater reliability and, therefore, using Fisher's Exact test risks reporting higher reliability than is actually present in the methods.

In addition to inspecting inter-operator agreement, a peer review process of generating errors was undertaken to deliberately create synthetic observations that would be

considered to have total, good, acceptable and poor agreement. This involved starting with a single set of observed results which were assumed to be correct and introducing impurities that would represent different levels of error. A basketball coach and the authors were involved in this process and came to an agreed set of observations representing different levels of error with the original observation. The disparity between the agreements was decided subjectively, but the values in the each case were revised with acceptable reasons for errors. The total frequencies varied between the four synthetic cases as it was assumed unacceptable agreement would be partly characterised by some events not being recorded by either operator. The severity of disagreements between different events was considered in terms of basketball performance data interpretation. Tables 1 to 4 show different levels of agreement based on the peer review process of introducing errors to represent different levels of agreement.

Table 1. Cross-tabulation of synthetic operators with total agreement.

First Operator	Second Operator													
	N/A	3pt Attempts	3pt made	Assist	Def. Rebound	FG attempt	FG made	FT attempt	FT made	Off. Rebound	Fouls	Steal	Turnover	Sum
N/A														0
3pt Attempts		4												4
3pt made			4											4
Assist				15										15
Def. Rebound					15									15
FG attempt						12								12
FG made							16							16
FT attempt								1						1
FT made									3					3
Off. Rebound										2				2
Fouls											4			4
Steal												1		1
Turnover													3	3
Sum	0	4	4	15	15	12	16	1	3	2	4	1	3	80

Table 2. Cross-tabulation of synthetic operators with good agreement.

First Operator	Second Operator													
	N/A	3pt Attempts	3pt made	Assist	Def. Rebound	FG attempt	FG made	FT attempt	FT made	Off. Rebound	Fouls	Steal	Turnover	Sum
N/A				6		1								7
3pt Attempts		4												4
3pt made			4											4
Assist	1			7										8
Def. Rebound					15					1				16
FG attempt	3					8								11
FG made							16							16
FT attempt								1						1
FT made									3					3
Off. Rebound										1				1
Fouls											4			4
Steal												1		1
Turnover	1												2	3
Sum	5	4	4	13	15	9	16	1	3	2	4	1	2	79

Table 3. Cross-tabulation of synthetic operators with acceptable agreement.

First Operator	Second Operator													
	N/A	3pt Attempts	3pt made	Assist	Def. Rebound	FG attempt	FG made	FT attempt	FT made	Off. Rebound	Fouls	Steal	Turnover	Sum
N/A				5	5	1					1			12
3pt Attempts		4				1								5
3pt made	1		3											4
Assist	3			5										8
Def. Rebound	3				10					1				14
FG attempt	2	1				7								10
FG made	1						15							16
FT attempt								1						1
FT made									3					3
Off. Rebound	1									1				2
Fouls											3			3
Steal												1	1	2
Turnover	1												1	2
Sum	12	5	3	10	15	9	15	1	3	2	4	1	2	82

Table 4. Cross-tabulation of synthetic operators with unacceptable agreement.

First Operator	Second Operator													
	N/A	3pt Attempts	3pt made	Assist	Def. Rebound	FG attempt	FG made	FT attempt	FT made	Off. Rebound	Fouls	Steal	Turnover	Sum
N/A			1	6	2	1					1			11
3pt Attempts		3				2								5
3pt made	1		2											3
Assist	4			1										5
Def. Rebound	4				7					1				12
FG attempt	3	1				5								9
FG made						1	15							16
FT attempt								1						1
FT made									3					3
Off. Rebound	1													1
Fouls	1										1			2
Steal												1	1	2
Turnover	1												1	2
Sum	15	4	3	7	9	9	15	1	3	1	2	1	2	72

3. Results.

Table 5 shows the inter-operator reliability statistics for each 6 pairs of observers within the set of 4 observers. It is evident that the relative rating of each pair of observers depends on the reliability statistic being used. All of the statistics find T1 v T3 to be the pair of observations with the greatest agreement, though Pearson's r gives T2 v T4 an equivalent reliability value. However, T1 v T4 would be deemed the pair of observers with the least agreement according to kappa while the other statistics find T3 v T4 to be the pair with the lowest level of agreement.

Table 5. The comparison of inter-operator reliabilities.

	Pearson's r	χ^2 (p)	Total % Error	Kappa
T1 vs. T2	0.84	6.87 (0.809)	24.0 %	0.75
T1 vs. T3	0.97	1.78 (0.999)	14.8 %	0.77
T1 vs. T4	0.79	9.18 (0.605)	31.4 %	0.59
T2 vs. T3	0.76	8.42 (0.675)	28.4 %	0.65
T2 vs. T4	0.97	5.02 (0.927)	20.2 %	0.67
T3 vs. T4	0.63	12.55 (0.324)	33.8 %	0.68

Table 6 shows the results of the peer review of different types of error levels. Pearson's r, Chi square and total percentage error give a greater level of reliability for the two synthetic observations of acceptable reliability than between the two synthetic observations of good reliability. Kappa, on the other hand, produced values that reflect the perceived order of agreement. Chi square is a particularly poor reliability statistic reporting excellent reliability ($P > 0.990$) in each case, even when two observations were deliberately created to synthesise unacceptable error between the observers.

Table 6. The summary of the peer reviews on different agreements.

	Pearson's r	χ^2 (p)	Total % Error	Kappa
Total Agreement	1.00	0.00 (1.000)	0.00 %	1.00
Good	0.95	1.96 (0.999)	13.7 %	0.81
Acceptable	0.98	0.96 (>0.999)	11.4 %	0.61
Not Acceptable	0.97	1.24 (>0.999)	13.5 %	0.50

Table 7 shows the chi-square and kappa values along with the percentage error values for observers 1 and 3. This is an example of the use of reliability statistics to undertake post hoc analysis of the reliability of individual event types. Where frequencies are very low, percentage error values of over 100% are possible. The percentage errors of 66.7% in Table 8 are for events recorded once by one observer and twice by the other observer. The overall reliability statistic would be expected to be between the lowest

and highest values reported for any event. This is not the case for chi square where the P value of 0.999 for the agreement between the two frequency profiles exceeds that for all individual event types. The percentage error for personal foul of 22.2% is particularly harsh when one considers that there was only 1 occasion where personal foul was confused with another event type. Kappa, on the other hand indicates good reliability for this event type.

Table 7. Pairwise comparisons between observers 1 and 3.

Event	%Error	χ^2_1, P	Kappa
3Pt Attempt	0.0	0.0 (0.971)	1.00
3Pt Made	0.0	0.0 (0.971)	1.00
Assist	28.6	0.8 (0.383)	0.75
Def Rebound	6.9	0.1 (0.781)	0.88
FG Attempt	8.7	0.1 (0.773)	0.85
FG Made	6.5	0.1 (0.782)	0.96
FT Attempt	0.0	0.0 (0.986)	1.00
FT Made	0.0	0.0 (0.975)	1.00
Off Rebound	66.7	0.4 (0.546)	-0.02
Personal Foul	22.2	0.1 (0.760)	0.88
Steal	66.7	0.3 (0.575)	0.66
Turnover	40.0	0.2 (0.630)	0.79
Total	14.8	1.8 (0.999) ^	0.77

^ Chi square for all events was calculated with 11 degrees of freedom.

Table 8 shows the pairwise comparisons for the two synthetic observations with acceptable agreement. This shows that some individual event types have kappa values of around 0.5 which would be interpreted as a poor strength of agreement if for all events. Therefore, a lower kappa value can be tolerated for individual event types where the overall observation is deemed to be acceptable.

Table 8. Pairwise comparisons between synthetic observers with acceptable agreement.

Event	%Error	χ^2_1, P	Kappa
3Pt Attempt	0.0	0.0 (1.000)	0.79
3Pt Made	28.6	0.2 (0.698)	0.85
Assist	22.2	0.3 (0.614)	0.51
Def Rebound	6.9	0.0 (0.835)	0.63
FG Attempt	10.5	0.0 (0.805)	0.71
FG Made	6.5	0.0 (0.839)	0.96
FT Attempt	0.0	0.0 (1.000)	1.00
FT Made	0.0	0.0 (1.000)	1.00
Off Rebound	0.0	0.0 (1.000)	0.49
Personal Foul	28.6	0.2 (0.698)	0.85
Steal	66.7	0.3 (0.559)	0.66
Turnover	0.0	0.0 (1.0000	0.49
Total	11.3	1.0 (>0.999) ^	0.61

^ Chi square was calculated with 11 degrees of freedom.

4. Discussion.

This study has examined the construct validity of reliability statistics for a real-time basketball event recording system by synthesising pairs of observations of known differences in level of agreement. This has revealed that kappa is the only statistic with construct validity for this type of data. The explanation of the poor reliability of chi square and Pearson's r comes from the calculation of their values. Consider 2 sets of numbers, one set containing values that are exactly double those of the other set. Pearson's r will be 1.0 as a straight line of gradient 2 can be drawn through the co-ordinates when plotted on a scatter graph. Chi square compares the proportion of each value within the two frequency distributions. These proportions will be identical and a chi square value of 0.0 and associated P value of 1.000 will be determined. Percentage error values of 13.5% and 13.7% were determined for pairs of synthetic observations created to exhibit unacceptable and good reliability respectively. This is evidence that the percentage error reliability statistic does not have construct validity for the type of

sports performance data used in the current investigation.

The interpretation of strength of agreement when using kappa in medical applications is "very good" if above 0.8, "good" if above 0.6, "moderate" if above 0.4, "fair" if above and "poor" if under 0.2 (Altman, 1991). It is possible to compute a kappa value of less than 0.0 if there is less agreement than would be expected by chance. Indeed the post hoc kappa values determined in the current investigation included one of less than 0.0. O'Donoghue (2005) extended Altman's (1991) interpretation scheme by classifying values of less than 0.0 as representing a "very poor" strength of agreement. The current investigation has revealed the need for a different means of interpretation that is different to that proposed by Altman (1991). The kappa values for the basketball system were about 0.8 rather than 0.6 where a good strength of agreement was synthesised and 0.6 to 0.8 when an acceptable level of agreement was synthesised. Values under 0.6 for the overall observations being compared would not be considered as acceptable. It is recommended that these values are used rather than Altman's (1991) interpretation method when dealing with performance analysis systems such as the basketball system used in the current investigation. Due to kappa values for individual event types being above and below the overall kappa value for the observations, values of under 0.6 for individual event types should not necessarily be considered as unacceptable. The synthetic observations deliberately created to represent acceptable reliability introduced errors up to the point where the disagreements would be considered to be of marginal acceptability. The lowest individual event kappa value was 0.49 suggesting that 0.5 might be a threshold value to use for acceptable reliability of individual event types.

The approach used in the current investigation of deliberately creating pairs of observations exhibiting good, acceptable and unacceptable levels of agreement has provided valuable information about the reliability statistics to be used with the CyberSports system for basketball. This information includes the construct validity of alternative reliability statistics, allowing the most suitable statistical technique to be selected, and the threshold values associated with different levels of reliability. These values can then be used in future reliability studies when the system is being operated by different users. It is recommended that this approach is used with other performance analysis systems to determine the reliability statistic with the greatest construct validity and the values of that statistic that reflect acceptable and good levels of agreement.

Kappa is a promising reliability statistic for use with nominal scale variables in performance analysis. Such nominal variables include event types, teams, players and outcomes of events. More research is needed with other sets of performance indicators from other sports to determine whether the kappa threshold values for acceptable and good agreement suggested by the current investigation apply more generally or not.

5. Conclusion.

The percentage error, chi-square and Pearson's r statistics do not have construct validity for reliability analysis of the Cybersports basketball system. The kappa statistic has exhibited construct validity and a means of interpretation has been specified. A further contribution of this research is that a post hoc analysis process using kappa has been proposed for individual event types. The peer review approach of introducing errors into synthetic data to model different levels of reliability is a promising method of selecting reliability statistics and determining a means of interpretation for performance analysis systems in general.

6. References.

Altman, D.G. (1991). **Practical Statistics for Medical Research**. London: Chapman & Hall, 404.

Bloomfield, J., Polman, R., & O'Donoghue, P. (2006). Reliability of the Bloomfield Movement Classification. In **Proceedings of the World Congress of Performance Analysis of Sport VII** (Edited by Dancs, H., Hughes, M. and O'Donoghue, P.G.), pp. 195-202. Cardiff: CPA Press, UWIC.

Choi, H. J., O'Donoghue, P., & Hughes, M. (2006). A Study of team performance indicators by separated time scale using a real-time analysis techniques within English national basketball league. In **Proceedings of the World Congress of Performance Analysis of Sport VII** (Edited by Dancs, H., Hughes, M. and O'Donoghue, P.G.), pp. 124-127. Cardiff: CPA Press, UWIC.

Hughes, M., Cooper, S. M., & Nevill, A. (2004). Analysis of notation data: reliability. In **Notational Analysis of Sport: Second Edition** (Edited by Hughes, M. &

Franks, I.M.), pp. 189-204. London: Routledge.

McLaughlin, E. & O'Donoghue, P. (2001). The reliability of time-motion analysis using the CAPTAIN system. In M. Hughes & I. M. Franks (Eds.), *PASS.COM; performance analysis, sport science and computers* (pp. 63-68). Cardiff: CPA, University of Wales Institute, Cardiff.

O'Donoghue, P. (2005). An Algorithm to use the kappa statistic to establish reliability of computerised time-motion analysis systems. In 5th International Symposium of Computer Science in Sport, Book of Abstracts (pp. 49). Hvar, Croatia.

O'Donoghue, P.G., Hughes, M.G., Rudkin, S., Bloomfield, J., Cairns, G., Powell, S. (2005), Work rate analysis using the POWER (Periods of Work Efforts and Recoveries) System, **International Journal of Performance Analysis of Sport (e)**, 5(1), 5-21.