Journal of Quantitative Analysis in Sports

Volume 6, *Issue* 3 2010 *Article* 3

An Examination of Judging Consistency in a Combat Sport

Tony Myers, *Newman University College*
Alan M. Nevill, *University of Wolverhampton*
Yahya Al-Nakeeb, *Newman University College*

Recommended Citation:
Tony Myers, Alan M. Nevill, and Yahya Al-Nakeeb (2010) "An Examination of Judging Consistency in a Combat Sport," *Journal of Quantitative Analysis in Sports*: Vol. 6 : Iss. 3, Article 3.
Available at: http://www.bepress.com/jqas/vol6/iss3/3
DOI: 10.2202/1559-0410.1178

An Examination of Judging Consistency in a Combat Sport

Tony Myers, Alan M. Nevill, and Yahya Al-Nakeeb

Abstract

Two related studies compared the consistency of two different methods of interpreting and applying scoring criteria in Muay Thai that are normally used by officials in the UK and that are used by officials in Thailand. In the first study, levels of consistency were determined by comparing judge's scores (n=270) from forty-five bouts judged by UK officials and forty-five judged by Thai officials. In the second study the original forty-five bouts judged by UK judges were compared with forty–five bouts judged by UK officials using Thai judging criteria. Consistency was examined in both studies using two methods. The first method compared differences in the range of the highest vs. lowest points awarded by judges for each bout. The second method compared homogeneity of variance between judges' scores. Results suggested that the Thai officials were more consistent than their UK trained counterparts but also that UK judges were more consistent when adopting the Thai judging criteria. It was suggested that the use of very clearly defined criteria and concrete operationalization of otherwise subjective concepts used in applying the system used in Thailand was the main reason for the findings.

KEYWORDS: Muay Thai, sport, boxing, judges, interrater agreement

The outcomes of sporting competitions are determined either by an objective measurement, an objective score or what is often considered a subjective judgement (Plessner & Haar, 2006). A considerable number of sports involve judges applying some type of performance rating (Stefani, 1998). Judges are trained to observe performances and apply specific judgement criteria to make value judgements. Judgement decisions are found in sports as diverse as synchronised swimming, gymnastics, snowboarding and boxing. Although often referred to as subjective, these judgements are really intersubjective as they do not depend on the purely idiosyncratic perspective of an individual judge but rather on the possibility of a consensus of opinion by a group of trained individuals (Annett, 2002; Manns, 1998; Muckler & Seven, 1992). To be able to provide a score with any degree of consistency, judgement criteria need to allow a level of intersubjectivity. With some exceptions (e.g. Rainey, & Larsen, 1988; Rainey et al., 1993; Duncan et al., 2005), limited attention has been given to differences in the application of judgement criteria and the impact this may have on outcomes.

While there is currently no data indicating the levels of consistency that are acceptable across sports, consistency in the application of specific criteria have been found to be low (Duncan et al. 2005), as have levels of inter-judge agreement (Weekly & Gier, 1989). An often cited example of the inconsistent application of judging criteria from professional boxing, is a world championship bout held on March 13th, 1999, between British boxer Lennox Lewis and American boxer Evander Holyfield at New York's Madison Square Garden. This was an important bout with significant media attention as it was the first unified heavyweight championship in over 5 years. The bout was declared a draw but resulted in strikingly different scores from the different judges. South African judge Stanley Christodoulu gave the bout to Lewis which reflected the opinion of a number of observers, British judge Larry O'Connell scored the bout as a draw, and most controversially, judge Eugenia Williams of New Jersey declared Holyfield the clear winner seven rounds to five. The result of this bout, along with other similar decisions, led the National Association of Attorneys General Boxing Task Force (NAAG, 2000) to suggest changes to the scoring system from the current '10-point must system' (described in Lee, Cork, & Algranati, 2002) to a 'consensus scoring system', where the median score of the three judges for each round is adopted (NAAG, 2000). Although, as Balmer, Nevill and Lane (2005) pointed out this recommendation only offers the possibility of controlling for a single biased or poor judge and not for other possible influences on scoring.

The Lewis Holyfield bout resulted in a tremendous outcry and charges of corruption and incompetence were levelled at New Jersey judge Williams. While it is certainly possible that a judge may be deliberately dishonest in their judgement of a bout, or any other subjectively judged activity, it is also possible that there are other explanations for differences in such decisions. Besides the

several major judging biases that have been established empirically in subjectively judged sports (Vanden Auweele, Boen, De Geest, & Feys, 2004) there is also the possibility that the level of subjectivity in the scoring criteria itself may play a role in the very different decisions given by judges. Lee, Cork and Algranati (2002) suggested discrepancies in boxing scoring may arise through the lack of standardization in applying the scoring criteria. For example, in international boxing judges consider four criteria for determining which boxer wins any particular round. The first criterion is the number of clean punches landed by each boxer (Kaczmarek, 1996). Clean punches are considered to be those thrown with a clenched fist, strike with the knuckle part of the glove and land above the waist on the front and side of the torso or the front or side of the head (Lee, Cork, & Algranati, 2002). While it seems a straight forward matter of judges identifying the ratio of successful punches each boxer delivers numerically and awarding the round to the boxer who has more hits on target, in practice a judge has other factors to consider. Not only can it be difficult for a boxing judge to determine which punches land cleanly and which do not, they must also attempt to determine the quality or power alongside the quantity of punches delivered by the boxer (Kaczmarek, 1996). Determining the relative power of punches without clear easily observable criteria is very subjective and may lead to differences in interpretation by different judges.

One possible influence on application of criteria is the application of different sets of normative rules. Normative rules as a set of standards different to the official rules of a sport (Silva, 1981). Normative rules have been found to impact on decisions in sports (Rainey, & Larsen, 1988; Rainey, et al. 1993). These unwritten rules or rule applications exist alongside the official rules (Plessner, 2005). One reason posited for this is that it is not always possible to articulate the exact nature of what needs to be judged in words, instead common understanding is demonstrated rather than described (Mumford, 2006). As such, sports officials tend to use unwritten conventions and it is these that often determine how official rules are actually applied in particular circumstances (D'Agostino, 1995). Groups of officials form what can be described as 'communities of practice' which are essentially formal or informal groups that generate and share practice, ideas and commitments while producing various mediums to carry this accumulated knowledge (Wenger, 1999). Where there is limited contact between these communities, local communities of officials and other stakeholders may arguably develop practices that vary from the other communities that they have little or no contact with. This can be the case in applying and interpreting particular rules. In sports that have regional, national and international governing bodies, such differences are likely to be held in check. However, in sports that have less formal organisation, language barriers or where communities are isolated geographically, differences in normative rules may exist.

DOI: 10.2202/1559-0410.1178

One sport in which appears to be an ideal vehicle for exploring differences in the unwritten application of judgement criteria is Muay Thai, the national sport of Thailand. Muay Thai is growing in global popularity with an estimated one million participants worldwide coming from five continents (Gartland, Malik & Lovell, 2004). The sport has two groups of experienced officials who use different normative rule applications of the written judging criteria (Myers, 2007). Muay Thai is a ring sport that allows competitors to use kicks, knees, elbow strikes and certain types of throws to defeat an opponent. The sport is judged in a similar way to international style boxing with competitors trying to win by influencing three ringside judges (in professional bouts) to award them more points than their opponent using a '10-point must system' (Myers, 2000). As in international boxing, competitors are also able to finish a bout prematurely by influencing the referee overseeing the bout and award them victory by knocking out or disabling their opponent (World MuayThai Council, 1995). Similar to the judging criteria used in international boxing, judging Muay Thai bouts include both quantifiable aspects along with more subjective elements. For Example, to award a particular round to a competitor the rules suggest judges should look for "The boxer with more heavy, powerful, and clear attacks" and "… who shows better offensive skills, defensive skills, elusive skills, or counterattacking skills" (Boxing Board of Sport, 2002, p.22). So similar to international boxing, judges have to decide which competitor strikes successfully on target with the highest number of techniques, while also considering the relative power of blows and making a judgement on the quality of each boxer's offensive and defensive skills. These criteria allow for quite different interpretations. For example, the concept of "effective" strikes could be interpreted in many different ways depending on how a particular judge determines effectiveness. The reason for differences in the interpretation and application of these different criteria between groups may be due to geographical isolation, a lack of a unifying single world body or may be cultural (Myers, 2007).

There has been a hotly contested debate in Muay Thai circles over which judging approach should be used in international competition: the normative rule applications used by Thai officials or those used traditionally by UK and other western officials. Those advocating the approach used traditionally in the UK suggest it is a more easily understood in the west, being similar to international style boxing, and is a more objective system because all technique score equally. Those that advocate the system used by Thai judges argue that particular techniques are innately more effective than others and this should be reflected in scoring. They also argue that the winner of a bout should be the competitor who won more of the whole match and not just be ahead in more segments of that match (Myers, 2007).

One of the strongest arguments put forward by those by those in favour of adopting the scoring practice used in Thailand for international competition, is the perceived level of inter-judge agreement in judgement decisions. This is seen as being important, offering consistency and direction to competitors. Anecdotal evidence suggests that the system used by Thai trained officials is more consistent than the system used by western trained officials. However, this suggested superior consistency is purely anecdotal and does not have the support from any published studies. Importantly, consistency may also suggest that judges have expert status (Einhorn, 1972, 1974) or, far less positively, are being influenced by a conformity effect (e.g. Sheer et al., 1983; Lee, 2007; Wander, 1987).

The aim of the two studies is to examine the consistency of two different applications of scoring criteria; Thai judging criteria and UK judging criteria. The first study will compare the consistency of Thai judges directly with UK judges. The second study will compare UK judges using Thai scoring criteria with the scores obtained in the first study where UK judges applied the criteria traditionally used in the UK. Given the anecdotal evidence, the study hypothesises that the Thai judging system will result in the more consistent scoring of bouts when compared with the system traditionally used in the UK.

Study 1

Method

Data

Judge's scores (n=270) were collected from forty-five Muay Thai bouts, judged in the UK by UK trained officials and forty-five judged in Thailand by Thai officials (i.e., with three judges per bout). UK judges' scores were collected over the period a year from a number of Muay Thai events across the UK. Thai judges scores were randomly selected from the same time period from results published in 'Muay Siam magazine', a weekly Thai newspaper that publishes results of Muay Thai bouts.

Analysis

Two different methods were used to calculate consistency. In the first method judge's scores for each competitor were entered into a spreadsheet in bout order (judge's scores for the red and blue corner in each bout). Differences in an individual judge's scores for each competitor were first computed. Then differences in the range of the highest versus lowest points awarded by judges for

DOI: 10.2202/1559-0410.1178

that bout were calculated (see figure 1.), with the means and standard deviations of this range determined; these being used to compare the consistency of the two groups.

Thai judges					
bout	judge	red	blue	diff between scores	diff range of the highest versus lowest points
1	1	50	47	3	
1	2	50	47	3	
1	3	50	46	4	1
2	1	49	47	2	
2	2	49	47	2	
2	3	49	47	2	0
3	1	49	48	1	
3	2	48	49	-1	
3	3	48	48	0	2
4	1	50	47	3	
4	2	50	47	3	
4	3	50	47	3	0

Figure 1. Example of the spreadsheet used to calculate calculating differences in the range of the highest versus lowest points awarded by judges for each bout

In the second method, the individual bouts were labelled as ninety unique bouts. Differences were calculated between the judge's scores for the boxers in the red and blue corners. Descriptive statistics were calculated to examine the differences in the points awarded by judges to boxers competing from red and blue corners. Given that each of the 90 bouts were likely to have a difference in points awarded depending on the quality of the boxers competing (and the corner they competed from red or blue) the residuals from these differences were calculated from a One-way ANOVA. These residuals were calculated to determine the variation in points that could not be explained by the quality of the boxers and colour of the corner. To determine the relative constancy of the UK and Thai judges, these residuals were used as a new dependant variable and a test for equality of variance was used to compare homogeneity of variance between Thai and UK judges. An F test and Levene's test were used to determine if any differences in homogeneity were statistically significant.

Results

Descriptive statistics calculated on all bouts suggested that, overall boxers competing out of the red corner (M = 43.28 ± 9.02) were awarded more points than boxers competing out of the blue corner (M=42.39 ± 8.65). This suggests over all the bouts, red corner boxers were better on average than boxers competing out of the blue corner being awarded 0.88 more points per bout on average.

Across all the forty-five matches, the Thai judges scores differed by a maximum of two points in any one match compared to a maximum of eight points difference in UK judges. The average range of the differences calculated using our first method of determining consistency suggested differences were lower in Thai officials (see table 1.). These differences were statistically significant ($t(55.29) = 8.458$, $P < .0001$) and the differences represent a huge effect (Cohen's $d = 1.79$; effect size $r = 0.67$) judges only disagreed on the outcome of four of the forty-five matches compared with eleven by UK judges.

Table 1. The average range (maximum- minimum score awarded by the three judges in each bout) and standard deviations and total point differences across all matches between Thai and UK judges

Judges	Mean	SD	Total differences
UK trained	2.38	1.54	107
Thai trained	0.3	0.56	14

The second method also suggested a higher consistency in the Thai judges. The One-way ANOVA identified that the difference in the quality of boxer from the red and blue corners varied significantly between the 90 bouts ($F(89, 180) = 3.45$, $P<.0001$). The residuals of the ANOVA reflect the variation or inconsistency between the three judges that could not be explained by the relative quality of the boxers. The standard deviations calculated from the Test for Equality of Variance, using the residuals as the dependant variable, suggested large variation between Thai trained (SD=.28) and UK trained (SD=1.2) judges. Thai trained judges were far more consistent in their decisions. This difference in variances was statistically significant ($F= 18.56$, $P<.0001$; Levene's Test statistic $= 125.22$, $P< .0001$). Confidence intervals for the difference in standard deviations between Thai and UK judges are shown below in figure 2.

DOI: 10.2202/1559-0410.1178

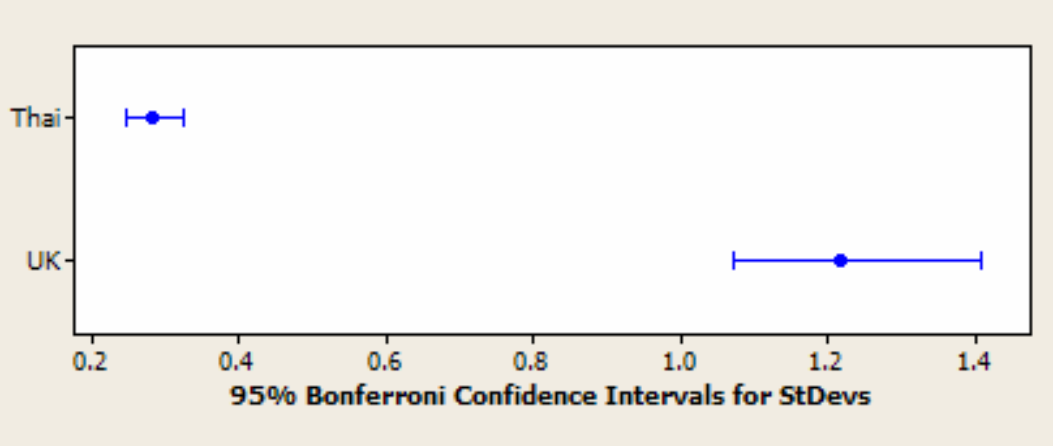

Figure 2. Confidence intervals for the difference in standard deviations between Thai and UK judges

Discussion

The results support the anecdotal evidence of the high level of consistency achieved by Thai judges. Both of the methods used to determine consistency suggested that Thai judges had a far greater agreement on the points awarded to particular boxers in bouts than did their UK counterparts. There was four times more variation in the UK judges' unexplained residual variation compared to that of Thai trained judges, suggesting that Thai trained judges were far more consistent in their decisions.

Such high levels of consistency can be seen as highly desirable sign of expertise if associated with performance factors. Consistency is one of the two key criteria that can be used to identify experts in a number of domains (Shanteau et al., 2003). Although Shanteau et al.,(2003) argued for intra-expert consistency rather than inter-expert consistency, Einhorn (1972, 1974) proposed that agreement between experts is a necessary condition for expert status. Certainly it can be argued, that without consistency in a judged sport, winning and losing becomes something of a 'lottery' where athletes and coaches have to try to guess what different judges may be looking for when determining an outcome. When judges are consistent in their application of judgement criteria competitors and coaches can be confident in their training and strategy decisions. The level of consistency demonstrated by Thai trained officials may suggest that the Thai judges were more expert than their UK counterparts.

Conversely, high consistency can be undesirable if it is the result of non-performance related factors. Consensus as a result of non-performance factors is as undesirable as any other non-performance bias. Conformity effects have been

identified by Scheer et al. (1983) in gymnastics, Wanderer (1987) and Lee (2007) in figure skating, and Vanden Auweele, Boen, De Geest, and Feys (2004) in synchronised swimming. Given that the Thai officials in this study regularly worked together, often on a weekly basis, it could be argued that the observed consistency may have been influenced by a nonperformance-based conformity effect. Vanden Auweele, Boen, De Geest, and Feys (2004) found synchronised swimming judges conformed to the group norm even when information about this group norm was no longer available suggesting that the conformity effect caused by information on other judge's scores is pervasive. In Muay Thai, at the end of each bout judge's scores are announced to audience, this regular feedback may be responsible in part for the observed consistency levels in Thai judges. However, UK judges often examine each other's score cards so obtain similar information to their Thai counterparts. Another non-performance factor that may have contributed to conformity is crowd noise (Nevill, Balmer, & Williams, 2002; Balmer et al, 2007). In Muay Thai stadiums in Thailand, gamblers offer very vocal support for their favourite contestant throughout a bout. This may be another non-performance factor that polarised judge's decisions. However, it can also be argued that similar factors were also present during the bouts judged in the UK by UK officials.

One methodological limitation of this first study that may have contributed to the findings was the difference in experience and familiarity between groups of officials used. The Thai officials were highly experienced and regularly worked together as frequently as on a weekly basis in some cases. In contrast, the UK officials worked more infrequently together and were less experienced. The secondly limitation of the present study was that it is conceivable that the bouts judged by the Thai judges may have been more "clear cut" that the fights judges by the UK officials. The Thai competitors were more experienced on the whole and often technically superior to their UK counterparts. Although the skill level is generally higher in the bouts judged in Thailand, it can be argued that they were actually not as clear cut as matchmakers organising the bouts like to please gamblers and have close bouts. They even go as far as handicapping the better competitor to make fights more even and the outcome less certain (Myers, 2000).

These limitations are addressed in the second study by directly comparing judges with similar levels of experience and on the same type of bouts held across the UK. This will be done by comparing the original scores obtained from UK judges applying their commonly used criteria with UK judges of similar experience using the Thai judging system.

DOI: 10.2202/1559-0410.1178

Study 2

Method

Data

To maintain consistency, the original UK judge's scores (n=270) used in the first study were used again in this second study. This included the forty-five bouts judged in the UK by UK officials. The second data set involved the scores of forty-five MuayThai bouts judged in UK by UK nationals, but using the scoring criteria applied by Thai officials (again with three judges per bout). All the UK nationals involved in this second data set had received training over a period of time and passed an assessment in applying Thai scoring criteria. Again, the score cards of the UK nationals using Thai scoring criteria were collected from a number of Muay Thai events across the UK.

Analysis

The same two methods of analysis used in study one were applied in this second study. The first method involved calculating differences in the range of the highest versus lowest points awarded by judges for each bout, with the means and standard deviations of this range calculated; these being used to calculate consistency of the two groups.

As in the first study, the second method employed in this study involved the individual bouts again labelled as ninety unique bouts with differences calculated between the judge's scores for the competitors from the red and blue corners. Descriptive statistics were again calculated for the differences in the points awarded by judges. Once more, given that each of the 90 bouts were likely to have a difference in points awarded depending on the quality of the boxers competing, the residuals of the differences were calculated from a One-way ANOVA to account for this. To determine the relative constancy of the judges using Thai criteria and using the criteria traditionally used in the UK, the residuals were used as a new dependant variable with a test for equality of variance used to compare homogeneity of variance between the two groups of judges. An F test and Levene's test were used to determine if any differences in homogeneity of variance were statistically significant.

Results

Descriptive statistics calculated on all bouts suggested that, overall boxers competing out of the red corner (M = 46.79 ± 5.77) were awarded more points than boxers competing out of the blue corner (M=46.09 ± 5.75). This suggests over all the bouts, red corner boxers were better on average than boxers competing out of the blue corner being awarded 0.7 more points per bout on average.

Across all the forty-five matches, the UK judges using Thai criteria differed by a maximum of three points in any one match compared to the previously identified maximum of eight points in UK trained judges applying their commonly used criteria. The average range of the differences calculated using our first method of determining consistency suggested differences were lower in those applying Thai criteria (see table 2.). These differences were statistically significant (t (70.87) = 5.511, P< .0001) and the differences represent a huge effect (Cohen's *d*= 1.17, effect size *r*= .503). The judges using Thai criteria only disagreed on the outcome of two of the forty-five matches compared with eleven previously identified in UK trained judges.

Table 2. The average range (maximum- minimum score awarded by the three judges in each bout) and standard deviations and total point differences across all matches between UK Judges using the Thai system and UK judges using UK system

Judges	Mean	SD	Total differences
UK criteria	2.38	1.54	107
Thai criteria	0.91	0.9	41

The second method also suggested a higher consistency in the Thai trained judges. The One-way ANOVA identified that the difference in the quality of boxer from the red and blue corners varied significantly between the 90 bouts (F (89, 180) = 13.93, P<.0001). The residuals of the ANOVA will reflect the variation or inconsistency between the three judges that could not be explained by the relative quality of the boxers. The standard deviations calculated from the Test for Equality of Variance, using the residuals as the dependant variable, suggested large variation between judges using Thai criteria (SD=.28) and UK criteria (SD=1.2). Judges using the Thai criteria were far more consistent in their decisions. This difference in variances was statistically significant (F= 0.21, P<.0001; Levene's Test statistic = 54.79, P< .0001). Confidence intervals for the

DOI: 10.2202/1559-0410.1178

difference in standard deviations between UK judges using Thai criteria and UK judges using traditional UK criteria are shown in figure 3. Confidence intervals for the difference in standard deviations between Thai judges, UK judges using traditional UK criteria and UK judges using Thai (UKthai) criteria are shown in figure 4.

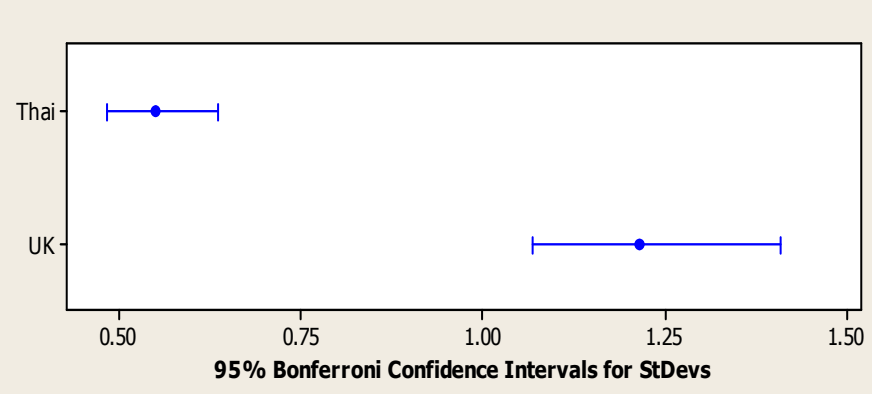

Figure 3. Confidence intervals for the difference in standard deviations between UK judges using Thai criteria and UK judges using traditional UK criteria

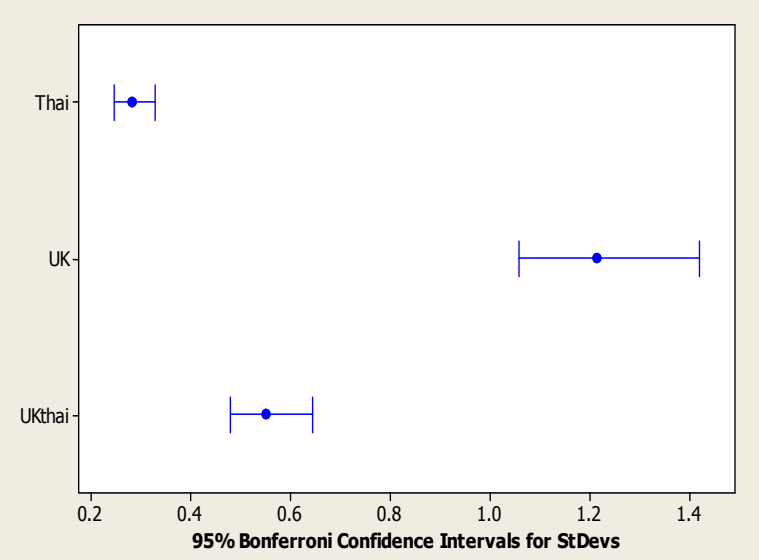

Figure 4. Confidence intervals for the difference in standard deviations between Thai judges, UK judges using traditional UK criteria and UK judges using Thai (UKthai) criteria

Discussion

As in the first study, there was a greater level of consistency in scoring when judges applied Thai scoring criteria. Again, both of the methods used to determine consistency suggested this was the most consistent method. Overall, when comparing all judges' scores together, the Thai judges were the most consistent of the three groups (see figure 4). Given the Thai judges' greater judging experience this may have been anticipated. However, the results of this second study finding that the UK judges were far more consistent when using Thai judging criteria, suggests that the criteria rather than merely the experience and ability of the judges is likely to be responsible for the higher consistency levels. The results appear to support those who advocate that the Thai scoring criteria should be adopted as the international standard.

With the consistency demonstrated by applying Thai judgement criteria, adopting this system internationally has the potential to offer international competitors with a clear direction in selecting and applying appropriate techniques. Without consistency in judging, coaches and competitors do not have a clear direction on technique selection and competition strategy. Uniform judging may well have an impact on the consistency of performers. Myers and Nevill (2008) conducted notational analysis on fights involving UK elite competitors and elite Thai competitors examining frequency particular of techniques delivered and quantitative aspects of delivery. The results suggested that those competitors regularly competing in Thailand were homogeneous in both technique selection and application. One of the reasons speculated for this was the consistent application of criteria by Thai judges in Thailand. Conversely, UK contestants were heterogeneous in technique selection and application. Again this was attributed, in part, to the lower levels of consistency in UK officials applying the traditional criteria applied internationally. The consistent application of criteria by Thai officials has, arguably, given a clear direction to athletes and coaches in technique selection and application and contributed to Thai nationals' international success in international competition. Something that may suggest the consistency evident in this study may be based on performance related factors rather than non-performance related factors.

General Discussion

One plausible reason for the consistency of judges using the Thai judgement criteria is concrete nature of the criteria. This is something that may be transferable to some other subjectively judged sports. Mortimer and Collins (1997) suggested that individuals may have a particular scaling value for pertinent cues, recognising relevant cues and weighting the relative value of each criterion

DOI: 10.2202/1559-0410.1178

in reaching a decision. In the Thai style of judging the weighting of cues is agreed in the form of normative rules used by judges. This means that potentially very subjective criteria are simplified to clear working definitions that can be easily operationalized.

Most translations of Muay Thai rules include a suggestion that techniques should be strong or delivered with power to score. One of the issues that divide those that advocate different judging criteria is how judges should make an assessment of a strong or powerful technique. If judges only make an assessment of power from the actions of the competitor who delivers the blow they have only a range of very subjective cues to make that assessment and agreement of observers can be quite low. However, if judges assess the effect of the technique on an opponent, they have much more easily agreed upon cues with which to make their assessment. These cues include moving an opponent or causing them to lose balance or show pain. Thai judges make an assessment of the relative effect of blows. For example, if one boxer is landing body kicks that have no visual effect on their opponent and the other lands fewer kicks that have a visual effect, the boxer landing the more effective techniques wins that exchange. Officials applying Thai criteria tend to focus on what is easily seen visually, and what can be easily agreed upon (Myers, 2000).

Conversely, in the traditional criteria applied by UK trained officials, all techniques score equally and judges make a more subjective assessment of power. For example, in coming to a judgment decision, UK officials try to determine the number of all the punches that land with force, similar to judges in international style boxing. However, when a judge observes a combination of punches thrown in quick succession, it can be very difficult to determine which of those punches land and which hit the arms or are blocked and therefore do not score. Different judges may well have quite different opinions on which punches landed and which did not. This has been highlighted in amateur boxing where concerns have been aired over difficultly of agreeing on the number of rapid punching blows that land during a bout. This is something that has surfaced recently again at the Beijing Olympic Games where a number of decisions were disputed with claims that judges had not recorded a number legitimate scoring blows (Clark, 2008). Using Thai criteria officials do not attempt to do this, instead they only attempt to determine which of the punches have a physical effect causing a competitor to be moved (Myers, 2000). Not only are there fewer instances of this for judges to monitor, but it is also easier for them to determine this visually and therefore agree upon the number instances this occurs during a round.

Subjectively judged sports often use used very detailed criteria to rate performance and determine outcome. For example, ski-jumping judges combine what can be considered an objective measure of distance with an aesthetic assessment of the style and form of the jump (Federation Internationale De Ski,

2004). Ice skating judges make an aesthetic judgement on each skater's performance and combine this with a technical mark (International Skating Union, 2008). However, there are still issues with consistency in these sports and still frequent disputes. Some sports may benefit by following Muay Thai's example and consider not only the detail of decision criteria but also the number of choices involved in each decision, and how easily outcomes can be agreed upon by a range of judges. Certainly boxing would benefit from a re-evaluation of its scoring and consider having clearer working descriptions of the more subjective elements used to score fights.

The results from the two studies suggest that the use of Thai criteria is significantly more consistent than the criteria commonly used by UK judges and frequently applied internationally. Although there may be a possibility of non-performance influences impacting on this level of consistency, it is likely given the results of the second study that the use of very clearly defined criteria and concrete operationalization of otherwise subjective concepts had a major impact on this. Future research could compare the judges' scores and consistency when they judge the same bout under experimental conditions using a repeated measures design. .In addition it may be useful to examine the impact of modifying judging criteria on consistency in a range of subjectively sports.

References

Annett, J. (2002). Subjective rating scales: science or art? *Ergonomics, 45, 14*: 966 -987

Balmer, N.J.,Nevill, A.M, Lane, A.M., Ward, P. Williams, M. Fairclough, S.H. (2007). Influence of Crowd Noise on Soccer Refereeing Consistency in Soccer. *Journal of Sport Behavior*. *30(2),* 130-145

Balmer, N.J., Nevill, A.M. & Lane, A.M. (2005). Do judges enhance home advantage in European championship boxing? *Journal of Sports Sciences, 23*, 409-416

Board of Boxing Sport. (2002). *Standard Rules and regulations for Boxing Sport Competitions B.E.* 2545. Bangkok: Office of Professional Sports, Sports Authority of Thailand.

Clark, N. (2008).The first post. 2008. Dirty tricks in Beijing? Accessed on 4 January 2009 at http://www.thefirstpost.co.uk/45129,features,there-is-growing-concern-about-partisan-judges-at-the-olympics

D'Agostino, F. (1995) 'The ethos of games', in W. J. Morgan and K. V. Meier (eds.) Philosophic Inquiry in Sport, second ed., Champaign, IL: Human Kinetics

DOI: 10.2202/1559-0410.1178

Duncan R.D. Mascarenhas, Collins, D. Mortimer. P. (2005.). The Accuracy, Agreement and Coherence of Decision-Making in Rugby Union Officials. *Journal of Sport Behavior*, *28* (3), 253-271

Einhorn, H. J. (1972). Expert measurement and mechanical combination. *Organizational Behavior and Human Performance*, 7, 86-106.

Einhorn, H. J. (1974). Expert judgment: some necessary conditions and an example. *Journal of Applied Psychology, 59*, 562-571.

Federation Internationale De Ski (2004) *Book III Ski Jumping, The International Ski Competition Rules (ICR) Approved by The 44th International Ski Congress, Miami (USA)* Thunersee: FIS

Gartland, S., Malik, M.H. A. and Lovell, M. E. (2001). Injury and injury rates in Muay Thai kick boxing. British Journal of Sports Medicine 35: 308-313

International Skating Union (2008) *Special Regulations and Technical Rules Single and Pair Skating and Ice Dance* as accepted by The 52nd Ordinary Congress June 2008 (accessed at http://www.isu.org/vsite/vfile/page/fileurl/0,11040,4844-191592-208815-140518-0-file,00.pdf)

Kaczmarek, T. (1996) *You be the Boxing Judge: Judging Professional Boxing for the TV Boxing Fan*. Pittsburgh: Dorrance.

Lee, J. (2008) Outlier Aversion in Subjective Evaluation: Evidence from World Figure Skating Championships. *Journal of Sports Economics*, *9*, (2), 141-159

Lee, H. K. H., Cork, D. L., & Algranati, D. J. (2002). Did Lennox Lewis beat Evander Holyfield? Methods for analysing small sample interrater agreement problems. *Journal of the Royal Statistical Society, Series D (The Statistician), 51*, 129 – 146

Manns, J. W. (1998). *Aesthetics*. New York: M.E. Sharpe.

Mortimer, P. W., & Collins, D. J. (1997). Coherence of decision-making in team sports. *Paper presented at the BASES Annual Conference,* York.

Muckler, F. & Seven, S. (1992) Selecting performance measures: `objective' versus` subjective' measurement, *Human Factors, 34*, 441 -455.

Mumford, S. (2006). Truth makers for judgement calls. *European Journal of Sport Science*. *6*(3): 179-186

Myers, T.D. (2000) Judging a Thai Boxing contest. *United Kingdom Muay Thai Magazine* July/August, 35-38.

Myers, T.D. (2007). Cultural differences in judging Muay Thai BASES abstracts, *Journal of Sports Sciences, 25*(3), 235 – 369

Myers, T.D. & Nevill A.M.(2008). The effects of different judging styles on technique selection of elite Thai and UK Muay Thai competitors. *Poster presentation at the BASES annual conference Brunel University* 2 - 4 September

NAAG (2000). National Association of Attorneys General Boxing Task Force Report, May 2000 (accessed at: http://www.oag.state.ny.us/press/reports/boxing_task_force/report.html).

Nevill, A.M., Balmer, N.J., & Williams, A.M. (2002). The influence of crowd noise and experience upon refereeing decisions in association football. *Psychology of Sport and Exercise*, *3,* 261-272.

Plessner, H Haar T. (2006). Sports performance judgments from a social cognitive perspective. *Psychology of Sport and Exercise 7*, 555–575

Plessner, H. (2005). Positive and negative effects of prior knowledge on referee decisions in sports. In T. Betsch, & S. Haberstroh (Eds.), *The routines of decision making* (pp. 311–324). Hillsdale: Lawrence Erlbaum.

Rainey, D., & Larsen, J. D. (1988). Balls, strikes, and norms: rule violations and normative rules among baseball umpires. *Journal of Sport and Exercise Psychology*, 10, 75-80.

Rainey, D., Larsen, J. D., Stephenson A., & Olson, T. (1993). Normative rules among umpires: the "phantom tag" at second base. *Journal of Sport Behavior*, *16* (3), 147-155.

Scheer, J., Ansorge, C. J., & Howard, J. (1983). Judging Bias induced by viewing contrived video tapes: A function of selected psychological variables. *Journal of Sport Psychology*, *5*, 427-437.

Shanteau, J., Weiss D. J. Thomas R. P. Pounds, J. (2003) *How Can You Tell if Someone is an Expert? Empirical Assessment of Expertise. Emerging perspectives on decision research.* Cambridge, U.K.: Cambridge University Press.

Silva, J. (1981). Normative compliance and rule violating behavior in sport. *International Journal of Sport Psychology, 12*, 10-18.

Stefani, R. (1998). Predicting outcomes. In J. Bennett (Ed.), Statistics in sport (pp. 249–275). London: Arnold.

Vanden Auweele, Y., Boen, F., De Geest, A. and Feys, J. (2004) Judging bias in synchronized swimming: Auditory feedback leads to nonperformance-based conformity. *Journal of Sport and Exercise Psychology* 26, 561-571.

Wanderer, J. J. (1987). Social factors in judges' rankings of competitors in figure skating championships. *Journal of Sport Behavior*, *10*, 93 – 102.

Weekley J.A., & Gier, J.A. (1989). Ceilings in the Reliability and Validity of Performance Ratings: The Case of Expert Raters. *The Academy of Management Journal, 32*, 1: 213-222.

Wenger, E (1999) *Communities of Practice. Learning, meaning and identity*, Cambridge: Cambridge University Press

World MuayThai Council (1995). *Constitution, by-laws, rules and regulations*. Bangkok: WMTC.

DOI: 10.2202/1559-0410.1178