1-1-2005

A Naturalistic Approach to Training Accurate and Coherent Decision Making in Rugby Union Referees

Duncan R. Mascarenhas
Glyndwr University, d.mascarenhas@glyndwr.ac.uk

D Collins

P Mortimer

R L. Morris

This is the published version of an article that was first published in the Sport Psychologist Journal in 2005 by Human Kinetics. The journal website can be located at http://hk.humankinetics.com

The Sport Psychologist, 2005, 19, 131-147

A Naturalistic Approach to Training Accurate and Coherent Decision Making in Rugby Union Referees

Duncan R.D. Mascarenhas, Dave Collins, Patrick W. Mortimer, and Bob Morris
University of Edinburgh, UK

The purpose of this investigation was to pilot the use of a video-based training program designed to develop referees' shared mental models. A group of English Rugby Football Union (RFU) national referees, divided into a control group (n = 15) and experimental group (n = 41) made their immediate decisions on pre and posttests of 10 video recordings taken from real game scenarios. Over a six-week period the experimental group studied training tapes consisting of 5 sets of 5 tackles, in each case with an expert providing his interpretation of the correct decision. Each clip was filmed from the referee's perspective and taken from real game situations in order to maintain high ecological validity in accordance with naturalistic decision-making theory. The lowest ranked referees on the national panel significantly improved their percentage of correct decisions, becoming 17.43% more accurate in their decisions at the posttest. These results suggest that such shared mental model training is an appropriate method for improving referee decision making. **\bb**164 words.

Referees are responsible for maintaining flow and control (FIBA, 2004) and ensuring fair play both in accordance with the laws and the spirit of the game (iRB, 2003). Their performance is critical. Not only are referees responsible for maintaining a safe environment, but they are also expected to ensure that the game's result is just. The increasing use of video replay to assist officials at the highest level signifies just one of the many attempts to ensure that referees are accurate

Duncan R.D. Mascarenhas, Dave Collins, and Patrick W. Mortimer are with the Department of PE, Sport, and Leisure Studies at the University of Edinburgh, St Leonard's Land, Holyrood Road, Edinburgh EH8 8AQ, Scotland. Email: duncan.mascarenhas@education.ed.ac.uk. Bob Morris was with the Department of Psychology at the University of Edinburgh.

in their decisions. Furthermore, refereeing professional sport requires a variety of skills (Anshel, 1995; Mascarenhas, Collins, & Mortimer, 2002a). Expert referees need to be physically fit enough to keep up with play, able to apply an accurate interpretation of the laws of the game and have the personality and management skills to "sell" these resulting decisions to the players.

However, while governing bodies have focused on physiological assessments and fitness training for their referees, structured training in decision making is rare (see Garcia, 2003). There is also very little scientific research that has investigated referee decision making and none that has attempted to train referees' understanding and application of the laws. Since this area has been identified as the most important competency for referees (Anshel & Webb, 1991; Anshel 1995), ranking well above fitness and communication skills, this paucity of research and the governing bodies' current emphasis needs to be rectified.

Currently, referees appear to rely on experience to develop this expertise. Since a large body of research suggests that expertise requires 10, 000 hours of deliberate practice to master, both in the sporting environment (Helsen, Hodges, Van Winckel, & Starkes, 2000; Helsen, Starkes, & Hodges, 1998; Starkes, 2000) and in more cognitively orientated activities such as chess (Simon & Chase 1973), this purely experiential approach would equate to about 7,000 rugby union games, the equivalent of refereeing one game per day for 35 years, assuming the season is 200 days long (Glaser, 1984). Furthermore, research by Williams and Davids (1995) suggests that mere experience will not necessarily lead to expertise. Refereeing itself may not provide a sufficient number of challenging scenarios and in close succession to develop expert performance (cf. Means, Salas, Crandall, & Jacobs, 1993; Starkes & Lindley, 1994). In short, purely "learning on the job" appears to be a questionable strategy, and some form of specific training in this crucial component of the referee's art is essential.

Typically, to train expert decision-making performance, traditional, prescriptive training models have been employed. Under this approach, the trainee would be asked to watch an event, generate a range of alternative solutions, evaluate each against one another, and choose the most appropriate course of action. However, while these traditional, normative models of decision making have been successfully used for routine decisions (Beach & Lipshitz, 1993; Edwards & Newman, 1982), they have often been erroneously applied to ill-structured, dynamic, and time-pressured tasks that typify the real-world environment (Lipshitz, 1993).

Reflecting these challenges, decision-making research has experienced a paradigm shift from the traditional, normative models to a naturalistic approach (Cannon-Bowers, Salas, & Pruitt, 1996). Modern decision researchers have become more interested in studying real-world decisions made under realistic conditions, leading to the emergence of naturalistic decision making (NDM). This approach specifically sets out to test and train expert decision making in environments high in ecological validity (Klein, 1997**\bb**).

How the Experts Make Decisions: Applying NDM to Refereeing

Ill-structured problems; uncertain dynamic environments; shifting, ill-defined, or competing goals; action/feedback loops; time stress; high stakes; multiple players; and competing organizational goals and norms characterize the NDM environment (Orasanu & Connolly, 1993). Not all of these challenging characteristics have to

be present to deem the environment appropriate for naturalistic study (Orasanu & Connolly, 1993), nonetheless a quick analysis of refereeing in team invasion games presents an extreme case, as all nine of the characteristics are present. Refereeing often presents complex and unstructured scenes with players moving at high speeds, where a referee has flow and control issues to balance, making decisions that are not isolated but wrapped into evolving patterns of play. They are expected to react instantaneously, often under intense media scrutiny, while accounting for the input from other refereeing team members (in rugby union's case, the touch-judges, who control the sidelines and have a microphone link to the referee). They also have to balance their interpretation of law with the organization's philosophy on the way the game should be officiated. Given such demanding circumstances and extreme time pressures, NDM seems to provide the most appropriate paradigm of study for referee decision making.

NDM proposes that an expert's decisions are governed by their superior organization of knowledge into knowledge structures (Druckman & Bjork, 1997**bb**\), enabling a more rapid response. For example, Stokes, Kemper, and Kite (1997) discovered that while experienced aircraft pilots do not have any cognitive advantages over their less experienced colleagues (such as better reasoning ability), their access to knowledge representations from long-term memory enables them to more readily recognize relevant cues and achieve superior performance.

There are many theories that consider how this knowledge is organized (for a review see Cooke, Salas, Cannon-Bowers, & Stout, 2000), however the literature is replete with inconsistent terminology and speculations that lack empirical data (Rouse & Morris, 1986). Nevertheless, it seems clear that experts use their superior organization of such knowledge to guide rapid decision making. As refereeing involves both perceptual and cognitive DM components (Ste-Marie, 1999), the most inclusive explanation of how this knowledge is organized is that experts use mental models.

Mental models are a special type of knowledge (Rouse & Morris, 1986), defined as symbolic representations of conceptual knowledge, holding information on the task, the situation, and the team and are responsible for providing information on future status. Consequently, they help individuals to make sense of situations, predict the likely consequences, and thus guide them into following appropriate courses of action (Rouse & Morris, 1986). The construct has been widely accepted for investigating human understanding of physical systems (Gentner & Stevens, 1983) and employed in the study of experts making rapid decisions (Stout, Cannon-Bowers, & Salas, 1996). Its visual, symbolic nature also lends itself to the study of referee decision making.

In an effort to explore the shared nature of knowledge structures, researchers have examined the extent to which mental models are common among team members (Webber, Chen, Payne, Marsh, & Zaccaro, 2000). Cannon-Bowers, Salas, and Converse (1990) suggest that such shared mental models (SMMs) are the key to team decision making, allowing implicit coordination through a shared understanding of (a) the problem definition, (b) the plans and strategies for solving the problem, (c) the interpretation of cues and information, and (d) the roles and responsibilities of the team members (Orasanu, 1990). Similarly, these SMMs provide knowledge of the situation to direct attention, classify information, form an understanding of meaning (Stout et al., 1996), and create expectations regarding future states of the environment (Endsley, 1997). Thus, SMMs seem to assist both interdependent

decision making, which is essential to a referee performing in a team of three (that is when accompanied by two touch-judges), and independent decision making, assisting different referees to officiate in the same way, week after week. In addition, they provide one solution to measuring refereeing performance success.

Clearly the accuracy of referees' law application is the critical marker for success. However, mere conformity to the same decision outcome, which may be accurate in law, may not be built on the same mental model. For example, if two different referees have arrived at the same decision through different decision-making processes they do not share the same understanding of the situation and are unlikely to repeat this concordance when one or more of the parameters change. Accordingly, it is important not just to evaluate the decision, but also the reasons that underpin each decision, which will reflect the SMM (see Langan-Fox, Code, & Langfield-Smith, 2000) in order to generate improvements in decision-making performance (Mascarenhas et al., 2002a). When both decision and reasoning are correct, the referee has a complete mental model of the situation and can truly be described as accurate. When different referees share these two-levels of accuracy they can be described as coherent. Consequently, players will experience more consistent and understandable decisions, increasing both player and spectator satisfaction.

Training Accuracy and Coherence With Scenarios

Effective teams need a variety of complex mental models to form an understanding and generate predictions about likely events (Salas, Cannon-Bowers, & Johnston, 1997). Accordingly, presenting "typical," filmed scenarios, with cues and patterns that characterize the domain, may be an attractive alternative to on-the-job learning (Means et al., 1993; Cohen, Freeman, & Thompson, 1997).

Using scenarios offers the flexibility to provide clear and timely feedback rather than waiting until sometime afterward, which may be long after the incident under scrutiny has occurred (Means et al., 1993). In addition, a variety of scenarios can be provided that might otherwise take years to encounter. Practice problems can also be tailored specifically to challenge the individual decision maker's current level of performance. Calderwood, Crandall, and Klein (1987**\bb**) suggest that novice decision makers overlook or are unable to see the important information, thus preventing them from generating accurate and complete mental representations of the situation (Helsen & Pauwells, 1993; Williams, Davids, Burwitz, & Williams, 1994). So, since the experts' ability to discern the important from the irrelevant information is a key determinant that distinguishes them from novices (Abernethy & Russell, 1987; Williams, Davids, & Williams, 1999) training scenarios can be designed to account for such process measures (the reasoning) by directing attention to the significant features of a problem (Rouse & Morris, 1986), rather than by using outcome measures alone (Brannick & Prince, 1997). Typically this would be done explicitly by building a SMM of how situations should be refereed, providing less experienced referees with the implicit knowledge to search for the salient cues, how they relate to each other, and how they should impose a weighting to this information in a variety of different situations (see Balke, Hammond, & Meyer, 1973). Rasmussen (1985) termed this weighting scale hierarchical task analysis, suggesting that exposure to this functional hierarchy of subtasks serves to direct attention, reduce mental workload, and enhance accuracy.

Purpose of the Study

In the specific context of rugby-union, the diverse and dynamic environment of the tackle (law 15) provides a uniquely demanding task that is appropriate for elucidating a referee's decision-making weighting scale (E. Morrison, personal communication, July 5, 1998), requiring a clear understanding of the refereeing priorities (Bunting, 1999). Therefore, the primary aim of this investigation was to pilot the use of a video-based coherence training program, designed to accumulate and reinforce referees' SMMs. Following NDM guidelines, the aim was to use very specific, realistic, knowledge-rich scenarios, presenting ambiguous information with auditory interference (represented by crowd noise) and demand a time pressured response (Johnston, Poirier, & Smith-Jentsch, 1998). We hypothesized that by presenting a variety of scenarios with detailed reasoning, provided by a "high-status" expert (see McCullagh, 1986) to reinforce a decision-making hierarchy (Rasmussen, 1985; Eylon & Reif, 1984), referees would show improvements in both the accuracy and coherence of their decisions from pre to posttest. A final aim was to collect feedback from the referees to obtain more qualitative insights into the efficacy and mechanisms of such training.

Method

Participants

RFU referees (n = 56), ranging in age from 26 to 51 years and who had officiated on the national panel from 1 to 16 years, volunteered to take part. As a feature of their continuing professional development, these referees attended 6 weekly development meetings held at four regions across England. This study reflects one aspect of the sports science support program, conducted by the first author at a series of these regional meetings.

A group of 41 referees, who attended both the pre and posttests held at two consecutive meetings, were assigned to the experimental (coherence training) group. The remaining referees were assigned to either a passive control group (n = 7) or an active control group (n = 8). The passive control referees were those who were unavailable between the two tests as they were either on holiday between the meetings (n = 5), moving house (n = 1), or unable to get access to a video recorder (n = 1). Based on their national ranking, the group of active control participants were asked to watch the same tackles as the experimental group but without the expert's interpretations, on the understanding that they could complete the training after the study. Referees who failed to attend both tests or failed to complete all aspects of the training were eliminated from the study.

Specifically, we were interested in noting the differential effect of training referees performing at different levels. A referee's career progression is determined by improving his or her standing on a national ranking system, made by a group of referee development officers from the periodical evaluations from 37 advisors. This placed referees into 1 of 3 groups: a top-20 group (n = 11), who were responsible for refereeing at premier league level; a mid-panel group ranked from 21-47 (n = 17), responsible for national leagues 1 and 2; and a lower-panel group ranked from 48-65 (n = 13), who officiated at national league level 3 and 4. Each of these groups was further subdivided into two balanced groups with similar numbers and rankings, enabling a reverse baseline test. Although lacking in scientific rigor (as

with so many similar systems), this ranking system represents the basis for reward and recognition and even progression to the pinnacle of international officiating. As such, it holds considerable ecological validity and allowed us to examine the effects of the training package on referees performing at different levels.

The control group consisted of 15 referees from the various ranked groupings (top-20, n = 3; mid-panel, n = 7; lower-panel, n = 5), again subdivided into two balanced groups. An independent t-test between the collective mean ranking of the training groups (M = 35.54, SD = 16.71) and the control groups (M = 35.13, SD = 19.34) showed no significant differences, t(1, 54) = .077, p = .939.

Developing the Assessment and Coherence Training Tapes

Using a Panasonic AGDP800HEG S-VHS camcorder, 12 English RFU premier league games were recorded onto a series of Super VHS videotapes by the first author, who had worked as a professional sports camera operator and videotape editor over a period of 12 years. The camcorder was mobile, supported by a monopod for stability, allowing the camera operator to move up and down the sideline of the pitch and stay level with the play. At every break in play—when a penalty, scrum, or lineout was awarded—the camera operator moved adjacent to the action, providing an angle similar to that which the match-day referee adopted. This own-point-of-view recording provides minimum distortion of the complexity and dynamics of naturalistic environments (Omodei, McLennan, & Whitford, 1997**\bb**).

From these tapes, 126 tackle incidents were edited onto a master tape. An independent expert panel consisting of the three highest ranked referees in England (who were not participants in the study) rated each tackle on ambiguity and quality of information available. Then, all the tackles that yielded a consensus opinion from the three independent experts, and were adjudged to exhibit sufficient information and present realistic match situations for refereeing, were ordered in terms of ambiguity. The experts rated these remaining 45 most ambiguous tackles on difficulty, providing two balanced groups of 10 tackles (labeled A1 - A10, and Z1 - Z10) for the reverse baseline pre and posttest and 5 sets of 5 training tackles. To further validate the two groups of tackles as balanced, an independent t-test showed no significant differences between these A and Z tackles on pretest accuracy scores, t(1, 54) = 1.735, p = > .05).

Pre and Posttest Assessment Tapes. Each of the 20 pre and posttest clips commenced with a voice-over that introduced the two teams competing, indicating which team had possession and the direction in which they were attacking. The tackle incident then began with approximately 5 seconds of lead-in, the period confirmed by pilot and previous studies to be necessary to allow the participants to orientate themselves to the scene. After the tackle incident, the videotape image froze, presented the title "make your decision now" and cut to a blank screen after about 5 seconds.

The Coherence Training Tape. This contained the 25 tackles used for training, edited in exactly the same manner, only arranged into 5 sets of 5 clips. After each set of five clips, the tackles were replayed, followed by Ed Morrison (at that time the number 1 ranked RFU referee), giving a detailed explanation of his interpretation to the camera, concluding with a further rerun of each tackle clip. Consistent throughout his interpretations was the hierarchical theme of (a) getting the tackler to roll away immediately, (b) allowing the ball carrier to release the ball, and (c) then ensuring that support players arrived on their feet. Ed Morrison

was chosen as the expert to provide the model interpretations as an active referee whose decisions would reflect the way in which the English game should be officiated. Furthermore, having refereed the world cup final in 1995, and as England's most capped international referee, it was anticipated that his opinions would have credibility with the participants (see McCullagh, 1986).

Instrumentation

A response sheet was developed to enable participants to quickly and easily signify their decision. The first section of this sheet consisted of a series of six boxes and asked the respondent to tick the appropriate box. This presented options to (a) play on, (b) award a penalty to the defending team, (c) award a penalty to the attacking team, (d) award a scrum to the defending team, (e) award a scrum to the attacking team, and (f) other, with a space to explain this "other" decision. The response sheet then asked participants to explain the reasons behind each decision and indicate on a Likert scale their confidence in the accuracy of each decision, scored from 1 (*low*) to 5 (*high*).

A coherence training booklet was given to each participant in the experimental group, which in addition to having 25 response sheets (one for each of the training tackles) identical to those for the pre and posttests, included a viewing log to record how often and on which dates they watched the two training tapes. Additionally, the booklet had an "acceptance sheet" asking the participants the extent to which they accepted the expert's model interpretation for each tackle and to provide their reasoning. At the end of the booklet, a feedback sheet asked the participants to comment on the quality, value, and ecological validity of the video training package.

Procedure

Pre and Posttests. All the participants completed the pretest at their regional development meeting. The posttest was carried out in exactly the same fashion at the next meeting 6 weeks later. Across all four regions, there were no more than 18 participants watching the assessment clips at any one time, and each was informed that their own personal responses would remain confidential and that the results would only be presented on a group basis.

After the participants familiarized themselves with the response sheet, they adopted a position where they could comfortably see the tackle incidents, projected onto a screen via a standard VHS video recorder and data-projector. This presented an image about 6 feet wide and 4.5 feet high. The assessment videotape was then replayed, showing either clip A1 or clip Z1, depending on the group to which the referees were assigned. The video was paused immediately after each clip and participants were asked to make an immediate decision by ticking one of the six boxes. They were explicitly told not to change their decision once made. Both inspection of the response sheets and observation of participants revealed that they all conformed to these instructions. The participants were then given three minutes to write down the reasons for their decision and indicate their confidence scores.

This procedure was conducted by the first author and followed consistently for both experimental and control groups in all four regions. With permission from the RFU, all the participants conducted no additional, structured decision making or video-based training during the whole testing period. This might typically have

included group discussions prompted by match-day recordings, or the RFU releasing updates on interpretations of law.

Experimental Group. Each experimental group referee was given a copy of the coherence training tapes and booklet and asked to watch the tape in the same manner on a set night each week and to record each inspection in the booklet's viewing log. After viewing the first five clips once and completing response sheets, they then watched the model interpretation section and reviewed these same clips as many times as they felt necessary during the week to understand the model interpretation. The viewing logs revealed that all the referees were both reliable in conducting their first viewing on the same night each week and viewed the model interpretation section at least one other additional time before the following week's training. These participants then completed the acceptance sheet, and finally after all the training, they completed the feedback section, commenting on the efficacy of the training package.

Control Groups. To control for any expectancy and potential training effects, which might occur merely as a consequence of watching a greater number of tackle incidents, the active control group watched exactly the same tackle clips as the training group but without the model interpretations. They were asked to reexamine the clips at least one other time later in the week. Examination of their viewing logs revealed that they conformed to this request. All the other control group members were passive, simply watching the pre and posttest clips in the same fashion as all the other participants.

Data Analysis

The primary analysis was quantitative, measuring the accuracy and coherence of participant's responses. Secondarily, the qualitative data from the acceptance sheets and feedback sheets were used to identify trends, providing evidence to show how changes may have occurred.

Quantitatively, the participant's responses were only considered to be accurate if they provided both the correct decision and the correct reasons underpinning that decision, as deemed by the expert. Thus, if a referee arrived at the correct decision but with incorrect reasoning, this was considered to be inaccurate, and was grouped with responses that reflected the wrong decision. Referees sharing these two levels of accuracy were considered to be coherent; hence, coherence was the percentage of accurate participants.

The referees' qualitative feedback on the validity and value of the test and the reasons for their acceptance levels were transcribed and analyzed by group. In addition to noting the comments, simple evaluations were made using a frequency count of similar responses. A second researcher, unconnected to the study, examined the frequency and distribution of these comments and felt that all were valid, credible, and reliable, thus reaching 100% agreement with the primary researcher.

Results

Quantitative Results

Before examining the effects of the intervention, it was first necessary to establish the equivalence (or lack thereof) of the two control groups. Accordingly, a 2×2 (group $\times$ time) ANOVA with repeated measures on the second factor, comparing

pre and post performance achieved by the active and passive controls. No significant differences were apparent on any of the effects, time: $F(1,12) = 1.29$, $p = .28$; time × group: $F(1, 12) = .33$, $p = .58$; or group: $F(1, 12) = .12$, $p = .78$. Accordingly, these two were combined as a single control group in all subsequent analyses. These nonsignificant results also enabled us to accept any changes in the experimental groups as due to the intervention and not just an artifact of viewing the tapes, either through expectancy or just through a greater amount of deliberate practice.

The pre and post intervention accuracy scores achieved by the different groups (presented in Table 1) show very high standard deviations signifying very large variance within each group. Changes in performance were examined by a 4 × 2 (group × time) ANOVA, with repeated measures on the second factor. The dependant variable was the percentage accuracy achieved. Two significant effects were apparent: the main effect of time pre and post the intervention, $F(1, 52) = 5.06$, $p = .029$ and the group by time interaction, $F(3, 52) = 3.01$, $p = .038$. Effect sizes (eta squared) were .071 and .146, respectively, reflecting medium and large values (Clark-Carter, 2001, p. 254). The main effect of group was nonsignificant: $F(3, 52) = 1.26$, $p = .298$, effect size = .065, power = .32. Follow up using Scheffé tests suggested that the interaction effect was due to the significant improvement in performance of the lower ranked referees from the control group. No other changes achieved significance.

A second 4 × 2 (group by time) ANOVA was used to examine changes in confidence scores (see Table 1), which revealed a significant group effect: $F(3, 52) = 3.34$, $p = .026$, effect size = .162. Follow up with the Scheffé test showed that this was due

Table 1 Percentage Accuracy by Group

	Top-20		Mid-panel		Lower-panel		All-controls		All training groups	
	M	SD	M	SD	M	SD	M	SD	M	SD
Pre % Accuracy	46.91	11.40	47.06	13.24	39.75	15.24	45.87	13.16	44.70	13.56
Post % Accuracy	50.55	14.81	51.73	13.61	57.18	12.53	41.25	14.44	53.14	13.57
Accuracy Change	3.64		4.67		17.43*		-4.62		8.44	
Pre Confidence	4.05	1.21	4.05	1.04	3.69	1.15	4.17	.92	3.94	1.13
Post Confidence	4.31*	.95	4.07	1.00	4.02	1.15	4.11	1.03	4.12	1.04
Confidence Change	.26		.02		.32		-.06		.18	

*p = < .05

to differences between the top-20 referees and the lower-panel referees. The time effect of confidence across all participants showed a small increase approaching significance: $F(1, 52) = 2.94$, $p = .092$, effect size = .053, changing from M = 3.95% (SD = 1.13) to M = 4.10% (SD = 1.05), whereas the control group's time effect showed a non-significant decrease from pre to posttest $F(1, 12) = 1.08$ $p = .52$.

Finally, to offer a preliminary insight into the mechanisms underlying the changes observed, a post hoc examination of the percentage accuracy increases was conducted on each individual tackle clip. Two tackles showed the most marked improvements from pre to post testing, clip A8 (pre = 33%, post = 79%) and clip Z2 (pre = 35%, post = 91%). Both these tackles presented situations whereby, in the expert's interpretation, "the ball carrier has become slightly isolated from his support and failed to release the ball." These data pertaining to the collective understanding of the situation are considered later in the discussion.

Referee Qualitative Feedback

All the participants in the experimental groups reported the process as valuable and worthwhile with comments such as "the video is much better than words" and "it helps me visualize what players should and should not do." In addition, the referees felt the assessment tackles to be good representations of game situations, with each tackle clip yielding sufficient information to make a decision.

The viewing logs revealed that all the referee groups watched each tackle approximately the same number of times, on average about 6 times per clip. Nevertheless, there were 16 comments from the lower-panel referees on having to see the tackle more than once, whereas none of the higher ranked referees noted such. For example, one lower-panel referee recognized his initial mistake explaining, "I can see more clearly on the 2nd/3rd/4th viewing," and another remarked, "On the 2nd viewing I agree [with the model interpretation], it is clear that the tackler makes little effort." Seven of these lower-panel referees also commented on gaining a greater understanding of the priorities at the tackle as summarized by one referee who said "I feel I became more aware of my priorities at the tackle after watching the clips." This improved clarity of the priority system was described by another referee who suggested that "the training package does very well, reinforcing the sequence: (a) did the tackler move, (b) did the tackled player release the ball, and (c) did the next players arrive on their feet." By comparison the mid-panel group only made only one such comment, while the top-20 referees made no comments at all. When responding to the training tackles, the higher ranked referee groups (top-20 and mid-panel) tended to offer more detailed reasons underpinning their decisions than the lower-panel referees.

Discussion

The findings of the present study suggest that this video-based training package is appropriate for developing SMMs in pre-elite referees. Referee comments reinforce the efficacy of using video scenarios to train accurate and coherent decision making. Specifically, the lower-panel group who experienced the most improvement felt that their exposure to the model's priority system (cf. hierarchical task analysis; Rasmussen, 1985) was the reason for their enhanced performance.

In contrast, it is perhaps surprising that both the top-20 and mid-panel ranked referees showed only small, nonsignificant improvement in both per-

formance and confidence scores. However, support for the training from the participants was overwhelming, and all groups adjudged the exercise to be valuable. Importantly, the referees reported the assessment tackles to be a fair test of referee decision making, accurately reflecting decisions that have to be made on the field of play.

Encouragingly, this approach now provides a means to identify and train problem areas in refereeing. Subsequently, the RFU have employed this type of training as a method to both reinforce the philosophy of refereeing the tackle and initiate new interpretations that descend upon them from the International Rugby Board (the sport's governing body), sometimes midway through a season. As a consequence, the tapes that provided such detailed descriptions of law interpretations have also been used to assist players' and coaches' understanding of the refereeing philosophy. Other similar invasion games such as basketball, soccer, and hockey could equally benefit by identifying controversial laws, presenting a variety of referee perspective video clips of these laws, together with an expert's detailed decisions and the reasons underpinning them in order to develop shared weighting scales among referees. Such systematic training would not only speed up referee development but could also be used to increase player and coach understanding, which may lead to a reduction in controversial incidents. Moreover, coach and player education in the application of the law could potentially reduce the number of infringements and lead to more flowing and attractive games.

Further development of SMMs is facilitated through the increased interactions and discussions, which now occur at regular referee meetings, driven by coherence-based video exemplars (cf. Kraiger & Wenzel, 1997). Using video, together with developing a coherent language may be a solution to Rouse and Morris' (1986) concerns of capturing mental models that may be largely pictorial through words alone. These referees have developed a working language, using phrases such as SMMs, coherence and "priorities in the tackle," to conceptualize and describe the nature of their task. The development of this common vocabulary has become an essential component of RFU referee training and can be considered to be an intervention in itself via enhanced communication.

Why Did the Lower Ranked Referees Improve?

While lower-panel referees were often inaccurate in their initial interpretation of the training tackles, many explained that after hearing the expert's interpretation and viewing the clip again, the reasoning became more apparent. Thus, it may be that their improved coherence in the posttest is as a result of a richer store of incidents in long-term memory, accessible by retrieval cues that pre-prime these referees into making the appropriate decision (McLennan & Omodei, 1996). Such prior knowledge or pre-priming, which may have occurred as a result of improved cue-utilization (Stokes et al., 1997), has already been found to affect referees decisions in soccer (Jones, Paull, & Erskine, 2002).

This explanation seems even more plausible since post hoc analysis revealed that tackles in which the most improvement was made were those in which the referees appeared to be pre-primed toward the likely outcome. That is, incidents where the ball carrier was "isolated from his support" and became susceptible to illegally holding on to the ball.

Why Didn't the Higher Ranked Referees Improve?

Given the aims of this intervention, it is equally important to consider the reasons underlying its failure to increase the coherence of the higher ranked referees. This may be due to two reasons. First, the higher ranked referees were more explicit in their interpretations than their lower ranked peers, both at the pre and posttest. The top-20 referees tended to offer more alternative and face-valid interpretations for each tackle, although these were often markedly different to the model answers. In other words, while these higher-ranking referees had more complex and developed mental models, these were not shared across peers. The positive aspects of the elite group's performance mirrors Orasanu's (1990) findings where high-performance cockpit crews were more explicit and revealed a more complex understanding of situations than low-performance pilots, resulting in quicker and safer decisions in emergencies. It also highlights the importance of the model interpreter. The advantage of using just one expert is that it is more likely that a consistent message will be presented, which was important for this preliminary intervention; however, future studies may find it valuable to present a consensus opinion from a group of international referees to help ensure that the model answers are representative of the way the game should be refereed.

The second reason may be the expert's diminished influence on this group of referees. The top-20 referees were much more equivalent in status to the expert and as such may have been more resistant to change their own mental models. Taking these two points together, the more complex and robust mental models of the higher ranked referees may explain their nonsignificant improvements.

Finally, the lack of significant differences found between the groups must be considered against the comparatively low power, which given that four groups were examined was conservatively calculated as 0.32. In fact, to reach the generally accepted levels of 0.8 (Cohen, 1988), we would have needed approximately 40 participants per group. Even though this estimate is a conservative worst-case scenario, the possibility that the analysis lacked the power to discriminate between the group's performance levels should be borne in mind. This distinct possibility notwithstanding, however, the improvements generated by the intervention are clear.

Why Only 50% Accuracy?

While it may be surprising that none of the referee groups achieved even 60% accuracy, one must consider the particular nuances of the sport that was scrutinized in this paper. Recent referee decision-making research to date has explored more "matter of fact" scenarios, such as the offside decision in soccer, asking merely whether the player was offside or not? At the next level, Plessner and Betsch (2002**\bb**) and Jones et al. (2002) have considered "matter of opinion" decisions, asking soccer referees to judge whether a foul was committed and if so, by whom. In contrast, however, the present paper explored a third and hereto unconsidered level of complexity. Refereeing the tackle in more open sports like rugby union presents a unique situation where multiple, complex, and dynamic decisions are required, as there are timing elements, overlapping elements, and interactive elements (see Ackford, 2003). In essence, the degrees of freedom in this situation are so great that the level of accuracy demonstrated in this study may be appropriate.

Furthermore, a rugby union referee is much more than a mere regulator of the law. A feature of refereeing this sport is the notion of advantage. For example, the referee will recognize that an offence has occurred but may choose to ignore it if no advantage has been gained, or simply manage it, perhaps through communicating to the players, to balance the trade-off between game flow and control. Indeed, it is the referee's ability to allow the game to flow, but also maintain the control of the players, termed contextual judgment (Mascarenhas, Collins, & Mortimer, 2002b) that is crucial. Furthermore, the increased degrees of freedom in rugby union may make the contextual factors even more crucial. This may in part explain why the higher ranked referees were unable to improve upon their initial performance, as contextual judgment at premier league level may well supersede law application as the more critical factor. This area requires further investigation, but the importance of contextualizing results against the specific challenges inherent in the game are well evidenced by this situation.

Nevertheless this study, together with previous (Mascarenhas, Collins, & Mortimer, in press), and subsequent investigations, establishes the standard to be as low as just over 50% accuracy, despite these samples including several international and ex-international referees. Interestingly, when 12 of England's elite rugby league referees were shown clips of similar tackle situations in their sport (one which presents slightly fewer degrees of freedom or reasons for awarding a penalty than the rugby union tackle), they achieved only 63% accuracy by the same measures employed in this paper.

It must also be recognized that the results of this investigation present an extreme score, as the referees were only considered to be accurate if they achieved the correct decision and the correct reasoning. Although examining the decision alone may give results more similar to other sports (typified by yes/no and judgment decisions), if they are not accurate on at least these two levels in rugby union, you cannot assume any SMM to be apparent. In fact, since the tackles that proved to be most trainable were ones in which the referees appeared to be pre-primed into their decisions, this investigation seems to suggest there to be a third level of coherence. Referees should not only award penalties to the correct team, and for the same reason, but they should also have the same understanding of that situation, which may help to pre-prime them into anticipating the event before it occurs. This remains a consideration for future research. Another reason for such apparently low scores may be attributed to the tackles that were specifically chosen as a range of difficult situations that regularly occur in rugby union. Nevertheless, whatever the reasons for the levels of coherence found in this study, it highlights an intervention that is capable of developing pre-elite referees into line with the elite.

Finally, it is interesting to note, that there is also anecdotal evidence from advisors, coaches and players suggesting that such improvements have transferred to their application of law 15 on the field of play. This may not only be as a result of this test per se, but as a corollary of such work, and the increased interest and discussions that have subsequently taken place.

Conclusions

Refereeing performance is a crucial component of team sports and there is not enough DM training outside of match experience. This study presents a training program that helps pre-elite referees improve and develop experience that might

otherwise take some years to acquire. Such referee perspective video training has provided a tool to identify and overcome problems and bring developing referees into line with their top-flight peers.
Additionally, it is a valuable tool in developing and reinforcing interpretations so that when international governing bodies give new directives on interpretations mid-way through the season, as has recently occurred in elite English soccer as well as rugby union, referees can quickly adapt to the new guidelines.

References

Abernethy, B., & Russell, D.G. (1987). The relationship between expertise and visual search strategy in a racquet sport. *Human Movement Science*, **6**, 283-319.

Ackford, P. (2003, March 16). Ring of confidence from the whistle blowers. *The Sunday Telegraph*, p. 11.

Anshel, M.H. (1995). Development of a rating scale for determining competence in basketball referees: Implications for sport psychology. *The Sport Psychologist*, **9**, 4-28.

Anshel, M.H., & Webb, P. (1991). Defining competence for effective refereeing. *Sports Coach*, July-September, 33-37.

Balke, W.M., Hammond, K.R., & Meyer, G.D. (1973). An alternate approach to labor-management negotiations. *Administrative Science Quarterly*, **18**, 311-327.

Beach, L. R., & Lipshitz, R. (1993). Why classical decision theory is an inappropriate standard for evaluating and aiding most human decision making. In G.A. Klein, J. Orasanu, R. Calderwood & C.E. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 21-35). Norwood, NJ: Ablex Publishing Corporation.

Brannick, M.T., & Prince, C. (1997). An overview of team performance measurement. In M. T. Brannick, E. Salas, & C. Prince (Eds.), *Team performance assessment and measurement: Theory, methods, and applications* (pp. 3-16). Hillsdale, NJ: Lawrence Erlbaum.

Bunting, N.J. (1999) *Allied Dunbar premiership guidance for referees, players and coaches on the application of law* (RFU Tech. Rep. from the conference on the game). **\bb** Castlecroft, Wolverhampton.

Calderwood, R., Crandall, B.W., & Klein, G.A. (1987). *Protocol analysis of expert/novice command decision making during simulated fire ground incidents*. *Army Research Institute Publication*, July, 56.

Cannon-Bowers, J.A., Salas, E., & Converse, S.A. (1990). Cognitive psychology and team training: Training shared mental models of complex systems. *Human Factors Society Bulletin*, **33**(12), 1-4.

Cannon-Bowers, J.A., Salas, E., & Pruitt, J.S. (1996). Establishing the boundaries of a paradigm for decision making research. *Human Factors*, **38**(2), 193-205.

Clark-Carter, D. (2001). *Doing quantitative psychological research: From design to report*. East Sussex, UK: Psychology Press.

Cohen, J. (1988). *Statistical power analyses for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cohen, M.S., Freeman, J.T., & Thompson, B.B. (1997). Training the naturalistic decision maker. In C.E. Zsambok & G.A. Klein (Eds.), *Naturalistic decision making* (pp. 257-268). Hillsdale, NJ: Lawrence Erlbaum.

Cooke, N.J., Salas, E., Cannon-Bowers, J.A., & Stout, R.J. (2000). Measuring team knowledge. *Human Factors*, **42**(1) 151-173.

\bb De Keyser, V., Nyssen, A-S. (2001) The management of temporal constraints in naturalistic decision making: The case of anesthesia. In E. Salas & G. Klein (Eds.), *Linking*

expertise and naturalistic decision making (pp.171-188). Hillsdale, NJ: Lawrence Erlbaum.

Druckman, D., & Bjork, R.A. (1991**\bb**). Modeling expertise. In D. Druckman & R. A. Bjork (Eds.), *In the mind's eye: enhancing human performance* (pp. 57-79). Washington, DC: National Academy Press.

Edwards, W., & Newman, J.R. (1982). *Multiattribute evaluation*. Beverly Hills, CA: Sage.

Endsley, M. R. (1997). The role of situation assessment in naturalistic decision making. In C.E. Zsambok & G.A. Klein (Eds.), *Naturalistic decision making* (pp. 269-283). Hillsdale, NJ: Lawrence Erlbaum.

\bb Ericsson, K.A. (1996). The road to excellence: *The acquisition of expert performance in the arts and sciences, sports and games*. Hillsdale, NJ: Lawrence Erlbaum.

Eylon, B.S., & Reif, F. (1984). Effects of knowledge organization on task performance. *Cognition and Instruction*, **1**, 5-44.

FIBA – Fédération Internationale de Basketball Federation (2004). *Official basketball rules 2004: Referees' manual for two-person officiating*. Paris, France: FIBA.

Garcia, A. (2003). *The professional boxing referee: Evaluating the professional boxing referee - Part One*. Retrieved December 11, 2003, from http://www.secondsout.com/ringside/goodman_45692.asp

Gentner, D., & Stevens, A.L. (1983). *Mental models*. Hillsdale, NJ: Lawrence Erlbaum.

Glaser, R. (1984). Education and thinking: The role of knowledge. *American Psychologist*, **39**, 93-104.

Helsen, W.F., Hodges, N.J., Van Winckel, J., & Starkes, J.L. (2000). The roles of talent, physical precocity and practice in the development of soccer expertise. *Journal of Sport Sciences*, **18**, 727-736.

Helsen, W.F., Starkes, J.L., & Hodges, N.J. (1998). Team sports and the theory of deliberate practice. *Journal of Sport and Exercise Psychology*, **20**, 12-34.

Helsen, W., & Pauwels, J.M. (1993). The relationship between expertise and visual information processing in sport. In J.J. Starkes & F. Allard (Eds.), *Cognitive issues in motor expertise* (pp. 109-134). Amsterdam, Netherlands: Elsevier Science Publishers B.V.

iRB - international Rugby Board (2003). The laws of the game of rugby union. (2003 ed.). Retrieved December 11, 2003, from http://www.irb.com/laws_regs/laws/index.cfm

Johnston, J.H., Poirier, J., & Smith-Jentsch, K.A. (1998). Decision making under stress: Creating a research methodology. In J.A. Cannon-Bowers & E. Salas (Eds.), *Decision making under stress: Implications for individual and team training* (pp. 39-59). Washington, DC: American Psychological Association.

Jones, M.V., Paull, G.C., & Erskine, J. (2002). The impact of a team's aggressive reputation on the decisions of association football referees. *Journal of Sports Sciences*, **20**, 991-1000.

\bb Klein, G. (1997a). Naturalistic decision making: Where are we going? In C.E. Zsambok & G.A. Klein (Eds.), *Naturalistic decision making* (pp. 383-397). Hillsdale, NJ: Lawrence Erlbaum.

\bb Klein, G. (1997b). An overview of naturalistic decision making applications. In C.E. Zsambok & G.A. Klein (Eds.), *Naturalistic decision making* (pp. 49-59). Hillsdale, NJ: Lawrence Erlbaum.

Kraiger, K., & Wenzel, L. (1997). Conceptual development and empirical evaluation of measures of shared mental models as indicators of team effectiveness. In M.T. Brannick, E. Salas, & C. Prince (Eds.), *Team performance assessment and measurement: Theory, methods, and applications* (pp. 63-84). Hillsdale, NJ: Lawrence Erlbaum.

Langan-Fox, J., Code, S. & Langfield-Smith, K. (2000). Team mental models: Techniques, methods and analytic approaches. *Human Factors*, **42**, 242-271.

Lipshitz, R. (1993). Decision making as argument-driven action. In G.A. Klein, J. Orasanu, R. Calderwood, & C.E. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 172-181). Norwood, NJ: Ablex Publishing Corporation.

Mascarenhas, D.R.D., Collins, D., & Mortimer, P. (2002a). The art of reason versus the exactness of science in elite refereeing: Comments on Plessner and Betsch (2001). *Journal of Sport and Exercise Psychology*, **24**, 328-333.

Mascarenhas, D.R.D., Collins, D., & Mortimer, P. (2002b). The 'four cornerstones' model, *Referee*, 27(4) 38.

Mascarenhas, D.R.D., Collins, D., & Mortimer, P. (in press). Assessing the accuracy and coherence of decision making in rugby union officials, *Journal of Sport Behavior*.

McCullagh, P. (1986). Model status as a determinant of observational learning and performance. *Journal of Sport Psychology*, **8**, 319-331.

McLennan, J., & Omodei, M.M. (1996). The role of prepriming in recognition-primed decision making. *Perceptual and Motor Skills*, **82**, 1059-1069.

Means, B., Salas, E., Crandall, B., & Jacobs, T.O. (1993). Training decision makers for the real world. In G.A. Klein, J. Orasanu, R. Calderwood, & C.E. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 306-326). Norwood, NJ: Ablex Publishing Corporation.

Omodei, M.M., McLennan, J., & Whitford, P. (1998**\bb**). Using a head-mounted video camera and two-stage replay to enhance orienteering performance. *International Journal of Sport Psychology*, **29**, 115-131.

Orasanu, J.M. (1990). Shared mental models & crew decision making. *Technical report for Princeton University*, **46**, 19.

Orasanu, J., & Connolly, T. (1993). The reinvention of decision making. In G.A. Klein, J. Orasanu, R. Calderwood, & C.E. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 3-20). Norwood, NJ: Ablex Publishing Corporation.

Plessner, H., & Betsch, T. (2001**\bb**). Sequential effects in important referee decisions: The case of penalties in soccer. *Journal of Sport and Exercise Psychology*, **23**, 254-259.

Rasmussen, J. (1985). The role of hierarchical knowledge representations in decision making and system management. *IEEE Transactions on Systems, Man and Cybernetics. SMC-*15(2) 234-243.

Rouse, W.B., & Morris, N.M. (1986). On looking into the black box: Prospects and limits in the search for mental-models. *Psychological Bulletin*, **100**, 349-363.

Salas, E., Cannon-Bowers, J.A., & Johnston, J.H. (1997). How can you turn a team of experts into an expert team?: Emerging training strategies. In C.E. Zsambok & G.A. Klein (Eds.), *Naturalistic decision making* (pp. 359-370). Hillsdale, NJ: Lawrence Erlbaum.

Simon, H.A., & Chase, W.G. (1973). Skill in chess. *American Scientist*, **61**, 393-403.

Starkes, J. L., & Lindley, S. (1994). Can we hasten expertise by video simulations? *Quest*, **46**, 211-222.

Starkes, J. (2000). The road to expertise: Is practice the only determinant? *International Journal of Sport Psychology*, **31**, 431-451.

Ste-Marie, D. (1999). Expertise in sport judges and referees: Circumventing information-processing limitations. In J.L. Starkes & K.A. Ericsson (Eds.), *Expert performance in sports: Advances in research on sport expertise* (pp.169-189). Champaign, IL: Human Kinetics.

Stokes, A.F., Kemper, K., & Kite, K. (1997). Aeronautical decision making, cue recognition, and expertise under time pressure. In C.E. Zsambok & G.A. Klein (Eds.), *Naturalistic decision making* (pp. 183-196). Hillsdale, NJ: Lawrence Erlbaum.

Stout, R.J. Cannon-Bowers, J.A., & Salas, E. (1996). The role of shared mental models in developing team situational awareness: Implications for training. *Training Research Journal*, **2**, 85-116.

Webber, S.S., Chen, G., Payne, S.C., Marsh, S.M., & Zaccaro, S.J. (2000). Enhancing team mental model measurement with performance appraisal practices. *Organizational Research Methods*, **3**(October), 307-322.

Williams, M., & Davids, K. (1995). Declarative knowledge in sport: A by-product of experience or a characteristic of expertise. *Journal of Sport and Exercise Psychology*, **17**, 259-273.

Williams, A.M. Davids, K., Burwitz, L., & Williams, J.G. (1994). Visual search strategies of experienced and inexperienced soccer players. *Research Quarterly for Exercise and Sport*, **65**, 127-135.

Williams, A.M., Davids, K., & Williams, J.G. (1999). *Visual perception & action in sport*. London: E and FN Spon.

Acknowledgments

We gratefully acknowledge the financial support from the Rugby Football Union and the contributions made by Nick Bunting at the RFU Referees Center of Excellence and Colin High, Ed Morrison and the full time referees at the RFU Elite Referee Unit. We would like to dedicate this manuscript to the memory of Bob Morris.

Manuscript submitted: February 29, 2004
Revision received: December 3, 2004