NOTE TO USERS

AN EXPERT PERFORMANCE AND DELIBERATE PRACTICE ANALYSIS OF OPEN SPORT REFEREES

By

CLARE MACMAHON, B.A., M.A.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree

Philosophae Doctorate

McMaster University

EXPERTISE IN OPEN SPORT REFEREEES

PHILOSOPHAE DOCTORATE (2004) McMaster University

(Kinesiology) Hamilton, Ontario

TITLE: An Expert Performance and Deliberate Practice Analysis of Open Sport Referees

AUTHOR: Clare MacMahon, B.A. (McGill),
M.A. (University of Ottawa)

SUPERVISOR: Professor J.L. Starkes

NUMBER OF PAGES: x, 167

ABSTRACT

The deliberate practice theory and the expert performance approach were used to explore expertise in open-sport referees in four related studies. Study 1 used questionnaires focused on deliberate practice training activities. Findings showed that world-class level soccer referees specialize early in referee training activities, with an emphasis on physical training to the virtual exclusion of perceptual-cognitive training. Similar to studies with athletes, activities rated high in effort were also rated high in enjoyment. Distinct from studies with athletes, however, highly effortful activities were not always rated high in relevance. Study 2 combined biographical experience and practice data with performance on video-based laboratory tasks. Findings indicate that part of referee skill involves anticipation of movements, but not necessarily superior strategic decision-making. Studies 3 and 4 moved to the sport of basketball with more detailed tests of video-based laboratory tasks. Study 3 findings indicate that referee skill in video-based infraction detection and decision-making tasks may be influenced by features of the video display (e.g., speed and camera angle) and the inclusion of secondary, visual processing tasks. The visual processing style used to interpret and classify video clip actions may also interact with expertise level. Finally, study 4 indicates that the choice of game clips as well as their sequencing relative to the signals contained within needs to be examined closely when creating testing and training tools. As one of the first focused programmes of research on expertise in open-sport referees, this work highlights a number of areas of future research.

ACKNOWLEDGEMENTS

The acknowledgements may be the most difficult section of the thesis to write. It is always my favourite section to read, because it forces the writer to reflect, and provides a snapshot of time with insight into the person, the advisor, and the programme - something rare in research. These lofty expectations are what make it so difficult. Regardless, my main goal is to acknowledge that I did not do this work alone, and that I am grateful for the guidance and support that I received throughout.

First and foremost I would like to acknowledge Jan Starkes, my advisor. I am so proud to have been your PhD student. You gave me incredible exposure, with a trip to Belgium and beyond, always quietly aware of what additional training would benefit me. Through all the times I stalked you outside your office, and dropped in for "quick questions", you always made time for me (even when you had pneumonia), with a gentle touch that never stifled. You served as a consultant for my work, rather than driving the train that I was hitched to. This helped me stay excited, and I think it has built a very strong foundation for future work.

I was lucky to have a patient and insightful supervisory committee, made up of Tim Lee and Jim Lyons. I always felt that you were there to provide guidance and to challenge me, and not to beat me down. You've made me see things in different ways, and consider lots of avenues for future work. I'm flattered by your interest in my thesis, and feel privileged to have worked with you both.

No one has tested me more than Werner Helsen. I learned a great deal about collaboration, and productivity; lessons I will take with me for the rest of my career. Thank you for your generosity while I visited, and thanks to the lab that welcomed me with open arms.

I have to acknowledge the extremely positive influence of Digby Elliott. I admire the way that you run a lab, your overwhelming generosity, and the relationships that you build with your students. You create an exceptional working environment with a family dynamic. I have wise "older brothers" in Tim Welsh and Luc Tremblay from whom I seek advice, and younger siblings who I play and eat with (especially Steve: there were lots of evenings I kept working because I knew you were working too, and I'd have someone to walk home with!).

From this family, I also have to acknowledge Jae Patterson and Brad Young, who were "born in to the programme" at the same time as me. It was always good to have people to relate to as we did classes, comps and thesis work together. We all became more independent, but I'll still miss the craziness, the tremendous amount of laughing we did, and the support we created for each other. John Cullen also deserves special mention for being my safe haven and reassuring me that I was on the right path. It was invaluable to have someone I could consult about my work and especially stats. You were always so generous with your time, and I truly appreciate that. Thanks, as well, for being my ride to Tammy's on Fridays! I'll miss those breakfasts!

And now to the women: I am grateful for the support of Amy Latimer and Jen Salter, and the close friend and role model I see in Sian Beilock. Outside of academia, I was much stronger knowing that I had Robin and Ulka, my powerful friends in Montreal, who are still trying to figure out what "comps" means, and Robin in B.C., who inspires me with her strength. I loved knowing I could tell you all anything, and you would listen and not judge.

My family has been extremely supportive. I enjoyed escapes to Toronto for an occasional "MacMahon sisters night out" (yikes) and relaxing trips to see Mom, Dad and Brian in Montreal. Thank you all for being my safety net, and Dad, for telling me you're proud of me.

On the technical side of things, I am grateful for John Moroz's help with equipment, and Guy Cipriani's very generous help in recruiting referees. I would be nowhere without the willingness of this group, which was there because of Guy's support. I am amazed by the passion of the Hamilton board. It was truly a pleasure to work with such a great group of people, who gave me insight into my work.

Deanna Goral was always my first stop on hunting trips for "The Starkes." I enjoyed all of our chats, and the support that she gave me through this whole process, from parking passes, to helping me photocopy this document! Finally, a big thank you to Mary Cleland, who was always on top of forms, funding and anything else I needed. Mary is a wonderful person, and I'll miss seeing her everyday.

My first week in the PhD programme at MAC, I got a fortune cookie from The Wokery. The fortune said: "You will make a good lawyer". I immediately went and told Mary that I thought I should change programmes. I'm glad she laughed, and I'm glad that I stayed in the programme, because I think it was just the perfect fit, and I can't imagine having done anything else!

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

PREAMBLE

This dissertation explores expertise in sport referees through two main studies, each with a follow-up. The first study (study 1) applied the deliberate practice theory by using questionnaires with 26 world-class soccer referees. Referees indicated time spent in activities for three target years, and also rated these activities on the dimensions of relevance to improving performance, effort, enjoyment, and concentration. A follow-up study (study 2) combined biographical data from the first study with video-based laboratory tests of soccer skill. This study compared international soccer referees with youth academy players.

Two additional studies used the expert performance approach in further tests of laboratory tasks. This section of the thesis moves to the sport of basketball. This change in sports was made in order to replicate and extend a previous study (Allard, Parker, Deakin & Rodgers, 1993) and thus to make use of previously constructed testing materials. In study 3, basketball, players, coaches and referees were tested in video tasks designed to match their respective role-related demands. A follow-up study (study 4) focused on the lack of clear referee superiority in the referee task. Specifically, study 4 used two groups of referees to explore the potential influence of task and design features in the referee task.

This work is organized in the following manner: A general introduction provides a context for the work. This section presents an overall framework for the dissertation in the Starkes, Cullen and MacMahon (2004) model that describes the development of perceptual-motor expertise in sport. Sections introducing the expert performance approach, an overview of findings related to expert perceptual-cognitive

skill in sport, and a general description of the theory of deliberate practice follow. A final section provides an overview of studies conducted as part of the dissertation.

Three papers follow the introduction. Paper 1 presents study 1, an examination of deliberate practice activities in top-class soccer referees. Paper 2 presents study 2, a combined analysis of biographical data and video-based laboratory task performance in soccer players and referees. Finally, paper 3 contains studies 3 and 4, which explore expertise in basketball referees using role-based video tasks. The final section of the thesis is the general conclusion, used to draw the work together. This section presents a general summary of the findings, and refers back to the Starkes et al. (2004) model for reflection.

GENERAL INTRODUCTION

Recent work in the area of sport acquisition research has resulted in a descriptive model of the learning process involved in sport skills (Starkes, Cullen & MacMahon, 2004). The model can be used to describe skill acquisition across the lifespan as children learn a skill, and adolescents and adults refine its performance to expert levels, which are then retained over a number of years. The model is presented in figure 1, at the end of this section. At four anchor points in development, performance is characterized by the model as: 1) the **acquisition** of basic knowledge and skills, where learners "get the idea" of the skill, 2) the **condensation and elaboration** of knowledge and skills, where multiple actions are grouped and additional movement options are created 3) **routine expertise**, in which experience is used to predict and plan events, and 4) **transcendent expertise**, when performers have become innovative and transcend previous performance levels (e.g., Tiger Woods, Wayne Gretzky).

A key feature of the model is that it not only describes performance but also offers explanation. Skill improvements are seen to result from changes in the way an athlete thinks and moves in response to their environment (e.g., a cycling race course) and stimuli (e.g., a soccer player's "fake"). These two major areas of development (thinking and moving) are represented by two streams or channels of skill within the performer. Cognitive or information-processing changes take place in the **perceptual-cognitive stream**, and movement changes in the **perceptual-motor stream**. As skill is acquired, the output from each stream is combined to produce the final performance.

The Starkes et al. (2004) model was created through reflection on the existing literature in expert sport performance (sport expertise). Within sport research, however, there is a focus on athletes and coaches, with little emphasis on the role of officials. Notwithstanding, the task demands in officiating, particularly in open-sports, are both unique and complex, making this a ripe population for study. For example, a soccer referee is not only required to know the game rules and their implementation, but must also keep up with the play, and position him/herself appropriately in order to view the action. An added difficulty of this task is that decisions must be made quickly (under time pressure) in response to moving (dynamic) stimuli (e.g., players, the ball). These dynamic situations often result in decision-making based on incomplete information (Plessner & Betch, 2002).

While the Starkes et al. (2004) model reveals the lack of existing knowledge related to open-sport referees, it also provides a framework to direct studies into this population. Indeed, the model highlights a number of perspectives and paradigms that can be applied to the study of referees. The following sections will review three relevant areas of study. These three, interlinked areas of the sport literature were used to investigate two major topics in sport refereeing: 1) referee training, and 2) referee skills.

The Expert Performance Approach

While researchers have always been fascinated by superior sports performance, it was not until the application of a structured approach that knowledge in this area began to boom, and focused programmes of study appeared (see Starkes & Ericsson, 2003, for a review of critical findings related to expert performance in sport). This influential approach is the expert performance approach, discussed in

great detail by Ericsson and Smith (1991).

As Ericsson and Smith (1991) describe, the expert performance approach has three phases: the elicitation of expert performance in the laboratory, the identification of mediating mechanisms responsible for this superior performance, and the identification of methods used to acquire the mediating mechanisms. The authors further discuss key guidelines for each phase of this research approach.

For sport, one of the most essential and challenging steps of this process is the creation of laboratory tasks that adequately represent real-world performance. It is often unfeasible for researchers to replicate the entire task (e.g., playing a game of baseball) under controlled conditions. Thus, the choice of smaller components of performance, representative tasks, becomes an important step. In order to determine whether the task is representative, Ericsson and Smith (1991) advise examining the *stability* of performance. For instance, dramatic learning and practice improvements in performance of a laboratory task may indicate that it is overwhelmingly novel to the performer, and thus fails to replicate real-life sport skills. On the other hand, small improvements may indicate that performers are simply "warming up", improvements often typical in real-world task performances. In order to distinguish whether improvements are due to warm-up or learning, Ericsson and Smith advise measuring performance in both the laboratory and the real-world. In this manner, improvement rates can be directly compared. For instance, performance in a baseball batting laboratory task can be compared to performance during real-world batting practice to determine if this task is indeed representative of baseball hitting skill.

In the second step of the expert performance approach, the focus shifts to discovery of mechanisms responsible for superior skills and performance. Ericsson

and Smith (1991) first describe the use of in-depth analysis of the expert performance elicited at step one as a means of generating hypotheses related to potential mediating mechanisms. This may result in the repetition of step one, with modifications of the laboratory task in attempts to more clearly elicit the processing related to the hypothesized mechanism. For instance, researchers may hypothesize that a critical aspect of batting performance is the batter's point of gaze during the pitcher's wind up. To explore this, point of gaze may be measured, and visual information manipulated. Similarly, if decision-making and conscious processing are targeted, a think-aloud protocol may be incorporated to gain access to this processing. In this task-analysis component of step two, researchers attempt to identify the most critical aspects of superior performance in the representative task.

This in-depth analysis of performance can then be used to more clearly direct the activities during step two's exploration of potential mediating mechanisms. Two typical paradigms used at this step are 1) expert-novice comparisons, and 2) single subject case studies of experts. Expert-novice comparisons, in which differences in performance are used to identify mediating knowledge and processes, are by far the more widely used approach in sport.

The final step of the expert performance approach focuses on understanding the acquisition of mechanisms identified in step 2. For example, if expert baseball batters are shown to differ in their decision-making abilities, the focus at this stage of inquiry is on how this skill is developed or acquired. While there is an emphasis at this step on practice, Ericsson and Smith (1991) also emphasize what they describe as learning mechanisms, which are seen as the result of practice. A key example is the work of Chase and Simon (1973), which argues that encoding, storage and retrieval of

knowledge gained through practice and experience are responsible for expertise in chess. Indeed, this perspective has been adopted in sport with measurement of skills such as processing, storage, recognition and recall of game related information.

While Ericsson and Smith (1991) present three distinct steps or phases to the expert performance approach, research activities may target more than one step simultaneously. For instance, identification of a representative task may take place through comparison of expert-novice performance, and thus the simultaneous exploration of possible mediating mechanisms. This approach has often been used in sport. Indeed, as mentioned, the application of the expert performance approach in general has been fruitful in sport research. A summary of these findings is the focus of the next section.

Findings Related to Expert Perception and Cognition in Sport

As mentioned in the preceding, Chase and Simon (1973) account for expertise in chess through superior information processing skills. Previous to its application in sport, the expertise approach, and the study of expertise in general, addressed primarily cognitive tasks such as chess (e.g., deGroot, 1978; Chase & Simon, 1973), musical performance (e.g., Ericsson, Krampe, & Tesch-Römer, 1993), and physics problem solving (e.g., Newell & Simon, 1972). These tasks all involve very little gross movement. In contrast, movement is a central component in sport performance. Moreover, athletes often act under dynamic environments with time and space restrictions. Thus, while sport expertise research maintained a cognitive approach to accounting for expert performance, the specific *type* of cognitive processing was first addressed.

The findings in Starkes and Deakin (1984) have had a large impact on sport expertise research. Specifically, this study identified different *types* of processing, distinguishing between hardware and software skills. Hardware skills are those that are inherent to the "system", as it were, and are physically bounded. Examples of hardware skills are reaction time, depth perception, and peripheral vision. These skills cannot be improved very much with training. In contrast, software skills are those used to process information gained through the hardware system, and can be improved with training and practice. The most important finding in Starkes and Deakin (1984) is that software skills specific to the domain, and not hardware skills, differentiate experts in sport. This finding has since guided work on sport expertise.

Perhaps one of the best illustrations of the software advantage in sport experts comes from Starkes (1987). Using a multitude of measures, Starkes compared expert and novice field hockey players in both hardware skills (e.g., simple reaction time, coincidence anticipation) and software skills (e.g., shoot/dribble/dodge decision making speed, accuracy of shot prediction). Starkes showed that expertise was predicted by performance on two software tasks: recall of game structured information, and accuracy of shot prediction. This method of comparing the same performers in a large number of skills is rare. Indeed, the tasks used to examine expert-novice differences in the research vary a great deal, with experts showing advantages in software skills as seemingly simple as detection, and as complex as the use of tactics and strategy. For example, Allard and Starkes (1980) found that volleyball players were able to detect the presence of a volleyball much faster than non-players in rapidly displayed game slides. Similarly, expert gymnastics judges are superior in detecting form errors when compared to novices (Ste-Marie & Lee, 1991).

On the other hand, McPherson and colleagues' detailed analyses of verbal reports during tennis game play show that experts use elaborate stores of information based on both the current game and previously gained knowledge, resulting in sophisticated game plans. Novices show much less depth of analysis, resulting in less sophisticated plans (McPherson, 1999a, 1999b, 2000).

Perhaps the most popularly examined software skills used in sport research are recall and recognition of domain specific information. This may be due to the relatively robust nature of findings showing an expert advantage in these skills across a number of studies and sports. Indeed, in a review of the literature addressing expert performance in sport and dance, Starkes, Helsen and Jack (2001) indicate that "...the best athletes are able to recognize, recall, and retain more information about plays or structured game information than less skilled athletes." (p. 182).

While research in expert perception and cognition in sport emphasizes the identification of domain specific skills, there has also been a great deal of study focused on the third step of the expert performance approach: the acquisition of mechanisms responsible for expert performance. A component of this acquisition process is practice, which is discussed in the next section.

The Theory of Deliberate Practice

The exposure, experience, and practice necessary for the development of domain-specific skills have been alluded to in the previous sections. In sport expertise research, this is certainly a topic that has received a great deal of attention. An extremely useful tool in examining the role of practice in expert performance is the theory of deliberate practice.

The nature versus nurture debate has long raged within psychology. Recently, the case for a nurture explanation of superior performance, i.e. that great performers are made and not born, has gained theoretical and empirical support. Ericsson, Krampe and Tesch-Römer (1993) advanced the theory of deliberate practice in the acquisition of expert performance. The premise of this theory is that expertise is gained through the deliberate practice activities of the performer. As per Ericsson et al., to qualify as *deliberate practice*, activities must be monitored, with immediate and informative feedback, effortful, and not inherently enjoyable. Moreover, the motivation to engage in deliberate practice must be to improve performance, and not, for example, for financial or social rewards.

Findings have shown a monotonic relationship between amount of deliberate practice and performance levels, such that a minimum of 10,000 hours or 10 years of deliberate practice is required to attain expert levels of performance. This has been shown in music as well as various sports (e.g., Ericsson et al, 1993; Helsen et al, 1998; Hodges & Starkes, 1996; Starkes, Deakin, Allard, Hodges, & Hayes, 1996).

There are two notable points in the evolution of the deliberate practice theory that have come about through its application to sport. First, physical effort became distinguished from mental effort. Where activities such as chess and violin playing involve minimal physical performance, sport skills are achieved primarily through the use of larger muscle groups in both gross and fine movements. Thus physical effort gains a larger focus within sport skills, in contrast to activities such as chess and violin playing. Indeed, in wrestling and figure skating, participants differentiate activities according to amount of concentration required versus amount of effort required (Helsen, Starkes, & Hodges, 1998).

A second important development in the theory of deliberate practice in sport was the move to look at different types of sports. Expertise was first investigated in individual sports like wrestling and figure skating, and then brought into the study of team sports (Helsen et al., 1998). This move provided a more elaborated view of practice activities, differentiating individual and team as two different forms of practice. Thus, in examining the specific activities (and not simply the amount of time spent) of international, national, and provincial level soccer players, Helsen et al. examined individual practice, team practice, sport-related activities and everyday life activities. Each activity was rated by participants on the dimensions of relevance to improving soccer performance, effort, enjoyment, and concentration.

The work presented in this thesis draws on the developments and findings reviewed in the above sections. Specifically, the overall purpose of this work was to explore expertise in open-sport referees. Given that there is a great deal of research dealing with athletes, and very little with referees, research on refereeing expertise represents a somewhat 'new area' of study, with the advantage of a very strong foundation. With this in mind, this research concentrated on two main areas. The first area studied was practice activities, with the theory of deliberate practice used as a framework. The second area of study concentrated on the skills and characteristics of referees, framed by the first and second steps of the expert performance approach. The studies are reviewed briefly below.

Overview of Studies

The first paper in this series reports a two-part study of the deliberate practice activities of elite soccer referees. In part 1, retrospective recall was used to report on weekly amounts of time spent in the various activities during the first year of formal

refereeing, the year before the implementation of formal training programs (1998) and 2003 (the 'current year' for the data collection period). For these same target years, participants also indicated their perceptions of how relevant the activities were at that time to improving performance. These two sources of data provided a picture of the transitions in training activities over time, as well as the changes in perceptions of the activities. This is the first time that training in referees has been examined, and the first time that retrospective reports of relevance have been collected in any population.

The second paper provides a theoretical bridge between practice activities and skills. For a sub-group of referees from the first study, the biographical data collected during study one were combined with performance data for two laboratory tasks. As well, performance in these laboratory tasks was compared to that of a group of youth academy players.

The final paper in this dissertation switches to the sport of basketball in order to replicate and extend a previous study. Again, the focus of this last paper is on capturing and explaining skill in the laboratory. A less traditional approach is used in comparing skills between coaches, players and referees. In the first study, tasks were designed to replicate on-court demands of the three roles. The follow-up study focused more specifically on the design of the referee task and possible factors influencing performance.

While the results of the four studies contained in the papers that follow advance our knowledge of expert refereeing, they also provide clear avenues for future research.

References

Allard, F. & Starkes, J. L. (1980). Perception in sport: Volleyball. Journal of Sport Psychology, 2, 22-33.

Chase, W.G., & Simon, H.A. (1973). Perception in chess. Cognitive Psychology, 4, 55-81.

Ericsson, K. A., Krampe, R. T., & Tesch- Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. Psychological Review, 100, 363-406.

deGroot, A. (1978). Thought and choice in chess. The Hague, Netherlands: Mouton.

Ericsson, K. A., & Smith, J. (1991). Prospects and limits in the empirical study of expertise: An introduction. In K. A. Ericsson and J. Smith (Eds.), Toward a general theory of expertise: Prospects and limits (pp. 1-38). Cambridge: Cambridge University Press.

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.

Hodges, N.J. & Starkes, J.L. (1996). Wrestling with the nature of expertise: A sport specific test of Ericsson, Krampe, and Tesch-Römer's (1993) theory of deliberate practice. International Journal of Sport Psychology, 27, 400-424.

McPherson, S.L. (1999a). Expert-novice differences in performance skills and problem representations of youth and adults during tennis competition. Research Quarterly for Exercise and Sport, 70, 233-251.

McPherson, S.L. (1999b). Tactical differences in problem representations and solutions in collegiate varsity and beginner women tennis players. Research Quarterly for Exercise and Sport, 70, 369-384.

McPherson, S.L. (2000). Expert-novices differences in planning strategies during collegiate singles tennis competition. Journal of Sport and Exercise Psychology, 22, 39-62.

Newell, A., & Simon, H.A. (1972). Human problem solving. Englewood Cliffs, NJ: Prentice-Hall.

Starkes, J.L. (1987). Skill in field hockey: The nature of the cognitive advantage. Journal of Sport Psychology, 9, 146-160.

Starkes, J.L., Deakin J., Allard, F., Hodges, N.J., & Hayes, A. (1996). Deliberate practice in sports: What is it anyway? In K.A. Ericsson (Ed.). The road to excellence: The acquisition of expert performance in the arts and sciences, sports and games. (pp. 81-106). Hillsdale, NJ: Erlbaum.

Starkes, J.L., Cullen, J.D., & MacMahon, C. (2004). A model of the acquisition and retention of expert perceptual-motor performance. In A.M. Williams, N.J. Hodges, M.A Scott, & M.L.J. Court (Eds.) Skill acquisition in sport: Research, theory and practice. (pp. 259-281). London: Routledge.

Starkes, J.L. & Ericsson, K.A. (2003). Expert performance in sports: Advances in research on sport expertise, Champaign, IL: Human Kinetics.

Starkes, J.L., Helsen W., & Jack, R. In R.N. Singer, H.A. Hausenblas, & C.M. Janelle (Eds.), Handbook of sport psychology (2nd ed.) (pp.174-201). New York: John Wiley & Sons.

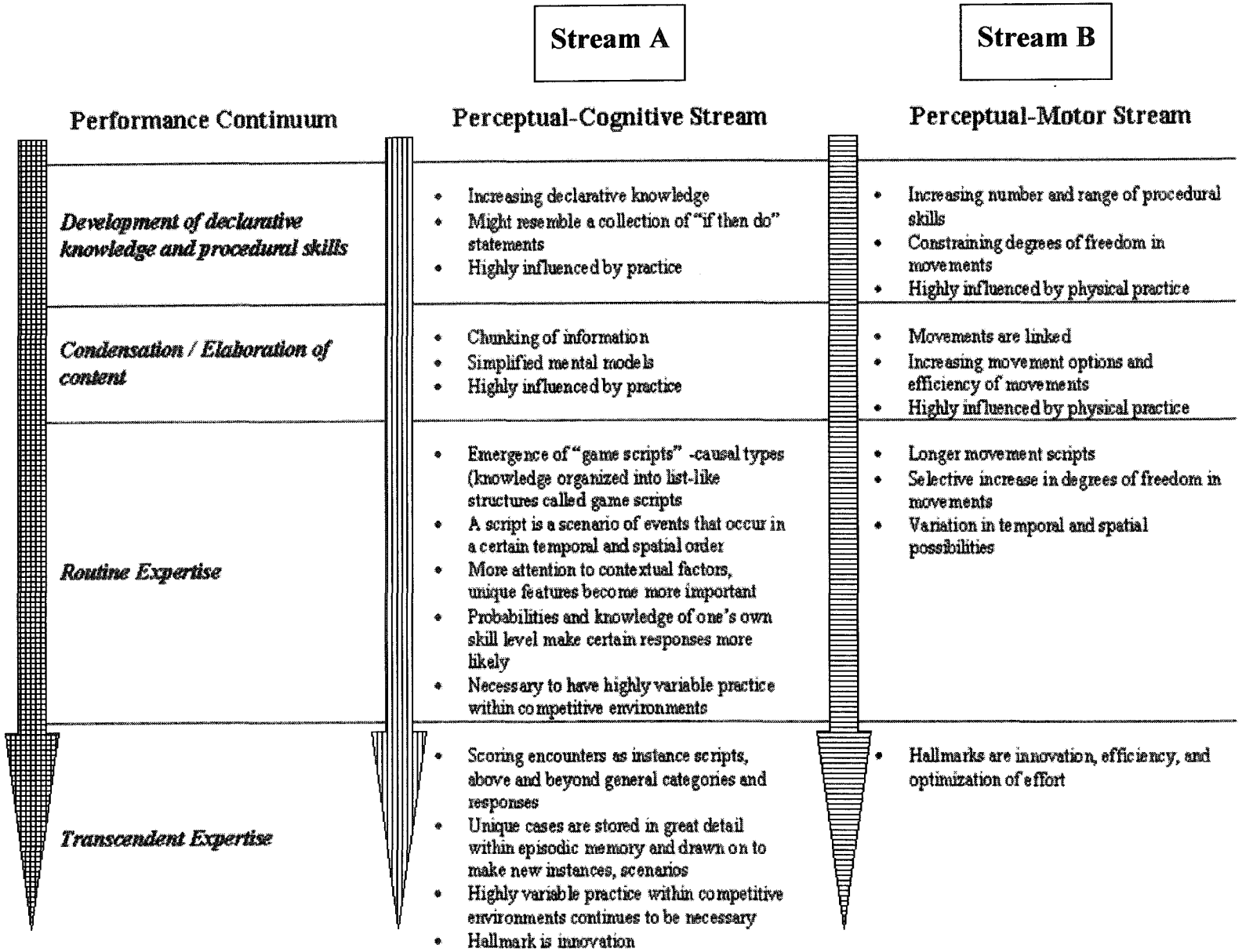
Ste-Marie, D. M., & Lee, T. D. (1991). Prior processing effects on gymnastics judging. Journal of Experimental Psychology: Learning, Memory, and Cognition, 17, 126-136.

Starkes, J. L., & Deakin, J. (1984). Perception is sport: A cognitive approach to skilled performance. In W. F. Straub & J. M. Williams (Eds.), Cognitive sport psychology, (pp. 115-128). Lansing, NY: Sport Sciences.

Plessner, H. & Betsch, T. (2002). Refereeing in sports is supposed to be a craft, not an art: Response to Mascarenhas, Collins, and Mortimer. Journal of Sport and Exercise Psychology, 24(3), 334-337.

Figures

Figure 1. The Starkes, Cullen, & MacMahon model depicting the development of expert perceptual-motor skill. Adapted from Starkes, Cullen & MacMahon (2004).


Stream A
Stream B
Performance Continuum
Perceptual-Cognitive Stream
Perceptual-Motor Stream
Development of declarative knowledge and procedural skills
Increasing declarative knowledge
Might resemble a collection of "if then do" statements
Highly influenced by practice
Increasing number and range of procedural skills
Constraining degrees of freedom in movements
Highly influenced by physical practice
Condensation / Elaboration of content
Chunking of information
Simplified mental models
Highly influenced by practice
Movements are linked
Increasing movement options and efficiency of movements
Highly influenced by physical practice
Routine Expertise
Emergence of "game scripts" -causal types (knowledge organized into list-like structures called game scripts
A script is a scenario of events that occur in a certain temporal and spatial order
More attention to contextual factors, unique features become more important
Probabilities and knowledge of one's own skill level make certain responses more likely
Necessary to have highly variable practice within competitive environments
Longer movement scripts
Selective increase in degrees of freedom in movements
Variation in temporal and spatial possibilities
Transcendent Expertise
Scoring encounters as instance scripts, above and beyond general categories and responses
Unique cases are stored in great detail within episodic memory and drawn on to make new instances, scenarios
Highly variable practice within competitive environments continues to be necessary
Hallmark is innovation
Hallmarks are innovation, efficiency, and optimization of effort

PAPER 1

This paper has been submitted for publication in the Journal of Sport and Exercise Psychology. While I was helped in data collection of this research project, I was the major contributor on the experimental design, data analysis, and write-up of the manuscript.

Running Head: DELIBERATE PRACTICE IN SOCCER REFEREES

Deliberate practice in elite soccer referees

Clare MacMahon
McMaster University
Hamilton, Ontario, Canada

Werner F. Helsen
Katholieke Universiteit Leuven
Leuven, Belgium

Janet L. Starkes
McMaster University
Hamilton, Ontario, Canada

Matthew Weston
Katholieke Universiteit Leuven
Leuven, Belgium

Corresponding author:
Clare MacMahon
Dept. of Kinesiology
McMaster University
1280 Main St. West
Hamilton, Ontario
L8S 4K1
Fax: (905) 523-6011
macmahc@mcmaster.ca

Abstract

Two groups of elite soccer referees completed deliberate practice questionnaires detailing career practice and current perceptions of training activities. Referees retrospectively reported time spent in activities for three periods: their first year of formal refereeing, 1998 (before the introduction of formal training programmes), and the current year. Current perceptions of activities were also provided on the dimensions of relevance to refereeing, effort, enjoyment and concentration of the activity. Overall, this research indicates that training patterns and perceptions of elite referees parallel what has been reported elsewhere for athletes. Like elite players, elite referees engage in a large volume of training activities. The most relevant activities are those that mimic the physical demands of refereeing for speed, endurance, and agility. Moreover, refereeing experience appears to contribute to expertise, while playing and coaching experience do not. Unlike the original theory of deliberate practice, and similar to other sport research, ratings of effort were correlated with enjoyment. There was no relationship, however, between ratings of relevance of the activity to refereeing and effort devoted.

Introduction

The original deliberate practice theory proposed that expertise is gained through an accumulation of time (typically 10 years or 10,000 hours) spent in activities that are monitored, provide immediate and informative feedback, are effortful, and are not inherently enjoyable (Ericsson, Krampe & Tesch-Römer, 1993). Later investigations applying the theory to sport, however, have shown relevant activities to be enjoyable (e.g., Helsen, Starkes & Hodges, 1998) rather than un-enjoyable, and have introduced a distinction between physical effort and mental effort, or concentration (e.g., Hodges & Starkes, 1996). A further elaboration of the theory is its application to different *types* of sports, from initial investigations in individual sports like wrestling and figure skating (e.g., Hodges & Starkes, 1996), to subsequent study in team sports such as field hockey and soccer (e.g., Helsen et al., 1998; Baker, Cotê, & Abernethy, 2003). The move to acknowledge differences between team and individual sport practice provided a more elaborated view of training activities, and distinguished individual and team aspects as two different forms of practice.

While a number of studies have looked at athlete training patterns through the lens of the deliberate practice theory (e.g., Hodges & Starkes, 1996, Helsen et al., 1998), this framework has yet to be applied to sport referees. Indeed, we know very little about referee training, including content, amount, and the influence of career stage. Since referees have such an influential role in the outcome of sporting contests, it behooves us to better understand their development. To explore this population and its practice activities, we must first acknowledge the unique demands and characteristics of performance.

An obvious requirement for referees is knowledge of the laws and rules of the sport. In addition, Plessner and Betsch (2001) describe a need for "game management", or interpretive skills in the application of this knowledge. Given that the population of interest in this study is that of elite soccer referees, recent work exploring officials' demands during match performance is relevant. For instance, Helsen and Bultynck (2003) have shown that referees make an average of 200 decisions during the course of a match, at a rate of three to four decisions per minute. In addition, soccer referees must move about the performance area with the athlete. This feature creates two very important demands. Firstly, Oudejans, Verheijen, Bakker, Gerrits, Steinbrücker, and Beek (2000) have illustrated that referees' field positioning is important for making correct decisions. Secondly, and likely a byproduct of this need for adequate positioning, referees have been shown to function at an average heart rate as high as 85 to 90% of their maximum throughout a match (Helsen & Bultynck, 2003).

In spite of the large physiological and perceptual-cognitive demands of performance, referees have traditionally not engaged in coach-prescribed training, practicing instead independently and in the absence of training guidelines. This characteristic presents a significant contrast to typical athlete training which is coach-directed. Recent referee professionalization, however, and the implementation of structured training programs at the elite level provide a timely manipulation of this factor, which we were able to examine. One feature of this study, then, is its comparison between training activities and perceptions prior to and subsequent to the introduction of structured training programmes.

All in all, the unique demands of soccer referees taken together with the current lack of information on training activities make this an ideal population to study. The goals of this paper, then, are 1) to understand and test the amount and types of practice accumulated before expert levels are achieved in a referee sample, comparing these findings to what is currently known about elite athlete training, 2) to compare differences in group training versus individual training, and 3) to examine perceptions of activities on the dimensions of relevance to improving performance, effort, enjoyment, and concentration. These goals are addressed in two parts. Part 1 examines two aspects of training: the amount of time spent in different activities, and the perceptions of these activities in terms of relevance to improving refereeing performance. These data are retrospective for 3 time periods: first year of formal refereeing, 1998 (before the implementation of training programmes) and current year (2003), when the programmes were still in place. Additionally, referee relevance ratings for the current year were compared to those of the referee-coaches. Biographical data are also examined in this section for those referees on the FIFA (Fédération Internationale de Football Association) list. Part 2 of the paper assesses the *current* perceptions of the different training activities on the dimensions of relevance, effort, enjoyment and concentration.

Part 1: Volumes and Relevance of Training Activities for Three Periods

Method

Participants

Two groups of elite referees participated in this study. The first group consisted of Belgian international referees on the FIFA list (n=7). Referees on the FIFA list are called upon to officiate during events such as the World Cup, and World

Youth tournaments, representing the highest levels of soccer played in the world. The second group consisted of English Premier League referees (n=19), some of whom were also on the FIFA list (n= 7). Both groups referee in the UEFA league (European competition), and represent the highest level of referees in their respective countries. Taken together, this sample consists of 26 of the top soccer referees in the world as recognized by FIFA and UEFA. In stating this, however, it must be acknowledged that many moderating variables may influence assignment to the FIFA and UEFA lists. For instance, the level of development of soccer within a country, and the number of FIFA referees allotted to represent a given country may result in a case where a Belgian referee who is not on the FIFA list is in fact superior to a FIFA referee who is from a country where the sport is less developed (e.g., Canada).

As shown in Table 1, the referees as a group had a mean age of 40.8 years, had begun refereeing at an average age of 19.5, and thus had accumulated an average of 21.3 years of refereeing. (These data are further broken down by nationality and experience level.)

Insert Table 1 about here

Materials and Procedure.

Participants were asked to complete a retrospective questionnaire on deliberate practice. Referees were first asked for biographic data related to their experiences in soccer, with an emphasis on refereeing (e.g., age began refereeing, number of years refereeing, number of matches in national leagues, international tournaments). Following this, retrospective recall was used to collect data on time spent in various

training activities (reported in minutes per week) and their perceived relevance to improving refereeing. Relevance was rated on a 10- point scale, with 1 labelled "*low*" and 10 labelled "*high*". The three years for which these data were recorded were: 1) the first year of formal refereeing, defined as the first year that the referee was a regular part of a league schedule and a refereeing federation, 2) 1998, before implementation of the FIFA training programmes, and 3) the current year (2003), in which training programmes were still in place. For retrospective ratings of relevance, participants were instructed to specify as accurately as possible how relevant they felt the activities were *at that time.* These two types of data (time spent and ratings of relevance) were collected for 33 activities from 6 different categories. These categories were: on-field training, off-field training, therapeutic, coaching/playing, and everyday life (see Table 2 for a list of the specific activities under each of these categories).

Insert Table 2 about here

Within the on-field training category, there were 5 running activities of different intensities (e.g. recovery versus high intensity). Given that the sample of referees used in this study make regular use of Polar™ heart rate watches in training, both a percentage of maximum heart rate (% HRmax) and a rating of perceived exertion (RPE) were associated with these activities. For example, high intensity running was described as running at 85-90% of HRmax, with an RPE of 6-10 on a 10-point scale with 10 indicating maximum exertion.

In addition to time and relevance data for overall activities and categories, 13 activities were divided into group and individual training. For example, referees were asked, for the three time periods, about time spent in, and relevance of strength training performed in a group, and strength training performed alone. This allowed us to examine both overall training volumes for activities (using time spent alone and in a group for total time in that activity), as well as training alone and in a group. For relevance ratings of these 13 activities, the alone and group scores were used to calculate an average overall relevance rating for that activity. (See Table 3 for a list of activities for which alone/group training was queried.)

Insert Table 3 about here

Data Reliability

During the questionnaire collection period, referees maintained their training activities and habits. A normal practice for both groups of officials (Belgian and English) is the use of Polar™ heart rate monitors. Each referee uses a monitor to record heart rate during training sessions for the week to two weeks between group sessions. The referee-coaches then download the data files stored on the monitors after group sessions (for which HR is also recorded) to ensure that referees are training within heart rate zones appropriate for that week (given time of season, upcoming games and other considerations). The referee-coaches compile these data to indicate time spent in different heart rate zones. As well, time spent during a session, and the classification of that session are indicated (e.g., speed endurance session).

The HR data files were used for sixteen referees to check the reliability of the questionnaire reports of average time spent in activities. In order to check the correspondence between these two measures (self-report vs. heart rate monitor data), we used Pearson correlations. The correlation between these two data sets yielded an r of 0.59 and a p-value of .017, indicating a significant correlation between the two measures, and thus reliability of the data collected through questionnaire. The high reliability of the data is even more striking when we consider the small sample size (16), and that the data were collected during a period of light training (whereas the questionnaire asked for reports of *average* amounts of time spent in activities).

Analysis and Results

The analysis of part 1 of the questionnaire consists of 4 sections: i) analysis of time and relevance data for the 6 activity categories, ii) analysis of the specific training activities in the on-field and off-field categories, iii) analysis of time and relevance data compared between training alone and in a group for a sub-group of activities, and iv) a comparison of the biographic data on the FIFA and non-FIFA referees.

i) Activity Categories

Total training and travel. One thing that influences training volumes is the amount of time already taken up by game related travel. Because geography can potentially affect the amount of time spent travelling, it was important to determine whether English and Belgian referees are required to travel differentially. Using a 2 Nationality (Belgian, English) x 3 Year (1st, 1998, current) mixed ANOVA, we found a main effect of Nationality $F(1, 2) = 13.29$, $p < .01$, with the English referees spending more time travelling overall than the Belgians (4.6 hr/week vs. 2.0 hr/week, respectively). Not surprisingly, there was also a main effect of year, $F(2, 48) = 13.34$,

$p < .01$, with travel time increasing significantly each year from an average of 1.2 hours a week in the first year, to 3.2 in 1998 and 5.5 in the current year.

In addition to geographical differences in our sample, refereeing international matches may also impact amounts of time spent travelling. We thus compared the FIFA and non-FIFA referees in a 2 Group (FIFA, non) x 3 Year (1st, 1998, current) ANOVA. While there were no significant group differences, there was a main effect of year, $F(2, 48) = 20.60$, $p < .01$, with significant increases in time spent travelling from the 1st year (1.6 hr/week) to 1998 (4.2 hr/week) to current year (5.9 hr/week). Travel time was thus eliminated from match/match-related activities because it was felt that it would misrepresent time spent in match activities.

Activity categories. While Table 4 shows total time spent in training and travel across years, an examination of training by *category of activities* provides a clearer picture of the microstructure of practice in this group. In order to understand how both perceptions of and types of activities in elite referee training have changed over time, a series of mixed ANOVAs was conducted based on the activity categories. (Tukey's post hoc tests were used on any statistically significant differences.)

Insert Table 4 about here

---

2 Group (FIFA, non) x 3 Year (1st, 1998, current) mixed ANOVAs were performed for on-field, off-field, therapeutic, match and coach/play categories of activities for both the time and relevance data. Given the lack of group differences in these analyses, the groups were then collapsed for all referees.

Figure 1 displays time spent in each activity category for all three years (1st, 1998, current). There is an overall pattern of increasing training volumes in all activity categories, with the exception of coaching/playing. Specifically, time spent in on-field activities increased significantly from 1.7 hours per week in the 1st year to 2.9 in 1998 and 4.9 in the current year, F (2, 48)= 64.797, $p< .01$. For off-field activities, volumes increased from 3.3 and 2.9 hours per week in the first year and 1998, respectively, to 7.3 hours per week in the current year, F (2,48)=12.85, $p< .01$. This same pattern was shown for therapeutic activities, with very little time being spent in the 1st year or 1998 (0.1 and 0.5 hr/week, respectively) and significantly more in the current year (1.4 hr/week), F (2, 46)=8.17, $p< .01$).

Insert Figure 1 about here

For match and match-related activities, while the same pattern of increasing volumes was shown, F (2, 48) = 26.83, $p< .01$, with 3.9 hours per week in the first year increasing significantly to 6.6 hours per week in 1998, and 8.5 hours per week in the current year, there was also an interaction between group and year, F (2, 48) = 3.85, $p = .028$). Tukey's post hoc tests revealed that for the current year, non-FIFA referees spend more time in match and match-related activities than FIFA referees (9.9 hr/week, versus 7.1 hr/week, respectively). This was also the case in 1998, with non-FIFA referees spending an average of 7.9 hours per week in match activities; a level significantly higher than the 5.2 hours on average spent by FIFA referees. There were no differences in the amount of time spent in match activities for the first year of

formal refereeing. In contrast to the findings for the first four activities, time spent in coaching/playing has not significantly changed over the years.

Figure 2 shows ratings of relevance to improving refereeing performance for each activity category across all three years. Once again, increases in ratings for all categories except coaching/playing indicate rising perceptions of relevance. For example, both on-field, $F(2, 48) = 43.02$, $p < .01$, and off-field $F(2, 48) = 31.98$, $p < .01$, activities' perceived relevance have increased significantly. Ratings have increased from the 1st year (2.3, 1.9, respectively) to 1998 (3.3, 2.5, respectively) to the current year (5.4, 3.9, respectively). Therapeutic activities' relevance ratings have similarly increased significantly each year, $F(2, 48) = 23.17$, $p < .01$, from 1.7 to 2.8 to 4.6 for the first year, 1998 and the current year, respectively. Ratings of relevance for match activities in the first year of formal refereeing are lower than those for the current year, $F(2, 48) = 6.70$, $p < .01$, but were not different when compared to ratings for 1998. While mean ratings of relevance for coaching/playing activities are significantly higher in the current year (2.9) as compared to 1998 (2.5) or the first year (2.4), $F(2, 48) = 5.48$, $p = .007$, there was very little variability between participant ratings for each year (current year, SD = 0.73; 1998, SD = 0.62; first year SD = 0.69), indicating a similar pattern to that shown in the training volume data: yearly increases in all categories except for coaching/playing activities.

Insert Figure 2 about here

Table 5 displays activity categories ranked in order from the highest ratings in perceived relevance to the lowest ratings across all three years. These comparisons

provide an indication of the movement across time along this dimension for the different categories. The most notable changes are for the coaching/playing and therapeutic activities. From the first year to the current year, coaching/playing activities move from being ranked the second most relevant to the least relevant group of activities. In contrast, therapeutic activities move from the lowest ranked relevance in the first year to the third most relevant in the current year.

Insert Table 5 about here

---

ii) Analysis of Specific Activities

On-field activities. Subsequent to comparing volumes and relevance of training by categories, we analysed changes in specific activities across time. This was an attempt to determine how the microstructure of training and relevance may change with increasing skill. First we examined the on-field category, composed of six different running-based activities (recovery, low intensity, high intensity, speed endurance, speed/agility, coordination). Given the lack of group differences in the previous analyses, we collapsed all referees into one group and used one-way repeated measures ANOVAs comparing three levels of Year (1st, 1998, current). These analyses showed main effects with yearly increases in training volumes for all activities (recovery: $F(2, 48) = 72.18$, $p < .01$; low intensity, $F(2, 50) = 4.00$, $p = .02$; speed endurance, $F(2, 50) = 21.06$, $p < .01$; speed/agility, $F(2, 50) = 23.44$, $p < .01$; coordination, $F(2, 50) = 5.21$, $p = .009$) except high intensity (HI) running, which did not change from year to year, $F(2. 50) = 2.21$, $p = .12$. Figure 3 illustrates these results.

Insert Figure 3 about here

We also compared ratings of relevance for the six on-field activities for each time period, again using one-way repeated measures ANOVAs. Figure 4 shows the results of these analyses. Relevance ratings increased across years, including those for HI running (recovery, F (2, 50) = 20.25, p< .01; low intensity, F (2, 50) =10.16, p< .01; high intensity, F (2, 50) = 23.33, p< .01; speed endurance, F (2, 50) = 43.47, p < .01; speed/agility, F (2, 50) = 37.12, p < .01; coordination, F (2, 50) = 8.79, p < .01).

Insert Figure 4 about here

Off-field activities. The same analyses were performed for the off-field training activities (strength, flexibility, technical referee skills, other training, video training, playing tactics psychological skills training (PST)). Figure 5 shows yearly volume (displayed in min/week) increases for all activities (flexibility, F (2, 50) = 5.91, p< .01; technical skills, F (2, 50) = 5.14, p < .01; video, F (2, 50) = 12.19, p < .01; playing tactics, F (2, 50) = 4.46, p < .05; PST: F (2, 50) =17.19, p<.01), with the exception of strength training and other training (e.g., cycling), F (2, 50) = 1.46, p =0.24, and F (2, 50) = 2.61, p = 0.08, respectively. In fact, while not significant, other training is the only activity that shows a pattern of decreasing volumes over the years.

Similarly, Figure 6 shows that perceived relevance increased across years for all activities (strength, F (2, 50) = 9.8, p < .01; flexibility F (2, 50) = 25.25, p< .01; technical skills, F (2, 50) = 11.10, p< .01; video training, F (2, 50) = 15.30, p <. 01;

playing tactics F (2, 50) =11.11, p< .01; PST, F (2, 50) =17.19, p< .01) except other training (F (2, 50) = 0.19, p = 0.83). Specifically, relevance ratings were higher in the current year when compared to both preceding years.

Insert Figure 5 and Figure 6 about here

Sleep. The deliberate practice literature has often acknowledged the need for rest in response to the physical (and mental) demands of large amounts of training (e.g., Ericsson, Krampe & Tesch-Römer, 1993). Changes in sleep thus reflect physical and psychological demands, as well as the potentially greater need for recovery due to aging. While amounts of time spent in sleep did not change significantly over the years (M = 42.9 hr/ wk in first year, 43.1 in 1998 and 44.9 in the current year), its perceived relevance did. A one-way analysis of variance for Year (1st 1998, current) showed a main effect, F (2, 50) = 5.26, p< .01, with ratings of relevance for sleep in the current year (M = 7.5) significantly greater than for the first year of refereeing (M = 5.7).

Referee-coach perceptions of relevance. We compared relevance ratings by the two coaches and selected the activities for which they agreed (i.e., ratings did not differ by more than 2.5 points). We then compared coach perceptions with those of the referees using a mean hypothesized t-test. Table 6 displays the results of these comparisons. Of the 10 activities analysed, coach and referee ratings were similar on four activities (low intensity running, coordination, strength, and technical referee skills). The six remaining activities were rated more relevant by coaches than by

referees (high intensity running, speed endurance, speed and agility, other training, video training, and psychological skills training).

Insert Table 6 about here

---

iii) Nature of Training: Alone Versus Group

Six on-field and seven off-field activities were divided into training done alone and training done as part of a group. Comparisons were made using 2 Category (on-field, off-field) x 2 Nature (alone, group) x 3 Year (first, 1998, current) repeated measures ANOVAs. For time data, results showed a main effect of the nature of training, F (1, 25) = 33.01, p<. 01. Over the three target years, referees reported more time spent training alone (M= 2.7 hr/week) than in a group (M =0.8 hr/week). There was also a main effect of year, F (2, 50)= 23.26, p <. 01, with significantly more time spent training overall in the current year (M = 2.7 hr/week) than in 1998 (M = 1.6 hr/week), and both significantly more than in the first year (M = 1.0 hr/week).
The same ANOVA design was used for ratings of relevance to improving refereeing performance. Not surprisingly, a main effect of year, F (2, 50) = 45.00, p <. 01, indicated that current ratings (M = 4.6) were significantly higher than those for 1998 (M = 2.9), which in turn were significantly higher than those in the first year (M = 0.1). A main effect of category, F (1, 25) = 32.33, p<. 01, showed that across all years and types of training, referees rated on-field training as significantly more relevant (M = 3.6) than off-field training (M = 2.8). As well, a main effect of the nature of training, F (1, 25) = 19.35, p <. 01, showed that referees rate training alone as significantly more relevant than training in a group (M = 3.6, and 2.8, respectively).

Finally, an interaction between activity category and year was found, F (2, 50) = 12.2, p< .01. Post-hoc tests revealed that on-field training was rated significantly more relevant than off-field training in 1998 (M= 3.3, 2.5, respectively) and the current year (M = 5.3, 3.9, respectively). There was no difference, however, in ratings of relevance for on- and off-field activities for the first year of refereeing (in order, M = 2.3, 1.9).

iv) Biographical Data

Just over half of the participants in the sample consisted of FIFA list referees. Being named to the FIFA list is an important event that can be pinpointed in time. This designation can thus be used as an indicator of expertise in refereeing. Using this sub-sample (n = 14), we explored the relationship between different sources of experience in soccer and appointment to the FIFA list.

Table 7 shows biographical data for both the FIFA and non-FIFA referees. Compared to non-FIFA referees, FIFA referees are significantly younger at present, t (23)= -2.20, p<. 038, with a mean age of 39.1 (versus 43.1 for non-FIFA referees). While the two referee groups do not differ in the number of years they have refereed, coached or played, a t-test comparing refereeing start age shows a significant main effect of group, t (23) = -2.63, p = .015, whereby the FIFA referees began refereeing at a significantly younger age than the non-FIFA referees (M = 18.1, 21.4, respectively).

Insert Table 7 about here

As a whole, the FIFA referee group had formally refereed for an average of 16.4 years before reaching the FIFA list. To understand the relationship between

different types of experience and reaching the FIFA list, we used Pearson correlations to compare number of years to reach FIFA with number of years accumulated in coaching, and playing. Number of years to reach FIFA was not significantly correlated with either accumulated number of years playing (M =9.43, SD = 5.14, r =. 08), or accumulated number of years coaching (*M* = 1.5 yrs, SD = 4.20, r = -.24). It should be noted, however, that only 3 referees indicated any experience whatsoever in coaching.

Part 2: Current Perceptions of Activities

Method

Participants

Participants for part 2 were the same referees as in part 1.

Materials and Procedure

A second part of the questionnaire was collected a week after the first part. In part 2, participants indicated current perceptions of relevance, effort, enjoyment and concentration for each activity. Given that RPE values for activities may influence ratings of effort, these were excluded from the second part of the questionnaire. Similarly, no activities were divided into training in a group versus alone.

Analysis and Results

For each activity on each dimension (i.e., relevance, effort, enjoyment, concentration), ratings were calculated and compared to the overall mean for all activities (on that dimension) using mean hypothesized t-tests. Significant differences were determined using Bonferroni's method for alpha adjustment, with alpha divided by the total number of activities (33). Results are shown in Table 8, with means

significantly above and below the overall mean indicated. These findings will be discussed by dimension.

Insert Table 8 about here

A large number of activities were rated higher in relevance than the grand mean for this dimension. As can be seen in Table 8, four out of six on-field activities, and four out of nine off-field activities were rated very high in relevance. Specifically, high intensity runs, speed endurance, speed/agility and coordination were rated very high. High ratings were also given for flexibility, and sleep, indicating the need for recovery and injury prevention in response to high intensity training.

Playing and coaching activities were rated lower than the average in relevance (player meetings, travel to coach, travel to play, playing, and reading about soccer). Referees distinguished specifically between the relevance of refereeing league games, and refereeing exhibition games. Officiating in league games was an activity rated not only significantly higher than the grand mean, but highest overall (9.3). Refereeing exhibition games, in contrast, was rated *lower* than the overall mean (3.7), though not significantly so.

The same on-field activities that were rated higher than the mean for relevance were also rated higher for effort (high intensity runs, speed endurance, speed/agility, coordination). Only two off-field activities were rated as requiring high effort (strength training, technical referee skills). Once again, league and exhibition refereeing were distinguished, with league refereeing perceived as the most effortful

activity (9.2) and exhibition refereeing ranked lower than the overall mean (3.8), although not significantly so.

Activities that were rated low in effort were those related to coaching and playing (player meetings, travel to coach, travel to play). Not surprisingly, the other activities with low effort were reading about soccer, sleep and non-active leisure. Referees indicated high enjoyment for the same four on-field activities (high intensity, speed endurance, speed/agility, coordination), and only one off-field activity (technical referee skills). Not surprisingly, refereeing league games was highest in enjoyment (9.1), with refereeing exhibition games comparatively lower, but again not significantly different to the overall mean (4.3 versus 5.4, respectively). Sleep and active leisure were also rated as highly enjoyable. Five out of the eight coaching and playing activities were rated low in enjoyment (travel to coach, travel to play, coaching, coach meetings, player meetings), with the exception of spectating soccer, which was significantly above the mean.

For concentration ratings, the same pattern was shown again, with the on-field activities of high intensity running, speed endurance, speed/agility, and coordination rated above the mean. Referees indicated that high concentration is also needed in technical referee skills, video training and referee meetings, with all three rated significantly higher than the mean. Referees once again distinguished between refereeing exhibition and league games with league games requiring high concentration (M = 9.4) in contrast to exhibition games (M = 4.2). Travel to referee is also an activity requiring high concentration, with a mean rating of 7.5. Finally, very little concentration was associated with recovery running, sauna/Jacuzzi, player meetings, travel to coach, travel to play, sleep and non-active leisure.

Relationships Between Relevance, Effort, Enjoyment and Concentration

To determine the relationship between the dimensions of relevance, effort, enjoyment and concentration, we converted the average ratings for each activity to z-scores and then correlated the dimensions using a Spearman's Rho analysis. A Bonferroni correction was again made to adjust the alpha level. Correlations showed a significant relationship between ratings of effort and enjoyment, $t(31) = 3.74$, $p < .001$, $R = 0.56$, and ratings of effort and concentration, $t(31) = 13.27$, $p < .001$, $R = 0.92$. Contrary to previous studies (e.g., Ericsson et al., 1993), the relationship between relevance and effort was not significant, $t(31) = 0.85$, $p = 0.4$, $R = 0.15$.

Discussion

The first important finding in these data is that both the volume of training and the relevance ratings, almost regardless of activity, increased over the three periods examined. While changes in ratings of relevance over time have not previously been examined, the increases in amounts of time dedicated to training reflect previous findings with athletes (e.g., Young & Salmela, 2001; Helsen et al., 1998) In this respect, then, elite referees show patterns similar to athletes in the training changes that accompany becoming more expert. They differ from athletes, however, in their very limited participation and low relevance ratings for coaching and playing activities. Indeed, we expected to find parallel decreases in time spent in playing and coaching activities with increases in time spent in refereeing activities. This was not the case, however. Participants spent very little time coaching and playing even in the first year of refereeing, which showed that they specialize very early. The reason for this early specialization can be seen in that, when we examined the FIFA referee sub-

sample, coaching and playing experience were not related to reaching higher levels of refereeing performance.

As referees become more elite and continue to specialize, not only does the volume of training increase, but this brings about a differentiation of practice activities. In the first years of refereeing, training is focused on high intensity (HI) running. As referees become more elite, they maintain a large volume of HI running, but also increase time in virtually all other activities, adding different *types* of training to their existing activities.

In all three years queried, referees report spending more time training alone than in groups. This is in contrast to findings tracking team sport athletes' careers, which show decreases in individual training that accompany increases in team practice (Helsen et al., 1998). There are two possible explanations for this finding. First, although soccer referees play a role in a team sport, they are essentially individual performers, creating little technical need for team or group training. Thus, their career practice pattern would more closely resemble individual rather than team sport athlete careers. A second possible explanation is that, until recently, referees had very little opportunity for group training.

To compare need versus opportunity for group training, we can turn to perceived relevance. Referees rated training alone as more relevant than training in a group for all years, including those in which group training was available. This indicates that the greater amount of time spent training alone over training in a group is based on needs, or lack thereof, rather than limited opportunity.

Another interesting finding is the comparison between relevance ratings for on-field and off-field training. When compared across the years, there is no difference

in perceived relevance for the first year of formal refereeing. In both 1998 and the current year, however, referees rated on-field training as more relevant than off-field training. This may reflect what McPherson (1993) reports, that, with expertise, performers proceed from acquiring factual/declarative knowledge and skills to sophisticated procedural knowledge consisting of condition-action sequences. In the case of the referee, anticipating and moving to where the play will be is an extremely important skill. It may be that, at elite levels, the off-field, basic knowledge tasks (e.g., video training for decision-making) are less important than procedural knowledge, especially in the faster paced level of play. On-field activities (recovery runs, LI runs, HI runs, speed endurance, speed and agility, coordination) are essential to executing good anticipation and positioning skills. While large cognitive demands are still present at these higher performance levels, they are inextricably linked to the motor components of the task, making the physical conditioning aspects of training and performing more demanding and thus more relevant.

Furthermore, the emphasis placed on the *motor* demands of performance can be seen when we compare relevance ratings for refereeing league games (actual performance) as compared to refereeing exhibition games, in both parts 1 and part 2. We can argue that the basic perceptual-cognitive demands of refereeing a league game are comparable to refereeing an exhibition game. However, a comparison of heart rate provides evidence that the physical demands may be very different. For example, one referee officiating in two matches in the same week spent 77% of a "schools match" at 60-75% of his HRmax, and 4% of the time at 76-85% of HR max. In contrast, when refereeing a league match, the referee spent 29% of the time in the 60-75% HR zone, and 58% at 76-85% of HRmax. This clearly shows the much higher

physical demands at higher performance levels. Our data show that referees are very aware of these loads based on the higher relevance ratings for tasks with similar *physical* demands to performance (e.g., speed endurance) as compared to tasks with similar *perceptual-cognitive* demands (e.g., refereeing an exhibition game). For instance, the relevance of video training has increased over the years, but is only rated at 5.0 for the current year.

The most striking findings with regards to perceived relevance, effort, enjoyment and concentration contrast refereeing and coaching/playing activities, and more specifically, refereeing league and exhibition games. Whereas several training activities are perceived as high on all dimensions (e.g., high intensity, speed endurance, speed/agility, coordination), several playing/coaching activities are low on all (e.g., player meetings, travel to play). While this may reflect that referees do not participate in playing and coaching activities, it is also indicative of their perception that playing and coaching have limited applications to refereeing. In addition to acknowledging the importance of training for the physical demands of the fast-paced level at which they referee, referees also recognize the effort and concentration that this training requires. A number of activities, while highly strenuous, are also very enjoyable to referees, a pattern previously shown with athletes (e.g., Young & Salmela, 2001). Indeed, this is reflected in the strong relationship between effort and enjoyment. As well, the strong relationship between effort and concentration indicates referees allow that some activities require both physical and mental effort. For example, speed endurance, speed/agility, and coordination are activities rated high in both effort and concentration.

As referees gain expertise and perform at higher levels with inherently increased physical demand, there is a greater need for physical recovery. This need is reflected in the increased relevance ratings for sleep and for therapeutic activities (e.g., physiotherapy) across the years. The increased relevance of flexibility over the years further reflects a need for recovery either as injury rehabilitation or prevention. It may be argued, however, that the increasing relevance of these activities reflects aging as much as heightened training and performance demands. Indeed, the impact of aging on performance and in injury/recovery is especially relevant in referees given their lengthy careers (mean years refereeing = 21.3) and average current age (M = 40.8). Additionally, this need for recovery from highly competitive and fast-paced games may explain why FIFA referees spent less time in the current year and in 1998 in match and match-related activities than their non-FIFA counterparts.

Conclusions

This study is the first time that practice activities have been examined in elite referees. Using the deliberate practice framework helps us to draw a picture of these elite performers as unique from athletes. These are not retired coaches or players who have gone on to excel in another role within their sport. Rather, they have specialized early on as referees, and committed a great deal of time to their skill development. As they become more expert, this time commitment not only increases, but practice becomes more diverse with a greater variety of training activities taken on.

While these referees officiate in some of the highest level of play in the world, they average 41 years of age. The importance of keeping up with the fast-paced games is reflected in the large amounts of time and high perceived relevance of speed and agility-type physical skills. In contrast, this group does not spend a great deal of

time training purely cognitive skills, and similarly does not rate refereeing exhibition games as relevant to improving performance. This surprising finding may be due to the slower pace of exhibition games and thus their limited replication of performance demands. A future examination of novice referees officiating at lower levels of play with a slower pace of action could be used to compare the relative perceived importance of cognitive, decision-making or declarative knowledge skills with physical conditioning at different stages of development/expertise.

While the sample consisted of 26 of the top referees in the world, a subsample of FIFA referees showed an average of 16 years to reach this expert level of performance. When compared to the "10 year rule", this seems to be a lengthier training period than shown in previous studies (e.g., Helsen et al., 1998). It should be kept in mind, however, that using appointment to the FIFA list may be a conservative measure of expertise, given that many non-FIFA referees officiate in European Cup play and the highest levels of play within their countries. Thus many referees may reach expert levels of performance within a shorter time using other criteria.

The efficacy of the deliberate practice theory in describing practice activities and thus explaining expertise in officiating is germane when seeking to understand how quickly referees become experts. Two main criteria of deliberate practice are that practice is monitored, with immediate and informative feedback (e.g., Ericsson et al., 1993). In the traditional practice paradigm of teacher and student or coach and athlete, this is a natural occurrence. However, while the referee groups in this study benefit from the guidance of referee-coaches, there is very little practice that receives performance feedback. The referee-coaches play a strength and conditioning role, rather than a skill acquisition role. Referee training may replicate the physical

conditioning demands of games, but there is little emphasis on responding to players, decision-making, and game positioning skills. Practice and acquisition of these skills, and feedback in training (as well as performance) may hasten expertise. Indeed, the analogy can be made to a gymnast training flexibility, strength, and gymnastic elements, but very rarely practicing element combinations or her full routine outside of competition. Where athletes receive feedback in training, referees receive performance feedback only from performances.

There are limits, however, to the feasibility of including performance replicating practice activities and feedback in referee training. First, the high physical demands of refereeing at this level limit the number of games that can be refereed outside of assigned league games. Second, even if shorter duration games were played to limit physical strain but replicate positioning and decision-making skills (providing a context for refereeing feedback) the logistics of such activities and their evaluation are discouraging and restrictive.

An encouraging development, however, is recent work in assessing the perceptual-cognitive demands of officiating high-level matches. While this process is more difficult than assessing physical demands through heart rate monitors, game analysis and the development of training activities are underway at the FIFA level. The development of perceptual-cognitive training tasks may present a partial solution to the current lack of refereeing feedback during practice. The goal of this training activity, then, is to enable referees to practice skills crafted to match typical on-field demands. The low relevance ratings for video training found in this study may reflect the traditional practice of reviewing a call and discussing it as a group in lieu of individually performed tasks that lead to skill acquisition and refinement. The

development and implementation of perceptual-cognitive training tasks may be reflected in a few years' time in both greater amounts of perceptual-cognitive training, and higher ratings for relevance.

Overall, the emphasis at the FIFA level on organized and monitored physical and perceptual-cognitive training is encouraging. The ultimate goal is for these initiatives and activities to trickle down to more novice performers and not only shorten the time taken to reach elite levels, but in the long run, to create "better" referees overall.

References

Baker, J., Cotê, J., & Abernethy, B. (2003). Sport-specific practice and the development of expert decision-making in team ball sports. Journal of Applied Sport Psychology, 15(1), 12-25.

Ericsson, K. A., Krampe, R. T., & Tesch- Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. Psychological Review, 100, 363-406.

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.F., & Bultynck J-B. (in press). Physical and perceptual-cognitive demands of top-class refereeing in association football. Journal of Sports Sciences.

Hodges, N.J. & Starkes, J.L. (1996). Wrestling with the nature of expertise: A sport specific test of Ericsson, Krampe, and Tesch-Römer's (1993) theory of deliberate practice. International Journal of Sport Psychology, 27, 400-424.

McPherson, S.L. (1993). The influence of player experience on problem solving during batting preparation in baseball. *Journal of Sport & Exercise Psychology, 15*, 304-325.

Oudejans, R.R.D., Verheijen, R., Bakker, F.C., Gerrits, J.C., Steinbrücker, M., & Beek, P.J. (2000). Errors in judging 'offside' in football. Nature, 404, 33.

Plessner, H. & Betsch, T. (2002). Refereeing in sports is supposed to be a craft, not an art: Response to Mascarenhas, Collins, and Mortimer. Journal of Sport and Exercise Psychology, 24(3), 334-337.

Young, B.W. & Salmela, J.H. (2001). Perceptions of training and deliberate practice of middle distance runners. International Journal of Sport Psychology 33(2), 167-181.

Acknowledgements

This research was partially funded through a SSHRC grant to Janet Starkes, and was supported through a SSHRC doctoral fellowship awarded to Clare MacMahon. Partial funds were also provided by a grant to Werner Helsen from F-MARC, the medical branch of FIFA.

Table 1.
Age and Experience of Referees by Nationality and Level

	Age		Start Age[a]		Total Years[b]	
Group	M	SD	M	SD	M	SD
Belgian (n = 7)	37.0	4.4	18.0	1.6	19.4	1.6
English (n = 19)	42.3	4.3	20.1	3.8	19.4	4.3
FIFA (n = 14)	39.1	4.4	18.1	2.2	20.9	4.5
Non-FIFA (n =12)	43.1	4.7	21.4	4.0	21.8	4.3
All (n = 26)	40.8	4.9	19.5	3.5	21.3	4.3

[a] Age at which refereeing was begun, whether formal or informal
[b] Total accumulated years refereeing

Table 2.
Activity Categories used in the Questionnaire

On-field Activities
- Recovery runs
- Low intensity runs
- High intensity runs
- Speed-endurance runs
- Speed and agility
- Coordination and running

Off-Field Activities
- Strength training
- Flexibility
- Technical referee skills
- Other training (e.g., cycling for fitness)
- Video training
- Game-playing tactics
- Psychological skills training (e.g., imagery, relaxation training)
- Referee meetings
- Language practice

Therapeutic Activities
- Physiotherapy
- Sauna/Jacuzzi

Match/Match-related Activities
- Refereeing league games
- Refereeing exhibition games
- Travel to referee

Coaching/Playing Activities
- Meetings-coach
- Meetings-player
- Travel to coach
- Travel to play
- Coach training courses
- Playing in an organized league
- Spectating soccer (live or televised)
- Reading about soccer

Everyday life Activities
- Sleep
- Study/work
- Active leisure
- Non-active leisure
- Travel (e.g., for work)

Table 3.
Questionnaire Activities Divided into Training Alone or in a Group

Recovery training
Low intensity running
High intensity running
Speed endurance
Speed and agility
Strength training
Flexibility
Coordination and running technique
Technical referee skills (e.g., body language, team management with assistant referee)
Video analysis of game play
Other training activities (e.g., cycling, swimming, squash, tennis)
Game-playing tactics (e.g., decision-making outside football match)
Psychological skills training (e.g., stress management skills, mental imagery)

Table 4.
Total Referee Training and Travel Across Year

	Training (hr/week)[a]		Travel (hr/week)	
Year	M	SD	M	SD
1st	7.4	4.6	1.7	1.9
1998	9.8	5.5	4.2	2.9
Current	14.4	5.9	6.2	3.3

[a] Includes time spent refereeing matches

Table 5.
Activity Categories Ranked By Relevance to Improving Refereeing Performance

Rank	Category	M	SD
	First Year		
1	Match/Match-related	5.0	1.4
2	Coaching/Playing	2.4	1.0
3	On-field	2.3	1.9
4	Off-field	1.9	1.5
5	Therapeutic	1.7	1.8
	1998		
1	Match/Match-related	5.7	1.9
2	On-field	3.3	2.0
3	Therapeutic	2.8	2.7
4	Off-field	2.5	1.6
5	Coaching/Playing	2.5	1.0
	Current Year		
1	Match/Match-related	6.0	1.8
2	On-field	5.3	1.3
3	Therapeutic	4.7	2.6
4	Off-field	3.9	1.4
5	Playing/Coaching	2.9	0.9

Note. Ratings are on a 10 point scale where 1 = *low* and 10 = *high*.

Table 6.
Results of Comparisons Between Coach and Referee Ratings for 10 Activities

Activity	t value	p value	Mean rating[a]	
			Referees	Coaches
Low intensity	t (25) = -.09	.93	5.0	5.0
High intensity	t (25) = -7.15	<. 001*	7.0	9.0
Speed endurance	t (25) = -4.36	<. 001*	7.3	8.8
Speed/Agility	t (25) = -5.87	<. 001*	5.8	8.5
Coordination	t (25) = -1.11	.28	3.4	4.0
Strength	t (25) = -0.77	.45	5.6	6.0
Technical skills	t (25) = -2.19	.04	4.2	5.3
Other (e.g., cycling)	t (25) = -8.15	<. 001*	2.4	4.5
Video training	t (25) = -8.29	<. 001*	5.0	9.0
PST[b]	t (25) = -5.08	<. 001*	3.5	5.6

[a] On a scale where 1 = *low*, and 10 = *high*
[b] Psychological Skills Training
* significant with alpha set at .0015

Table 7.
Biographical Data for FIFA vs. Non-FIFA Referees

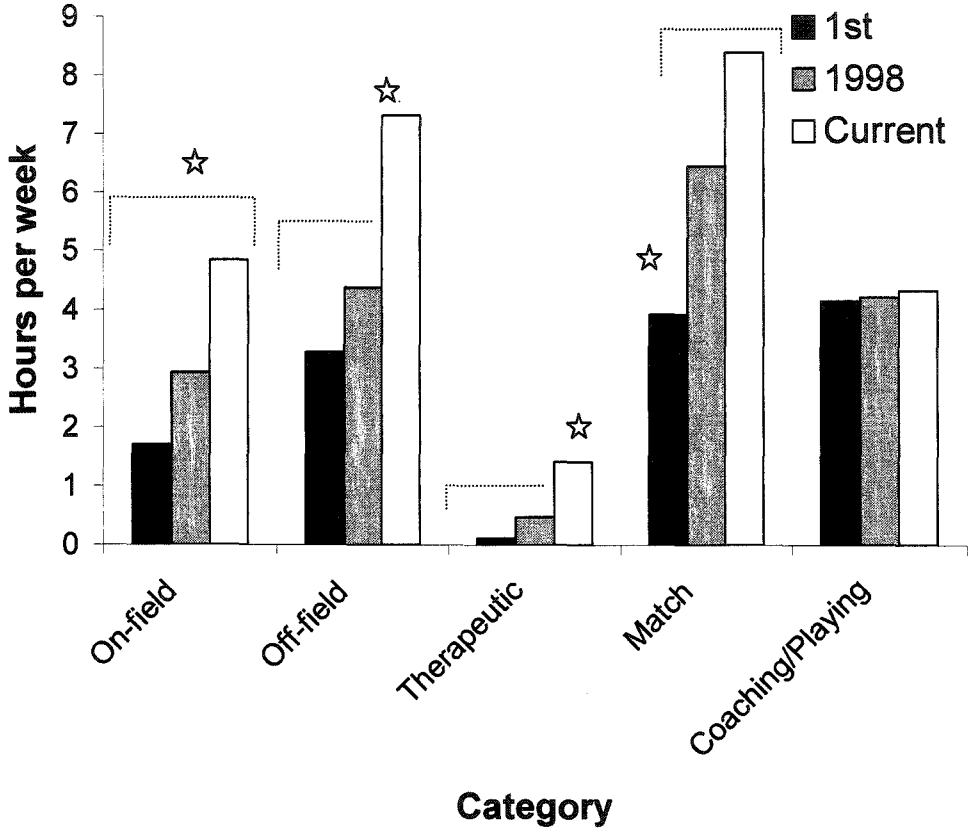
	Current Age		Start Age		Yrs to FIFA		Yrs on FIFA	
Group	M	SD	M	SD	M	SD	M	SD
FIFA (n = 14)	39.1	4.4	18.1	2.2	16.4	3.0	2.5	3.3
Non-FIFA (n= 14)	43.1	4.7	21.4	4.0	----	----	----	----

Table 8.
Mean Ratings of Training Activities Evaluated on Four Dimensions

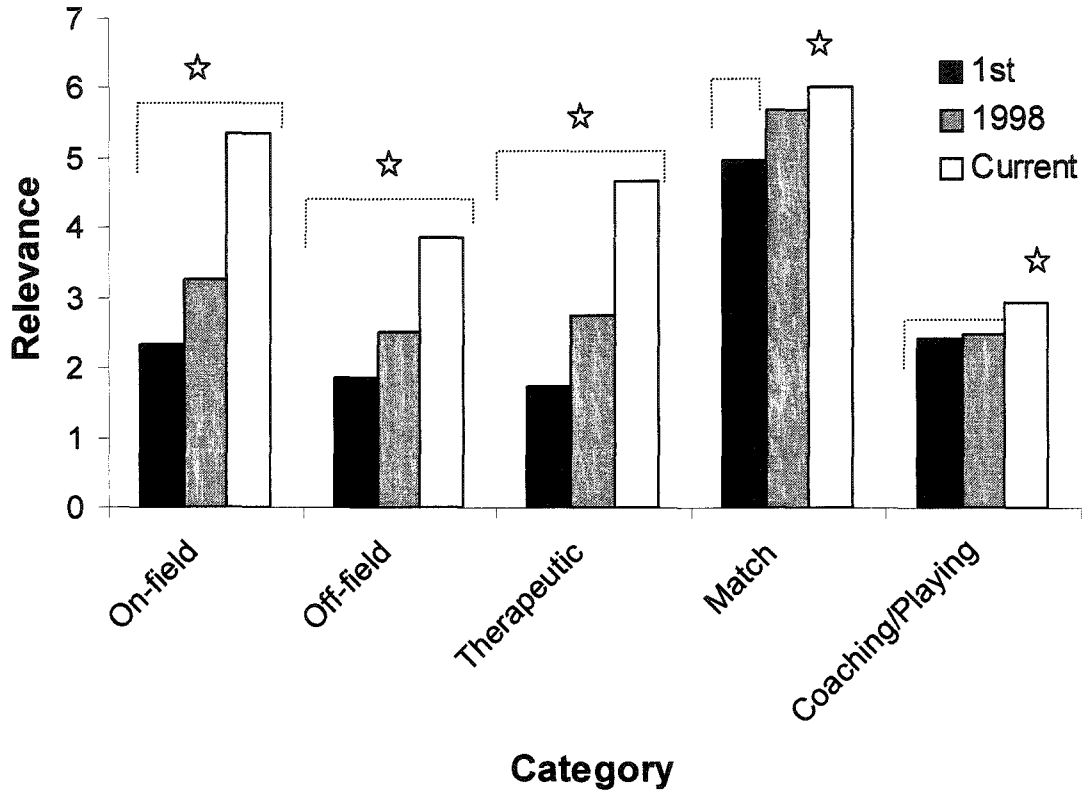
	Dimension[a]			
Category	Relevance	Effort	Enjoyment	Concentration
On-field training activities				
Recovery training	6.4	3.3	5.7	3.2* L
LI	6.9	4.9	6.2	4.5
HI	8.6* H	8.3* H	6.7* H	7.4* H
Speed endurance	8.7* H	9.1* H	6.8* H	7.9* H
Speed and agility	8.2* H	7.3* H	7.2* H	7.0* H
Coordination	8.0* H	7.3* H	7.0* H	7.4* H
Off-field training activities				
Strength	5.7	6.7* H	5.3	6.0
Flexibility	7.5* H	5.8	6.0	5.8
Technical ref skills	8.9* H	6.6* H	7.8* H	8.1* H
Other (e.g., cycling)	5.7	5.7	6.3	5.7
Video training	7.7* H	4.3	6.0	7.2* H
Game-playing tactics	6.0	4.7	5.3	6.2
PST	6.8	4.9	5.2	6.1
Referee meetings	7.6* H	5.9	5.8	6.7* H
Language practice	4.5	3.2	4.0	3.5
Therapeutic activities				
Physio	6.4	5.0	6.1	4.4
Sauna/Jacuzzi	6.2	2.1* L	5.6	1.9* L
Match activities:				
Refereeing league games	9.3* H	9.2* H	9.1* H	9.4* H
Refereeing ex. games	3.7	3.8	4.3	4.2
Travel for refereeing	7.9	6.8	4.3	7.5* H
Coaching/playing activities				
Meetings coach	4.0	3.4	3.2* L	3.8
Meetings player	2.4* L	2.0* L	2.0* L	2.2* L
Travel to coach	2.9* L	2.6* L	2.0* L	3.3* L
Travel to play	1.7* L	2.1* L	1.8* L	2.3* L
Coach training	4.1* L	3.2	3.3* L	3.7
Playing football	3.1* L	3.6	4.5	3.7
Spectating (live or tv)	7.1	3.9	6.7* H	5.2
Reading re: football	4.1* L	2.8* L	4.6	2.5
Everyday life Activities				
Sleep	8.3* H	2.4* L	7.9* H	2.0* L
Study/work	4.2	4.7	2.6	5.5
Active leisure	5.2	4.6	7.4* H	4.5
Non-active leisure	4.4	1.5* L	5.9	3.3* L
Travel	2.7	4.7	4.2	5.5
Overall Mean	5.9	4.8	5.4	5.1

[a]Rated on a 10-point scale where 1 = *low* and 10 = *high*
*significantly above or below overall mean (L = lower than mean, H = higher than mean)

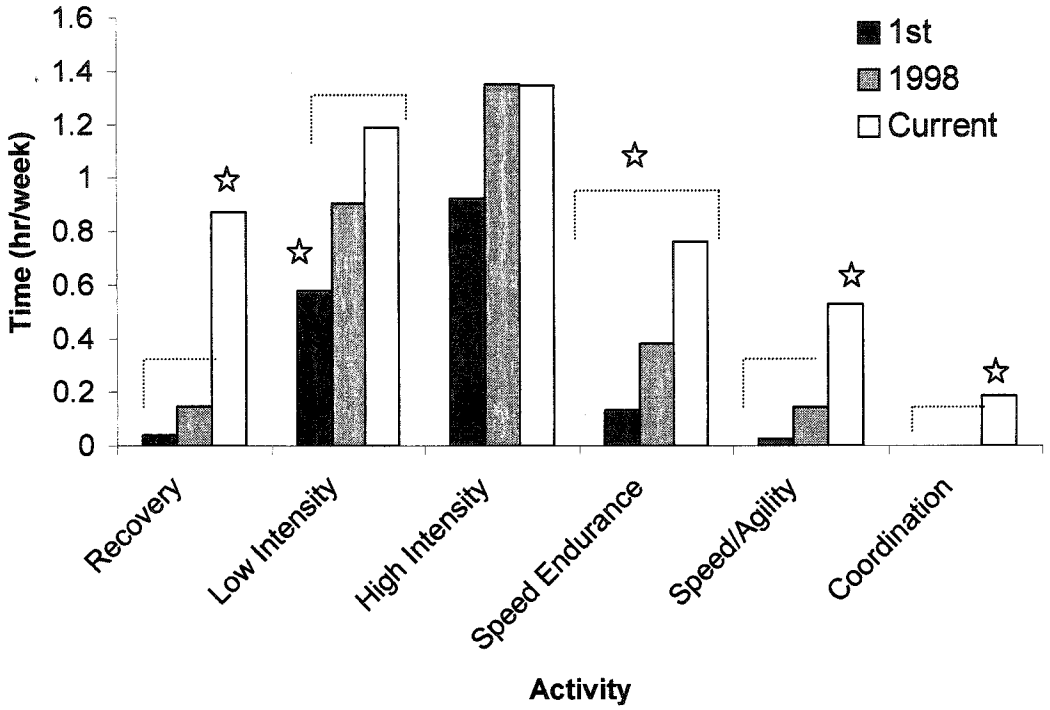
List of Figures



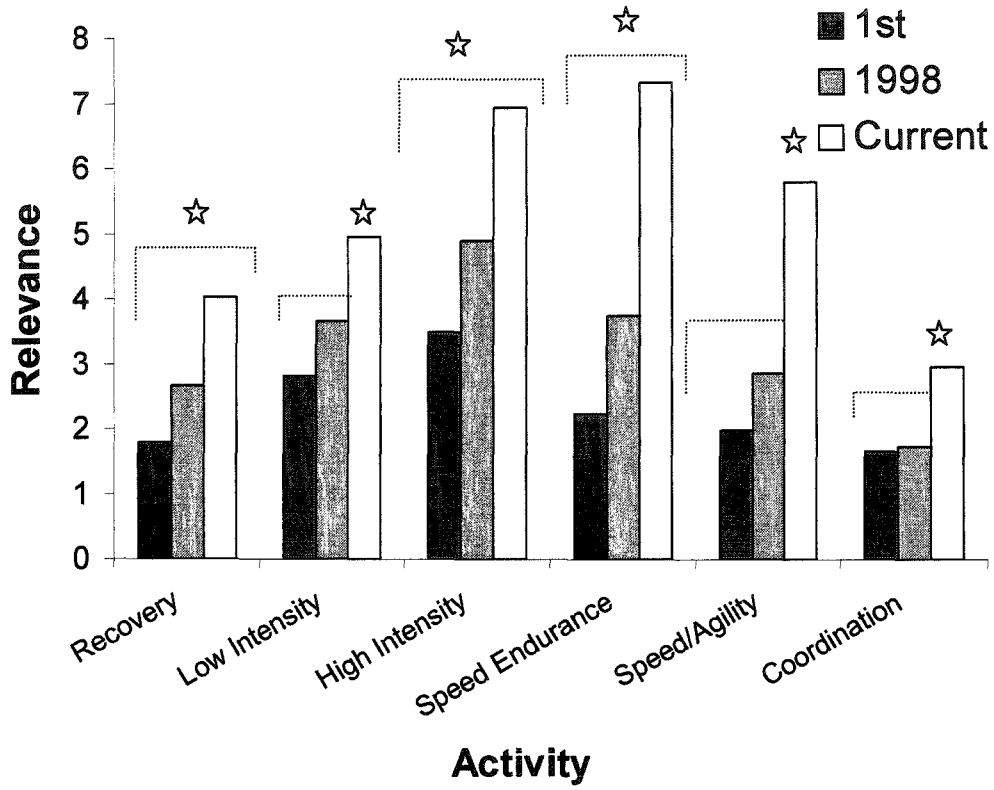
☆ Indicates significant difference at $p < .01$.



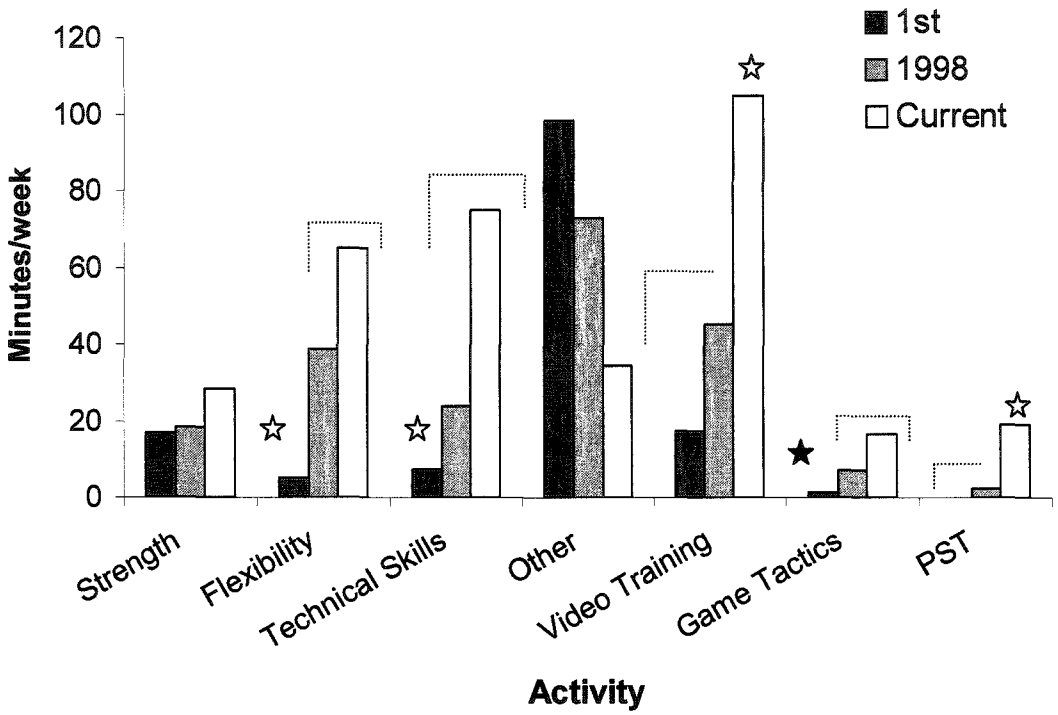
☆ Indicates significant difference at $p < .01$.



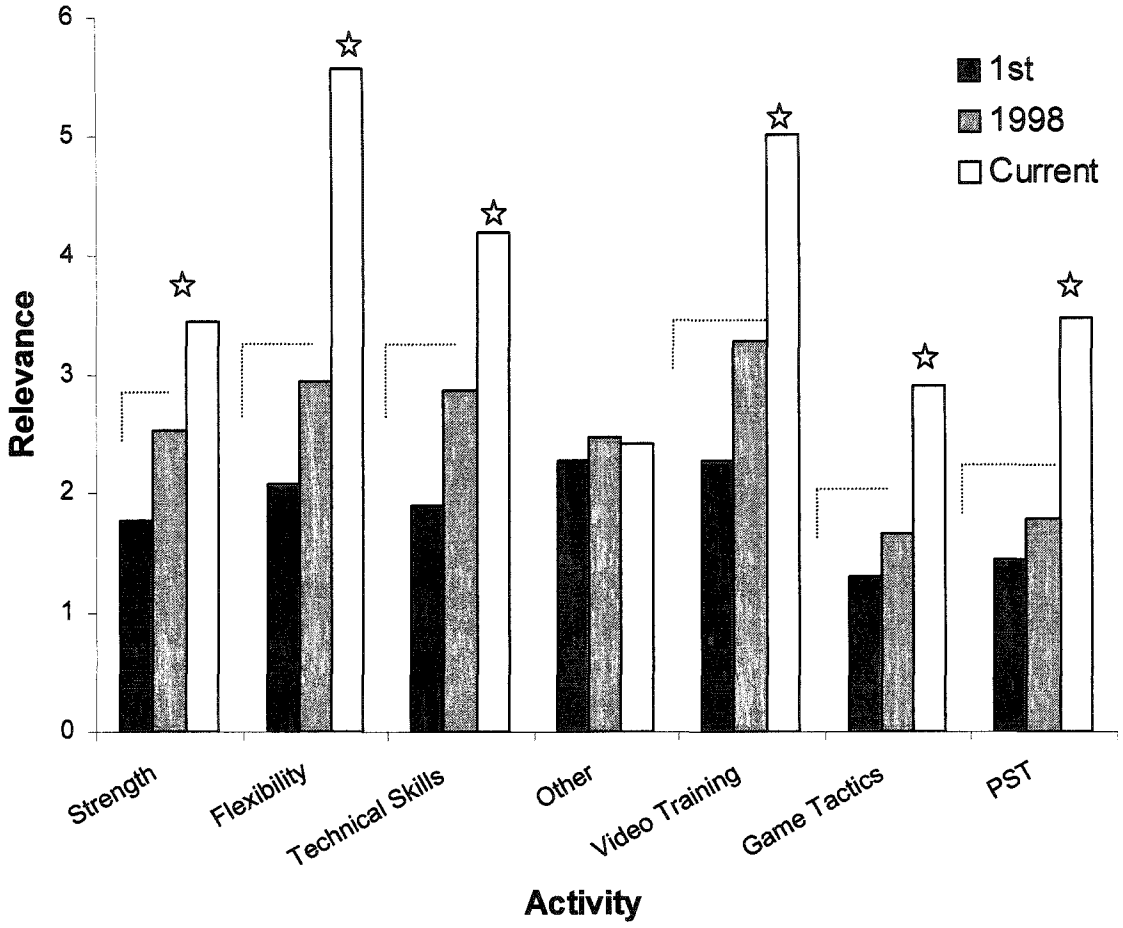
☆Indicates significant difference at $p < .01$.



☆Indicates significant difference at $p < .01$.



☆ Indicates significant difference at $p < .01$.
★ Indicates significant difference at $p < .05$.



☆ Indicates significant difference at $p < .01$.

Appendix 1: Additional Tables

Table A.1
Results of 2 Group (FIFA, non-FIFA) x 3 Year (1st, 1998, current) Mixed ANOVAs for Training in Activity Categories.

Category	Group F-value	Group p value
On-field	F (1, 24) = 0.004	0.95
Off-field	F (1, 24) = 0.13	0.72
Therapeutic	F (1, 23) = 2.56	0.12
Match	F (1, 24) = 2.38	0.14
Coach/Play	F (1, 24) = 0.145	0.71

Table A.2
Results of 2 Group (FIFA, non-FIFA) x 3 Year (1st, 1998, current) Mixed ANOVAs for Relevance of Activity Categories.

Category	Group F-value	Group p value
On-field	F (1, 24) = 0.40	0.53
Off-field	F (1, 24) = 0.26	0.62
Therapeutic	F (1, 23) = 0.08	0.77
Match	F (1, 24) = 0.07	0.78
Coach/Play	F (1, 24) = 0.04	0.80

PAPER TWO

This manuscript has not yet been submitted for publication. For the studies within this paper, a number of people helped me in data collection. I was the major contributor to the data analysis and writing of the manuscript.

Running Head: REFEREEING EXPERTISE

Refereeing expertise: Practice, experience and perceptual-cognitive skill in a laboratory task

Clare MacMahon
Werner F. Helsen[1]
Janet L. Starkes
Koen Cuypers[1]
Matthew Weston[1]

McMaster University, Hamilton, Ontario, Canada

[1]Katholieke Universiteit Leuven, Leuven, Belgium

Address all correspondence to:

Clare MacMahon
Dept of Kinesiology
McMaster University
1280 Main St. West
Hamilton, Ontario
L8S 4K1
Fax: (905) 523-6011
macmahc@mcmaster.ca

Abstract

Two laboratory tasks were used to assess skill related to soccer. The first task was specific to the role of the referee, and the second was a general soccer skill. In the refereeing task, participants viewed video clips of soccer tackles and decided on the appropriate call (i.e., no foul, free kick, free kick followed by a yellow card, free kick followed by a red card). The general soccer task involved the assessment of relative offensive threat of players shown in game clips. A group of seven Belgian FIFA list referees and a group of 34 youth academy soccer players performed both tasks. While there were no group differences in performance on the general soccer test, the referee group was more accurate than the player group in the tackle assessment task. Regression analyses indicated that weekly time spent in practice for playing interfered with performance of tackle assessment. Overall, the results of this study indicate that the role (e.g., player, referee) of the expert is an important factor in skill development, with on-field demands influencing the specific skills acquired. As well, support is provided for the finding that skill is a constituent of expertise rather than simply a byproduct of exposure to the domain. Finally, the potential for training tools for referee skill acquisition and refinement is discussed, as well as the need to consider the ecological validity of the task relative to one's role in the sport.

Introduction

The study of expert performance in sport has expanded tremendously in the last 15 years. The most influential factor for the relatively recent proliferation of knowledge in this area is an adoption of the expert performance approach. This approach, as discussed in detail by Ericsson and Smith (1991), has three phases of investigation. The first step is to elicit expert-novice differences in laboratory tasks. The ability to use a controlled and observable environment to distinguish performers is essential. The next step is to identify the mechanisms responsible for experts' superior performance. The final step involves accounting for the acquisition and learning of these mechanisms.

Through the expert performance approach, researchers have learned a great deal about the characteristics of elite performers in a variety of sports such as field hockey (e.g., Starkes, 1987) wrestling (e.g., Hodges & Starkes, 1996), and soccer (e.g., Helsen & Starkes, 1999). Sport scientists have considered factors such as the nature of the sport as team or individual, (e.g., Helsen & Starkes, 1999), the age and development of performers (e.g., Nevett & French, 1997; Ward & Williams, 2003), and gender (Thomas, Gallagher, & Lowry, 2003). One aspect of sport expertise research that has received very little attention, however, is that sport includes individuals who occupy different roles, performing unique tasks and functions. While the focus has overwhelmingly been on athletes, a particularly intriguing role is that of the referee. Indeed, referees have complex motor and cognitive demands placed on them through their role in a sport. Furthermore, these requirements differ from those of athletes and coaches.

As a result of the lack of existing research, the goal of this study was to explore refereeing expertise. The design was influenced by two key factors. The first factor is that research on refereeing skill is in its infancy and therefore must begin with step one of the expert performance approach. The second key factor is the critical influence of role in sport. Two past areas of research become influential when discussing sport role and expertise.

First, following from research in chess (e.g., Chase & Simon, 1973), sport studies have shown that the expert advantage is specific to the domain of expertise (e.g., Starkes & Allard, 1991). Thus, an expert volleyball player, for example, may show superior recall of a volleyball game play diagram as compared to a novice. This superiority will not be displayed in measures of general recall, memory, or IQ. Moreover, while experts in a particular domain have acquired skill in processing domain-relevant information, the expert advantage is shown only when the stimuli are in a meaningful arrangement. In this manner, then, the expert volleyball player will show no advantage when asked to recall a game play diagram that shows players in random formations unrelated to the game (e.g., warm-up, between halves).

Extending research on the domain-specific nature of expertise, researchers have also investigated the issue of general and specific knowledge related to performers' function. More specific, two studies indicate that within one area of performance, or one sport (e.g., soccer), there may be domain specific skill that is general to all roles (e.g., players and referees) and role-specific skill that differentiates within that domain (Allard, Parker, Deakin & Rodgers, 1993; Williams & Davids, 1995). For example, Allard et al.

(1993) found that officials and coaches in the sport of basketball do not differ in a test of general basketball knowledge. Coaches were superior to officials, however, when asked to recall schematic diagrams of basketball plays, and officials were superior to coaches in both a rules and a signals test. This indicates that information is processed and knowledge is developed as a function of a performer's task demands on the court. From this point of view, critical insight is gained by studying referee performance in both role-specific (refereeing) and role-general skill.

The numerous studies that have examined expert athletic performers provide guidance for the design of laboratory tasks. That is, key findings tell us that laboratory tasks should be perceptual-cognitive in nature. Indeed, we know that athletic performance is best predicted by sport-specific perception and decision-making skills (declarative knowledge) (Starkes & Deakin, 1984). For example, Helsen and Starkes (1999) demonstrated that skill differentiations in expert soccer players are best made on the basis of performance in complex game-simulated decision making, wherein players decide whether to shoot, dribble or pass a ball in response to large screen video of game action. Fast and accurate move selection is thus a key skill for athletic performance that differentiates level of expertise.

Using the sport of soccer, we created two tasks in order to assess both referee-specific and role-general skill. The referee-specific task was designed to represent real-world performance requirements in a well-defined activity, as per the recommendations of Ericsson and Smith (1991). We focused on the application of the rules or laws of the sport as a key responsibility in refereeing. More specifically, we hypothesized that the

detection of infractions and assignment of appropriate calls is a perceptual-cognitive skill specific to refereeing. Focused on this skill, we chose the soccer tackle as a key infraction. The task thus requires assessment of game play video involving tackle situations.

The second laboratory task was designed to assess role-general skill in soccer. Using freeze frames of game situations, participants were required to identify the relative offensive threat of players in the clips. This task involves both anticipation and the use of probabilities. That is, participants must evaluate the situation in order to anticipate a number of likely outcomes, and then determine which outcome poses the most threat to the defense. While directly linked to the performance goals of players (i.e., score goals against your opponent and prevent goals against your team), the skills involved in this task are also inherent to refereeing. Indeed, close proximity to the game action is critical for infraction assessments. Because the action can move from one location to another with extreme rapidity, referees use anticipation to begin their movements early and thus relieve time-pressures. This ability to "read the game" may thus contribute to *both* playing and refereeing performances, whether as a central skill in the former case, or a related skill in the latter. Given this perspective, the assessment of relative offensive threat was used to assess role-general soccer skill.

Taken together, we tested the performance of a referee group and a player group on both a referee task (tackle assessment) and a soccer general task (relative offensive threat). We predicted that the referee group would outperform the player group in the referee task, but show equally proficient performance in the soccer general task.

Furthermore, whereas studies consistently show a monotonic relationship between amounts of deliberate practice and expertise (e.g., Ericsson, Krampe, & Tesch-Römer, 1993), practice has not been directly linked to performance in laboratory tasks. We therefore combined practice, experience and referee task performance data for the referee group to explore whether greater amounts of practice and experience afford the expert referee superior role-specific perceptual-cognitive skill.

This study is part of ongoing research with high-level soccer officials. As such, the data are secondary analyses of data from two other studies. The first study (Cuypers, 2003) assessed perceptual-cognitive skill in groups of players and referees. The second study (MacMahon, Helsen, Starkes, & Weston, in preparation) examined the deliberate practice activities of 26 elite soccer referees. The secondary analyses on these two data sets allow an examination of referee expertise from a novel angle. Not only did this data re-analysis allow us to examine referee-specific and soccer-general perceptual-cognitive skill, we were also able to look at the relationship between referee experience and practice with performance on these tasks.

Method

Participants

Data were used from a total of 41 participants from two previous studies. One group consisted of 7 Belgian FIFA list (Fédération Internationale de Football Association) referees with an average of 19.4 years of refereeing experience (sd = 4.2), 8.7 years of playing experience (sd = 4.2) and a mean age of 37 (sd = 4.4). As FIFA referees, this group consisted of the top referees in Belgium. A second group was

comprised of 34 youth academy soccer players who attend a soccer academy full time and play in top Belgian soccer clubs, and on the national team for their age group. The mean age of the players was 16.3 years (range 14-18), with an average of 10.1 years of playing experience (sd = 1.4 years). None of the players had any refereeing experience. An independent t-test comparing the two groups (FIFA referees and academy soccer players) confirmed that they were no different in the number of years of playing experience accumulated, t (39) = 1.6, p = 0.12.

Materials and Procedure

Both groups completed perceptual-cognitive soccer tasks as part of a previous study (Cuypers, 2003).

Referee Task: Tackle Assessment

Digital film clips of soccer tackles were provided by FIFA and UEFA (Union of European Football Associations) and displayed on a large screen with an LCD projector. Two blocks of clips were shown, each with twenty clips. Participants were instructed to make a decision about the tackle shown in each clip, choosing one of four possibilities: 1) no foul, 2) free kick, 3) free kick followed by a yellow card, or 4) free kick followed by a red card. After each clip there was a window of 10 seconds to make a decision before the next clip was shown. The correct decision had previously been established for each play by a FIFA panel of experts. The clip number was announced before each trial to ensure that participants were attentive at the start of each clip. Participants were tested in groups: a referee group and a player group. Group members were instructed to refrain from communicating with each other during testing.

Soccer-General Task: Relative Offensive Threat

This task was borrowed from a previous study completed by Ward and Williams (2003), after obtaining the videotape stimuli from the authors. The videotape consisted of video clips of soccer game action with a red team playing a white team. All of the clips were used in this study. A TV/VCR and an LCD projector were used to display the clips on a large screen. Before each clip, a red box appeared on the screen to indicate the "start" position of the ball in the upcoming action. At a specific point in the clip, the action froze. During the freeze frame, participants were asked to assess the situation from a defensive perspective. After viewing the clip, participants were provided with a photocopy of the last freeze frame in order to record responses. The question was: as a defensive player, who are the most threatening players other than the ball carrier (e.g., which players are in a position to break down the defence and build a good attack, or have an opportunity to move closer to scoring a goal)?

To answer this question, participants were asked to draw a box around the ball carrier, and then circle the players they felt were a likely threat to the defence. There was no limit to the number of players participants could circle. Participants were also asked to rank the relative offensive threat of the circled players by writing the number one next to the most threatening player, the number two next to the second most threatening player, and so on. Key players and the correct order of threat for each play were determined by a panel of coaches (cf. Ward & Williams, 2003). Three practice trials were completed to provide task familiarization. These practice trials were followed by the 16 test trials.

Prior to completing both perceptual-cognitive tasks, participants were asked for biographical data on the number of years spent playing soccer, number of weekly hours practice to play soccer, number of years refereeing, and number of weekly hours practice to referee.

Biographical Refereeing Practice Data

In addition to the data related to experience, practice data for the referee group were collected as part of a larger study investigating career changes in practice patterns (MacMahon, Helsen, Starkes, & Weston, submitted). Average weekly time spent in referee practice was collected by retrospective questionnaire for 3 specific years: first year of formal refereeing, 1998 (five years prior to the questionnaire date) and the current year. Weekly amounts of training were used to calculate yearly amounts of training for each target year. In addition, an estimate of total accumulated training for three years was made by adding the three yearly amounts of training together.[1]

The referees in the study regularly recorded heart rate data for all training sessions. These data were used to log time spent in training. We compared questionnaire reports for the current year with heart rate training log data for the same time frame. A significant correlation between the self-reported time in training and the training log reports (Pearson r= .59, p= .017) indicated the reliability of the questionnaire data.

---

[1] This measure will be referred to throughout the paper as accumulated training although it is of course more accurately cross-sectional data using three target years.

Analysis and Results

The data were analysed in three phases: role-related differences in the referee tackle assessment task, expert performance in the referee tackle assessment task in relation to practice and experience data, and role-related differences in the general soccer task.

Referee Task: Tackle Assessment

Role-Related Differences

The first analysis compared performance by referees and players in the referee tackle assessment task. Given the complexity of the task, the first block of 20 trials was used as practice, with analysis performed on the second block of 20 trials. For the test block, accuracy of responses was calculated and expressed as a percentage. Figure 1 presents these performance data by group. This figure illustrates that the player and referee groups differ in performance on this task. Referees are quite proficient with a mean accuracy rate of 80.6% (sd = 6.7%). In contrast, players are accurate on only 55.1% of the trials with a large standard deviation of 13.0%. These performances are significantly different, as indicated by the results of a one-way ANOVA with Group as the two-level factor (referee, player), $F(1, 39) = 16.2$, $p < .001$.

Insert Figure 1 about here

Next, we combined the referee and player data to examine the influence of different types of experience on task performance, and pinpoint potential factors contributing to group differences. Because there were two groups (player, referee), we

first tested to see which variables reliably predict group membership. A stepwise multiple regression with group as the dependent variable indicates that years refereeing, hours per week in refereeing practice, years on FIFA, weekly playing practice, and years playing predict group membership. These five variables together account for 99% of the variance, ($R^2 = 0.99$, $F\,(5, 34) = 704.8$, $p <. 001$). A multiple regression using these five variables was then performed for the accuracy measure.

A stepwise regression identified hours per week in refereeing practice as a moderately good predictor of accuracy, $F\,(1, 5) = 7.69$, $p = 0.039$, with an $R^2$of .61, indicating 61% of the variance accounted for. Results show that weekly amounts (hours) in playing practice predicts 38% of the variance, $F\,(1, 38) = 23.18$, $p <. 001$). The negative Beta weight associated with this regression model (standardized Beta = -.62, unstandardized Beta = -1.29) indicates that accuracy decreases with more time spent in weekly training for playing.

Expert Performance

The referee group's performance on the referee tackle assessment task was examined in relation to referee-specific experience and practice. Experience was measured by the total number of international matches refereed (youth and senior), years on FIFA, and the number of years refereeing. Practice variables were weekly time spent in referee practice and accumulated referee practice. These five variables were correlated with the accuracy measure using Pearson correlations.

Table 1 displays the results of the entire correlation matrix. Not surprisingly, number of years on FIFA was significantly correlated with number of international

matches refereed (r = .81) and years spent refereeing (r = .76). There was also a significant correlation between years on FIFA and accumulated time in refereeing practice (r = .80). For the performance variable, the only significant correlations were between accuracy and years on FIFA (r = -.79), and accuracy and number of international matches refereed (r = -.78).

Insert Table 1 about here

Based on these first order relations, we performed a multiple regression analysis for the performance variable. The independent variables entered into the regression represented practice and experience. The practice variables were weekly refereeing practice and cumulative refereeing practice. For the experience variables, although years on FIFA did show a significant correlation with accuracy, it was not entered into the regression. Years on FIFA was felt to be a problematic measure of refereeing expertise due to the many moderating variables that may influence assignment, such as political factors, the level of development of soccer within a country, and the number of FIFA referees allotted to represent a given country. Furthermore, the number of years spent on the FIFA list is highly related to the number of years spent refereeing altogether, as indicated by the significant correlation (r = .76). This same reasoning was used to eliminate number of international matches as a measure of experience. Thus, years refereeing was determined to be the most representative measure of experience within a group of referees. The resulting regression analyses used weekly referee practice,

cumulative referee practice, and years refereeing. None of these variables resulted in predictive regression models.

Soccer-General Task: Relative Offensive Threat

Perceptual Sensitivity (PĀ)

Refereeing and playing experience were again contrasted by investigating their contributions to performing the soccer task: relative offensive threat. Signal detection theory (Macmillan & Creelman, 1991) was used to assess performance. Specifically, key "threatening" players correctly identified were classified as "hits", whereas non-key players incorrectly identified were classified as false alarms. The hit and false alarm rates were combined to calculate PĀ, a non-parametric estimate of *d'* for small samples. PĀ was thus used to indicate perceptual sensitivity and skill in this task.

A Pearson correlation matrix explored relationships between playing and refereeing experience with PĀ as a measure of task performance. Table 2 displays the results of these comparisons. Given that the player group had no refereeing practice and no refereeing experience, the correlations between these variables are not surprising, and simply confirm that the two groups can be distinguished using these measures. What is important to note, however, is the lack of significant correlations between key player identification (PĀ) and any of the other variables. That is, none of the variables that distinguish group membership are related to performance on this task. Indeed, a multiple regression using Group as the independent variable and PĀ as the dependent variable results in an R^2 of only .02, $F(1, 39) = 0.88$, $p = .36$.

Insert Table 2 about here

Comparing performance on this task by group shows that referees and players are very similar, with high mean PĀ scores of .74 (sd = .04) and .76 (sd = .05), respectively, with perfect detection indicated by a score of 1.00. Furthermore, a t-test shows that these scores are not significantly different, t (39) = .94, p = .35. Taken together, these findings show that, not only do the groups perform similarly; the high PĀ scores show that they both perform quite well.

Ranking of Relative Threat

To score the rankings of players by degree of threat, five points were awarded for the correct ranking of the first-ranked player, four points for the correct ranking of the second-ranked player, and so on. As well, one point was deducted for each ranking away from the correct ranking. For example, a player incorrectly ranked second instead of first resulted in a deduction of one point, and a score of four out of a possible five. For each trial, participants were given a ranking score expressed as a percentage of the total possible points for that trial. Table 3 shows Pearson correlations between performance using this measure and the group membership variables. Similar to findings using PĀ, there were no significant correlations between any of the refereeing or playing experience/practice variables and this performance measure. Regression analyses show that group membership was not predictive of ranking scores. Not surprisingly, then the

referee and player groups do not differ in this measure as confirmed by the lack of significant results using a t-test, t (39) = .68, p = .50.

Insert Table 3 about here

Discussion and Conclusions

When compared to players, referees showed superior performance in a referee tackle assessment task. When compared on a more general soccer task of identifying and ranking relative threat of players, however, there were no differences between the referee and player groups. On one hand this is not surprising, since both groups had equivalent playing experience. With referee expertise research focused on the first step of the expert performance approach, however, this finding has important implications for task design. Indeed, one of the most challenging aspects in this area of study is the creation of appropriate laboratory tasks. In their meta-analysis of expert sport research, Thomas, Gallagher and Lowry (2003) indicate that in order to elicit expert-novice differences, researchers need to use laboratory tasks with high ecological validity. Our findings suggest that considerations of ecological validity should include considerations of role. Indeed, we found that knowledge and performance in soccer refereeing (tackle assessment task) is specific to role. In contrast, role does not have an impact on performance in a more general soccer skill (relative offensive threat task).

The results of this study and its consideration of role also have implications for what we know about the development of knowledge within a domain. Only two previous

studies have addressed whether skill is a constituent of expertise or a result of observational learning within a domain (Williams & Davids, 1995; Allard, Parker, Deakin & Rodgers, 1993). Both studies found that skill is an essential characteristic of expertise, and not a byproduct of exposure. Said another way, "what you do" within a sport related to role, influences "what you know" in terms of perceptual-cognitive skill. Our findings support this position. Although players are constantly exposed to refereeing and refereeing decisions, the results of this study suggest that this exposure does not contribute to skill for a referee-specific task (tackle assessment). In fact, our findings indicate that playing experience in the form of weekly practice actually interferes with the ability to perform this task, perhaps because time spent playing takes away from the time available to train as a referee.

The role-specific nature of knowledge and skill in sport has important implications for individuals making the transition from one role into the other. Indeed, in a study of figure skating, Allard et al. (1993) found that, while experience as a competitor *contributes* to the ability to judge figure skating, this experience by itself is not sufficient for high level judging performance. Although being an athlete in a sport may contribute to skill as a judge, the accumulation of formal training and specific judging experience are necessary to develop skill and eventually expertise.

Not only are we able to make general recommendations for referee training, we have identified the referee tackle assessment task as a potential training tool. Indeed, with referee accuracy at 80.6%, and player accuracy at 55.1%, this task was free from any floor or ceiling effects. This indicates that the referee tackle assessment task is a tool

adaptable to different performance levels. For example, adjusting the temporal features of the clips to create higher ecological validity may create a tool for skill refinement in elite level referees. Indeed, as discussed, Thomas, Gallagher and Lowry (2003) indicate that ecological validity is important in eliciting expert-novice differences. They have also found, however, that ecological validity gains importance as the skill of the performer is increased. Similarly, research into the use of simulators for skill training advises that greater fidelity or realism of simulations is needed for operators or trainers with greater skill (Hays & Singer, 1989; Alessi, 1988; Allerton, 2000; Andrews, 1988). While adjusting temporal features would represent a move towards greater realism of the task, it must be acknowledged, that increasing simulator fidelity is a difficult task, made more complex by the different types of fidelity that can be addressed (e.g., psychological, physical). On the other hand, another important feature of this tool is its low cost and relative ease of access, as well as its adaptability to different "key" infractions (e.g., hand balls).

Our findings provide a direction for future research addressing laboratory and training tasks for refereeing skill. This work represents an important advancement of the knowledge related to expert refereeing as well as the acquisition of role-specific knowledge and skill in sport.

References

Allard, F., Parker, S., Deakin, J., & Rodgers, W. (1993). Declarative knowledge in skilled motor performance: Byproduct or constituent? In J.L. Starkes, & F. Allard (Eds.), (1991). Cognitive issues in motor expertise (pp. 95-108). Amsterdam: Elsevier.

Allerton, D.J. (2000). Flight simulation: Past, present and future. Aeronautical Journal, 104 (1042), 651-663.

Andrews, D.H. (1988). Relationship among simulators, training devices, and learning: A behavioral view. Educational Technology, 28 (1), 48-54.

Alessi, S.M. (1988). Fidelity in the design of instructional simulators. Journal of Computer-Based Instruction, 15, 40-47.

Chase, W.G., & Simon, H.A. (1973). Perception in chess. Cognitive Psychology, 4, 55-81.

Ericsson, K. A., Krampe, R. T., & Tesch- Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. Psychological Review, 100, 363-406.

Ericsson, K. A., & Smith, J. (1991). Prospects and limits in the empirical study of expertise: An introduction. In K. A. Ericsson and J. Smith (Eds.), Toward a general theory of expertise: Prospects and limits (pp. 1-38). Cambridge: Cambridge University Press.

Hays, R.T. & Singer, M.J. (1989). Simulation fidelity in training system design. Springer-Verlag.

Cuypers, K. (2003). Decision-making and skill training in top-class soccer referees: Tackling the perceptual-cognitive dimension. Unpublished master's thesis, Katholieke Universiteit Leuven, Leuven, Belgium.

Helsen, W.F., & Starkes J.L. (1999). A multidimensional approach to skilled perception and performance in sport. Applied Cognitive Psychology, 13, 1-27.

Hodges, N.J., & Starkes, J.L. (1996). Wrestling with the nature of expertise: A sport specific test of Ericsson, Krampe and Tesch- Römer's (1993) theory of "deliberate practice". International Journal of Sport Psychology, 27, 400-424.

Macmillan, N. A., & Creelman, C. D. (1991) Detection theory: A user's guide. Melbourne, Aus.: Cambridge.

MacMahon, C., Helsen, W.F., Starkes, J.L., & Weston, M. (submitted). Deliberate practice in elite soccer referees. Journal of Sport and Exercise Psychology.

Nevett, M.E., & French, K.E. (1997). The development of sport-specific planning, rehearsal and updating of plans during defensive youth baseball game performance. Research Quarterly for Exercise and Sport, 68, 203-214.

Starkes, J.L. (1987). Skill in field hockey: The nature of the cognitive advantage. Journal of Sport Psychology, 9, 146-160.

Starkes, J. L., & Deakin, J. (1984). Perception is sport: A cognitive approach to skilled performance. In W. F. Straub & J. M. Williams (Eds.), Cognitive sport psychology, (pp. 115-128). Lansing, NY: Sport Sciences.

Thomas, J.R., Gallagher, J. & Lowry, K. (2003). Developing motor and sport expertise: Meta-analytic findings. Presentation at the conference of the North American

Society of Psychology of Sport and Physical Activity (NASPSPA), Savannah, Georgia, June 2003.

Ward, P, & Williams, A.M. (2003). Perceptual and cognitive skill development in soccer: The multidimensional nature of expert performance. Journal of Sport and Exercise Psychology (25)1, 93-112.

Williams, A.M., & Davids, K. (1995). Declarative knowledge in sport: a byproduct of experience or a characteristic of expertise? Journal of Sport and Exercise Psychology, 7 (3), 259-275.

Acknowledgments

This research was partially funded by a SSHRC grant to Janet Starkes, and a grant from F-MARC, the medical branch of FIFA, awarded to Werner Helsen. Support was also provided by a SSHRC doctoral fellowship awarded to Clare MacMahon. The authors would also like to thank Paul Ward and Mark Williams for the loan of testing materials.

Table 1.
Correlations Between Experience and Performance on the Referee Task: Tackle Assessment.

	Accuracy	Weekly Ref Practice	Total Ref Practice	Yrs Referee	Matches
Weekly Ref Practice	-.32				
Total Ref Practice	-.68	.43			
Yrs Referee	-.24	-.17	.42		
Matches	-.78*	.08	.59	.57	
Yrs FIFA	-.79*	.14	.80*	.76*	.81*

* $p < .05$

Table 2.
Correlations Between Experience and Identification of Key Players in Soccer-General Task: Relative Offensive Threat.

	PĀ	Weekly Ref. Practice	Weekly Play Practice	Years Refereeing
Weekly Ref. Practice	-.14			
Weekly Play Practice	.10	-.92**		
Years Refereeing	-.12	.91**	-.95**	
Years Playing	-.11	-.33*	.27	-.25

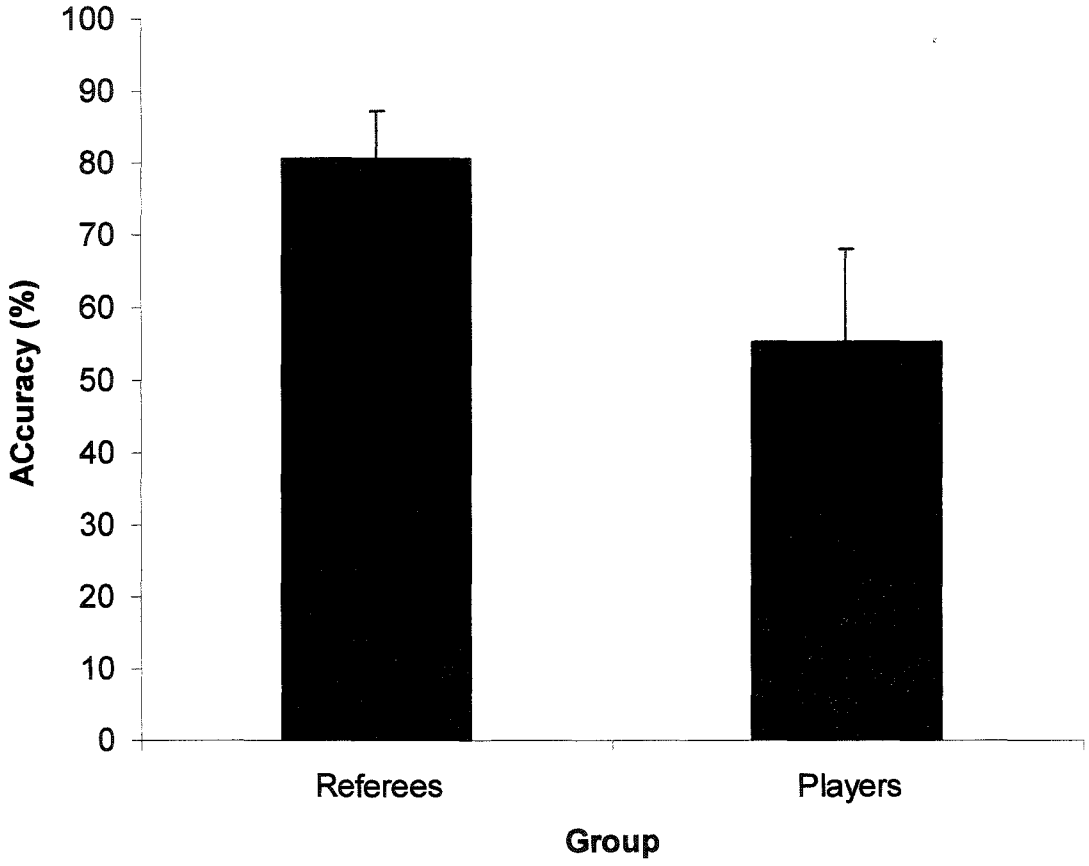
* $p < .05$
** $p < .01$

Table 3.
Correlations Experience and Ranking of Key Players in Soccer-General Task: Relative Offensive Threat.

	Ranking	Weekly Ref. Practice	Weekly Play Practice	Years Refereeing		Years Playing
Weekly Ref. Practice	.11					
Weekly Play Practice	-.15	-.92**				
Years Refereeing	.14	.91**	-.95**			
Years Playing	-.09	-.33*	.27	-.25	-.11	1.00

* significant at .05 level
** significant at .01 level

List of Figures


100
90
80
70
60
50
40
30
20
10
0
ACcuracy (%)
Referees
Players
Group

PAPER 3

This manuscript has not yet been submitted for publication. I was the major contributor to every aspect of this project including design, data collection and analysis and write-up.

Running Head: REFEREE DECISION-MAKING SKILLS

Capturing referee skill in the laboratory: Video-based decision making and the influence of priming

Clare MacMahon
Janet L. Starkes
Dept. of Kinesiology
McMaster University
Hamilton, Ontario, Canada

Janice Deakin
Dept. of Physical Education
Queen's University
Kingston, Ontario, Canada

Corresponding Author:
Clare MacMahon
Dept of Kinesiology
McMaster University
1280 Main St. West
Hamilton, Ontario
L8S 4K1
Fax: (905) 523-6011
macmahc@mcmaster.ca

Abstract

This paper outlines two studies of expertise in basketball refereeing from the expert performance approach. The first study replicates and extends Allard, Parker, Deakin, and Rodgers (1993) by testing groups of players, coaches and referees in three separate video-based decision-making tasks designed to correspond to the demands of each role (playing, coaching, refereeing). A re-test one week later also measured recognition of clips and recall of decisions. The results indicate that movement/action anticipation is a referee-specific skill, and play recognition is a coach-specific skill. For performance in the referee task at time 2, coaches and referees were influenced differentially by simultaneously performing clip recognition. This finding is linked to medical diagnosis literature and the influence of shifts in visual processing style on novice and expert performances. Given that study 1 differed to Allard et al. in finding only weak evidence of referee superiority in the referee task, we conducted a second study testing the influences of knowledge priming versus attentional priming on task performance. While knowledge priming did not influence performance, attentional priming appears to have had a detrimental impact. Implications for performance of visually based decision-making tasks are discussed. The results of these two studies raise a number of areas for future research such as the impact of individual clip/item selection as well as sequencing of clips on task performance.

Introduction

The complexity of sport, wherein performers must contend with both cognitive and motor demands has made it a ripe domain for the study of expertise. The majority of sport expertise literature has addressed learning and performance in athletes (e.g., Abernethy, Neal, & Koning, 1994; Starkes, 2000; Farrow & Abernethy, 2003), and to a lesser extent the coach (e.g., Côté, Salmela, & Russell, 1995; Salmela, Draper, & Laplante, 1993; Salmela & Moraes, 2003). The role of the referee, however, is virtually ignored (e.g., Janelle & Hillman, 2003). Yet the referee occupies an extremely complex and unique role, with both high cognitive and motor demands. Indeed, not only are referees required to possess a large store of declarative knowledge of the rules and laws governing a game, they are also responsible for enforcing the rules throughout the competition using the appropriate judgment, positioning, and signals. In light of the crucial role that referees play in sport contests, examinations of expert performance will add to our overall knowledge of elite sport functioning. In addition, studies of expert sport refereeing will advance the general understanding of human performance in complex psychomotor tasks.

Given the emphasis of sport expertise research in the past, we know a great deal about expert athlete performance, and very little about expert referee performance. This prompts the question: is there a difference between the skills of expert players, coaches and referees from one sport? Phrased another way, can athletes and coaches gain enough refereeing experience from observational learning to perform at the same level of skill as a referee, or does sport skill vary according to the role occupied (i.e., player, coach, referee)?

While the expertise and cognitive literature show us that superior knowledge is specific to the domain of expertise (e.g., Chase & Simon, 1973), or that doing and knowing are linked, the unique task demands in the roles of player, coach and referee raise consideration of the *role specificity* of knowledge within a given domain. If the tasks, or "doing" of referees differ from those of coaches and athletes, then skill characteristics and knowledge base ("knowing") should also differ. The purpose of this paper is to investigate sport refereeing using the expert performance approach (e.g., Ericsson, & Smith, 1991). In addition, our focus upon referee skill and characteristics also considers the impact of sport role and the distinction of "referee-specific skills" from skills related to coaching and playing.

Only two previous studies (Allard, Parker, Deakin, & Rodgers, 1993; Williams & Davids, 1995) have addressed knowledge varying as a function of task requirements, with only one including referees as a population of interest. Both studies have shown, however, that knowledge is specific to role, or "what you do", and a constituent of skill rather than a byproduct of exposure to the sport. The first of these studies, Allard, et al. (1993) addressed role-based skill differences by examining basketball officials, coaches and players on 5 different tasks. The tasks consisted of: 1) a test of general basketball knowledge, 2) a rules test, 3) a test of FIBA hand signals, 4) a recall test of pictures of both random and actual plays, and 5) a foul/violation detection and naming test. Results showed that performance, and by inference knowledge associated with these tasks, was a function of the expert's role within basketball. Specifically, officials performed best on the referee-specific tasks (rules, signals, and naming of fouls and violations), coaches on the coaching-relevant task (recall accuracy for real versus random plays), and players on recall of plays regardless of whether they were

real or random. The authors concluded that there are "...different types of declarative knowledge required by experts having different roles within a particular skill domain." (p. 106).

Williams and Davids' (1995) study is the second and only other research work available which addresses knowledge and skill as a byproduct of exposure or constituent of expertise. Williams and Davids examined the soccer knowledge of high-skill players, low-skill players, and disabled spectators. All three groups had comparable amounts of experience in soccer; however, the disabled spectators had acquired this experience by spectating and had never played soccer. In contrast, the players had gained the majority of their soccer experience through playing, with comparably less time spent spectating. Using these three groups, then, physical playing experience (players versus disabled spectators) and expertise (high versus low skill players) were both controlled. All three groups of participants completed tests of anticipation, recall, and recognition, with full screen video presentations of soccer plays used as stimuli.

In the anticipation test, in which participants predicted the destination of a soccer pass, both player groups were faster than the disabled spectators, and not significantly different from each other. Players anticipated prior to the ball kick, whereas the disabled spectators responded after ball contact. With regards to the accuracy of anticipation, high-skill players showed lower response error than both low-skill players and disabled spectators, who did not differ significantly.

For recall of structured game plays, Williams and Davids' (1995) high-skill players showed fewer recall errors than low-skill players, who were in turn superior to the disabled spectators. The domain specificity of this skill was confirmed by the lack

of group differences in recall of unstructured film clips[2]. The authors concluded, "...the experienced high-skill players have developed an extensive soccer-specific knowledge base that enables them to recognize meaningful associations between players' positions in game situations (p. 267)". Thus, for high-skill players, the ability to recall structured game play is a product of expertise, and developed in response to the task requirements that they encounter when playing.

Finally, in Williams and Davids' (1995) recognition task, the high-skill players were faster and more accurate than both the low-skill or disabled spectator groups in recognising previously viewed clips. As well, the low-skill players and disabled spectators did not differ significantly from each other. Overall, then, this study provides support for the idea that there are different roles within a particular sport that require different skills and/or knowledge bases.

Given that this paper targets referee skill in the laboratory, a study by Trudel, Dionne, and Bernard (2000) provides relevant insight into performance measures by considering that a critical feature of refereeing skill is consistency of decisions from one time to the next. Indeed, Trudel et al. showed that referees are more consistent than coaches, players, and parents when assigning penalties for ice hockey film clips. In this study, participants made decisions for film clips (e.g., penalty, no penalty) in two separate testing sessions three weeks apart. Trudel et al. showed that referees were more likely to make the same decision twice in response to the same clip,

---

[2] The authors acknowledge the possibility, however, that the composition of the unstructured clips was less complex, making them easier to recall, and decreasing the probability of error.

whereas players, coaches, and parents were more likely to change their minds. Referees were thus shown to be more reliable or consistent across time.

The first study in this paper compared performance on sport tasks between players, coaches and referees. Three separate video-based tasks were designed to mimic the demands in playing, coaching and refereeing. More specifically, this study partially replicates Allard, Parker, Deakin and Rodgers (1993), with additional considerations. For example, given the findings in Williams and Davids (1995) we incorporated both reaction time and accuracy measures in each task. As well, Trudel, Dionne and Bernard's (2000) findings with regards to decision consistency prompted us to include a time 2 testing session for all tasks. The time 2 measures are a unique aspect to this study, since apart from Trudel, Dionne, and Bernard (2000), no other researchers have provided test-retest reliability measures for video-based decision-making data. The second testing session also allowed us to measure the *consistency* of decisions when viewing the same stimuli on two separate occasions. Furthermore, measures of recognition for test clips were assessed, as well as recall of decisions. These measures allowed us to determine if explicit memory influenced decision-making in the second testing session.

Study 1

Method

Participants

The participants for this study consisted of three groups: basketball players, coaches and officials. The profile of participants provides a cross-section of individuals with primary experience in one of these roles, but often with experience in one or even both of the other two roles. However, the *current* role being fulfilled,

regardless of past experience, was the primary factor used to group participants. There were 12 participants in each group (players, referees, coaches), for a total of 36 participants overall. Table 1 displays the experience of participants in each group. The player group consisted of male and female varsity basketball players with an average of 10.6 years playing experience (sd = 3.4) and an average age of 21.8 (sd = 1.9). The players had only 2.1 mean years of coaching experience (sd = 3.5) and virtually no refereeing experience. The coach group was comprised of primarily high school coaches with a mean age of 36.7 (sd = 9.5), and 16.4 years accumulated playing experience (sd = 9.5). While this playing experience is greater in number than that of the player group, coaches were substantially older, with their current emphasis on coaching rather than playing. This group had coached for an average of 10.8 years (sd= 5.3), and refereed an average of 1.5 years (sd = 2.3). The referee group was composed primarily of university men's league referees with an average of 14 years' playing experience (sd = 8.9) and only 2.1 years coaching (sd = 4.4). The average age of this group was 44.3 (sd = 6.3). Like coaches, the referees' most current emphasis was refereeing, and they had accumulated an average of 20.3 years (sd = 7.9) in this role.

Insert Table 1 about here

Materials

Participants were tested using digital video clips in four separate tasks. The first task was a simple decision task general to the sport of basketball. This was used to familiarize participants with the general display and response features of the

remaining "target" tasks. Three tasks followed, with each crafted to mimic the demands of the respective sport roles of player, coach and referee. All of the tasks used video clips of basketball game footage. The video clips for task one, (warm-up task), task two (player task), and task three (coach task) were adapted from a study of perceptual training in basketball (Starkes & Lindley, 1994) and thus had the same temporal and spatial features. That is, clips in these tasks were all wide-angle game clips from Canadian University league women's games displayed in normal time.

Task four (referee task) video clips were adapted from the Allard, Parker, Deakin and Rodgers' (1993) referee-based task requiring the detection and naming of fouls and violations. These clips differed both spatially and temporally to those used in the first three tasks. Referee task clips were taken from NCAA game footage, showing a closer camera angle. As well, these clips were often played in slow motion with action freezes and re-starts. All of the clips used in the experimental tasks (player, coach, referee) were pre-tested for the correct responses by expert raters in basketball. Only those clips with 100% agreement across evaluators were used.

Participants were seated at a table in front of an IBM laptop computer, which ran the digital video clips. At the beginning of each video clip an internal auditory signal triggered a timer that was then stopped by participants' verbal responses. A headpiece was used to capture verbal responses, with a microphone in front of participants' mouths. This procedure allowed a measure of decision time. Participants were provided with a pencil and response sheets to further record responses and indicate their level of confidence in their choice and the degree of difficulty of each item (both rated on a five point Likert scale with one labelled *"low"* and five labelled *"high"*).

Procedure

Participants were first asked to read and sign the information and consent form. If there were no concerns expressed at this point, the study continued, otherwise, any issues were addressed and questions answered at this time. Each task was explained before its initiation. All participants began the study with completion of task 1, the warm-up task. This was followed by the three role-related tasks: playing, coaching and refereeing. The order of completion of tasks two to four was randomized to control for any between-task effects.

Task 1: Warm-up/ Decision time (DT) task. The purpose of this task was to provide participants with a warm-up to the study and familiarity with the general display and response demands for the three target tasks. This also provided a measure of basic decision time, ensuring that there were no group differences. For the DT task, participants were asked to view videotape of game footage and to indicate the first time that they saw the ball passed from one player to another with the verbal marker of "now". They were asked not to anticipate the pass, but to react to it as quickly and as accurately as possible, with the voice reaction timer measuring decision time. There were 12 clips used in this task.

Task 2. Player task: Optimal offensive move. In this player task, participants were asked to identify the next optimal offensive move following the cut-off point of a video clip. Tape occlusion in these clips occurred coincident with the ball handler's initial movement to execute the optimal move. An audio tone at the beginning of the clip initiated the voice reaction timer, which was then stopped once the participant spoke. After this, participants were asked to indicate their response on the answer form as a confirmation, along with their level of confidence in that response, and how

difficult they felt the task to be for that clip. Confidence and difficulty were measured on a Likert scale from 1 (*low*) to 5 (*high*). No feedback was provided for performance on these trials. Participants completed four lead-in trials for familiarization, and then 13 test trials.

Task 3. Coach task: Offensive formation identification. Task three incorporated the same videotape format as task two, with a play sequence cut off before its completion. Once again, participants' decision time and accuracy were measured. In this coach task, however, participants were asked to identify the offensive formation in the video footage. Where participants felt the play had transitioned from one formation to another, they were asked to identify the formation at the cut off point, and not prior. Written responses were recorded, along with measures of confidence and difficulty. Four lead in clips were used for familiarization, followed by 15 test trials. Participants received no knowledge of results during or after performance of the task.

Task 4. Referee task: Foul/violation detection. In the fourth task, taken from Allard et al. (1995), participants viewed game play and indicated whether a foul/violation had occurred. They were asked to either identify the foul/violation as quickly as possible, or to indicate "no call."[3] Again, decision times were recorded for detection. For decision time, however, only those items correctly identified as containing an infraction were used (hits). "No call" decisions, or correct rejections, in

---

[3] This procedure differs slightly to that used originally in Allard, Parker, Deakin and Rodgers (1995). In the original task, participants were asked to indicate whether a foul or violation was present first, and then to name the foul or violation where detected. Speed of responses was not recorded, nor was there any apparent time-pressure. Our method asks participants to detect and name simultaneously. We felt this procedure would be more naturalistic, and a closer replication of real game refereeing performance.

contrast to "call" decisions (hits) necessitate a delayed response while participants watch the entire clip. In this "wait and see" approach, participants may make a "no call" decision but delay verbal response until the entire action has been viewed and the absence of an infraction is confirmed. It follows, then, that the time taken to verbalize "no call" decisions does not reflect decision time.

Participants also rated their level of confidence in their responses as well as the difficulty of the items. There were four lead-in clips, and 16 test trials, with a maximum of 11 items for the reaction time measure based on a maximum of 11 possible "hits". There was again no knowledge of results during or after performance of the task.

All four tasks were repeated one week after initial testing. Each task contained the same test and lead-in items as in time 1, with an additional four new items that were not seen at time 1. These new items were placed randomly within the time 1 sequence of clips. New clip sequences for time 2 were thus created, with the same new sequence of clips used for all participants. Participants followed the same procedure as in time 1. In time 2, however, participants were also asked if the item was "new" (never seen before), or "old" (seen in time 1). Participants were also given the option of indicating that they did not know whether the item was new or old. For the player, coach and referee tasks, when participants judged an item as "old", they were then asked to recall whether their time 2 decision was the "same" as or "different" than the decision at time 1. For each clip and task, these recognition and recall judgments were made after the main decision (e.g., optimal offensive move). Participants were also instructed to focus upon the "central" decision for each clip and to treat the recognition and recall decisions as secondary. Testing at time 2 began with task 1

(DT/warm-up), with the remaining tasks presented in a randomized order different to the order of presentation at time 1 for that participant.

Analysis

Performance in all three target tasks (player, coach, referee) was assessed using a number of measures. The first measure, decision time (DT), was calculated using the data recorded from the voice reaction timer. Response time was measured as the time taken from the start of the video clip until the participant's response. The clip length was then subtracted from this value to provide a measure of the amount of processing time and length of stimulus viewed before verbalization of responses (DT). This calculation allows capture of responses made before the end of the clip, which are indicated by a negative DT value. DT was calculated for every trial, with averages and standard deviations then calculated for each task.

Response accuracy was measured as the percentage of correct responses. For the referee foul/violation detection task, a foul or violation response was "correct" only if it was both detected, and then named correctly. That is, if an offensive foul was detected as a foul but erroneously labelled a defensive foul, no point was awarded, and this was deemed an incorrect response. For those items that did not contain an infraction, a no call response was awarded one point, and any other response (i.e., foul or violation) was awarded a 0.

In addition to this measure of accuracy, we applied signal detection theory (SDT) (Macmillan & Creelman, 1991) by labelling responses as one of four possibilities: 1) hit, 2) miss, 3) false alarm, or 4) correct rejection. These designations were based solely on the detection of fouls/violations without consideration to the naming component of the task. Therefore, a clip containing a foul or violation that was

detected (i.e., participant indicates foul/violation) was classified as a "hit", a foul/violation for which the participant indicated no call was labelled a miss, a foul/violation response for a clip with no infraction was labelled a false alarm, and a clip with no infraction for which the participant indicated "no call" was a correct rejection. We then used the hit rate, calculated as the percentage of actual foul/violation clips that were detected, and the false alarm rate, calculated as the percentage of no call clips for which a false alarm response was made. The hit and false alarm rates were then used to calculate PĀ, an estimate of *d'* for small sample sizes, and a measure of perceptual sensitivity.

Performance in the referee task was also measured by participants' ability to correctly name the detected infractions. By way of example, if 10 infractions were detected, the percentage of these infractions correctly named was the measure of naming. Taken together, hit rate and PĀ were used to indicate the ability to detect an infraction and overall perceptual sensitivity, the naming measure was used to indicate the ability to classify detected infractions, and the accuracy measure was used to indicate a participant's ability to *both* detect and name an infraction (and thus included measurement of correct rejections).

Altogether, the main infraction detection measures used were DT, accuracy, hit rate, false alarm rate, PĀ, and naming. This performance assessment differs somewhat to that used by Allard et al. Indeed, Allard et al. assessed detection (comparable to hit rate in this study) and naming (comparable to naming in this study). We have additionally incorporated signal detection measures, as well as an overall accuracy measure encompassing both hit rate and naming. This assessment provides a more detailed, and in some instances more conservative indicator of performance.

For all tasks, outliers were removed for the DT and accuracy measures to eliminate data from trials in which participants may have missed the action (e.g., due to lack of attention). Two standard deviations above the average was used to identify DT outliers and two standard deviations below the average was used to identify accuracy outliers. Outliers accounted for less than 2% of the data overall.

Three additional measures were derived solely from the time 2 testing sessions. First, recognition was measured as the percentage of video clips correctly identified as new or old. This measure was used for all tasks, including the DT/warm-up task. Second, for clips identified as old, participants were asked to indicate whether the decision made at time 2 was the same or different as at time 1. The percentage of correct same/different responses provided a measure of recall. Third, the decisions made at time 1 and time 2 were compared to indicate the consistency of decisions, with the same decision at both times for a given clip awarded a 1 and a different decision awarded a 0. This resulted in a percentage score for consistency. The recall, recognition, and consistency measures were calculated only for the player, coach, and referee tasks.

Results

Each task was examined separately using analyses of variance to compare performance between groups (players, coaches, referees) and between testing sessions (time 1, time 2). Any significant interactions were tested using Tukey's post hoc testing procedure. We also examined relationships between different types of experience (years spent playing, coaching, and/or refereeing) and performance on each task. In order to guide these analyses for each task, we first explored which experience variables predict membership in player, coaching and refereeing groups.

This resulted in the finding that a model using years refereeing predicts 63% of the variance, $R^2 = 0.63$, $F(1, 34) = 58.7$, $p = <.001$. For each task, the relationship between experience variables and performance was first explored for first order relations using Pearson correlations, and then followed with tests of forward regression models.

Task 1: Warm-up/Decision time (DT) task

The warm-up/DT task was used to familiarize participants with the basic visual display and the voice reaction timer. It also served to ensure that the groups did not differ in basic decision time responses to video displays. Indeed, this was the case, with a 3 Group (player, coach, referee) x 2 Time (1, 2) mixed ANOVA revealing no differences by Group, $F(2, 31) = 0.78$, $p = 0.47$), or by Time, $F(1, 31) = 1.41$, $p = 2.44$. Thus, performance in this task did not change over time or vary by group. A one-way ANOVA showed that there were also no group differences with regards to recognition of the clips used in this task, $F(2, 33), = .96$, $p = .39$, with mean percentages of correct new/old decisions at 73% (sd = 10%), 75% (sd = 10%) and 78% (sd = 9%) for the player, coach and referee groups, respectively.

Task 2. Player task: Optimal Offensive Move

A 3 Group (player, coach, referee) x 2 Time (1, 2) mixed analysis of variance was used to examine performance on the playing task with the DT, accuracy, confidence and difficulty measures. As illustrated in figure 1, a main effect of testing session for DT, $F(1, 31) = 4.16$, $p = 0.05$, showed that all participants were faster at time 2 ($M = 1.57$s, sd = 0.59) than at time 1 ($M = 1.78$s sd = 0.59).

Insert Figure 1 about here

For the accuracy of responses (in determining the next optimal offensive move), figure 2 illustrates a main effect of Group, F (2, 32) = 6.73, p = .004, with the accuracy rate of the referee group (M = 65%, sd = 15%) higher than that of the player group (M = 51.3%, sd = 17%), but not higher than the coach group (M = 54.7%, sd = 15%). There were no differences between players and coaches.

Insert Figure 2 about here

There were no significant differences by Group or Time for ratings of confidence or difficulty (Confidence: Group F (2, 33) = 1.19, p = 0.32, Time F (2, 33) = 0.8, p = 0.46. Difficulty: Group F (2, 33) = .08, p = 0.92, Time F (2, 33) = <0.001, p = 0.99.). Indeed, confidence ratings were quite high for all three groups, with player, coach and referee ratings averaged across both sessions resulting in ratings of 4.4 (sd = 0.4), 4.1 (sd = 1.2) and 4.4 (sd = 0.4), respectively (on a 5 point scale where 1 = *low* and 5 = *high*). In parallel with confidence ratings, difficulty ratings averaged across both sessions were quite low, with averages of 2.0 (sd = 0.5), 1.9 (sd = 0.6) and 2.1 (sd = 0.8) for players, coaches and referees, respectively.

For recognition of clips as old or new, there was a main effect of Group, F (2, 31) = 3.56, p <. 041. Post Hoc analyses show that coaches were more accurate at classifying clips as old or new (M = 77.7%, sd = 10.3%), as compared to players (M = 64.4%, sd = 17.3%). There were no differences, however, between referees (M

=74.8%, sd = 7.7%) and coaches, or referees and players. This finding is shown in figure 3. An interesting pattern to note is that when examining the standard deviations in accuracy for clip recognition, the player group varies the most (17.3%) followed by the coach group (10.3%) and then the referee group (7.7%). These differences in the ability to recognize clips did not translate into differences in the ability to recall decisions, as shown by the results of the ANOVA, F (2, 33) = 0.48, p = 0.62. Coaches did show the greatest accuracy in this task, however (M = 80.9%, sd = 9.7%) as compared to players (M = 78.3%, sd = 18.7%) and referees (M = 75.4%, sd = 10.8%).

Insert Figure 3 about here

The consistency of decisions from time 1 to time 2 was compared between groups using a one-way ANOVA with 3 levels (player, coach, referee). This analysis showed that players, coaches and referees did not significantly differ, F (2, 32) = 1.16, p = 0.33, with consistencies of 71.2% (sd = 11.3%), 72.7% (sd = 15.5%) and 65.6% (sd = 8.1%), respectively.

Final analyses of the player task explored relationships between performance and experience in each of the roles of playing, coaching and refereeing. Pearson correlations revealed a significant positive relationship between number of years spent refereeing and accuracy for both time 1 (r = 0.49, p< .01) and time 2 (r = 0.42, p < .05). Indeed, a forward regression analysis with accuracy at time 1 as the dependent variable resulted in an R^2 of 0.26, showing that years refereeing accounts for 26% of the variance in this variable, F (1, 34) = 11.91, p < .01. The same pattern was shown for accuracy at time 2, with a regression model using years refereeing resulting in an

R^2 of 0.14, F (1, 34) = 5.64, p= 0.023. At time 2, then, years refereeing accounts for 14% of the variance.

When looking at the consistency of decisions from time 1 to time 2 in this task, the regression model using years refereeing produces an R^2 of 0.12, F (1, 33) = 4.5, p = 0.042. Examination of the beta weights, however, shows that there is a negative relationship such that, as number of years refereeing increases, consistency decreases (Unstandardized Beta = -.005, Standardized Beta = -.346).

Task 3. Coach task: Offensive Formation Identification.

Mixed analyses of variance showed no Group, F (2, 32) = .57, p = .57, or Time differences, F (1, 32) = .37, p = .55, in DT for performing the coach task. With regards to accuracy of responses, there were no main effects of Group, F (2, 32) = 0.29, p = 0.75, or Time, F (1, 32) = 1.64, p = 0.21. There was an interaction, however, between Group and Time, F (2, 32) = 3.60, p = .04. Tukey's Post Hoc analyses show that the coach group was more accurate at time 1 (M = 51.5%, sd = 6.7%), than at time 2 (M = 44.8%, sd = 5.9%). As well, at time 2, the referee group's performance was significantly more accurate (M = 51.1%, sd = 8.8%) than the coach group's performance (M = 44.8%, sd = 5.9%). These results are illustrated in figure 4. For their part, while not different either from time 1 to time 2, or when compared to the other groups, player accuracy in the coach task was 49.4% (sd = 7.8%) at time 1, and 49.0% (sd = 6.6%) at time 2.

Insert Figure 4 about here

With regards to ratings of confidence, there were no differences by Group, F (2, 33) = 0.65, p = 0.53, or Time, F (2, 33) = 1.00, p = 0.38. Ratings indicated that participants were very confident, with averages across both sessions of 4.1 (sd = 0.4), 4.1 (sd = 0.6), and 3.9 (sd = 0.6) for players, coaches, and referees, respectively. Difficulty ratings were also similar between the three groups, F (2, 32) = 1.48, p= 0.24, and the two testing sessions, F (2, 32) = 0.36, p = 0.70. Average ratings of difficulty for both sessions were 2.2 (sd = 0.5), 2.1 (sd = 0.7) and 2.5 (sd = 0.9) for players, coaches, and referees, in order.

The three groups did not differ in recognition of clips, F (2, 27) = 0.32, p = 0.73, nor in recall of decisions, F (2, 27) = 2.2, p = 0.13, with the average number of clips recognized at 61.1% (sd = 20.4%), 66.3 % (sd = 13.1%) and 65.4% (sd = 14.8%) for players, coaches and referees, and the average recall scores at 81.2% (sd = 14.7%), 75.3% (sd = 16.6%) and 68.9% (sd = 16.2%) for the same groups. Finally, the consistency of decisions in this task also did not differ by group, as shown by the results of the one-way ANOVA, F (2, 32) = 2.33, p =0.11. Players made the same decision at both sessions on 68.1% of the clips (sd = 15.4%), coaches on 72.7% (sd = 15.5%), and referees on 59.6% (sd = 10.7%).

Pearson correlations examining accuracy at the coaching task showed no significant relationships with experience in playing, coaching, or refereeing. An examination of regression models, however, showed that years refereeing accounts for 19.5% of the variance in consistency of responses for the coaching task, F (1, 34) = 8.21, p = 0.007. Examination of the Beta weights shows that this relationship is negative, however, (Unstandardized Beta = -.006, Standardized Beta = -.441).

Task 4. Referee Task: Foul/Violation Detection and Naming.

Decision time for the referee task did not show any main effects of group. Figure 5 shows that, when comparing performance between time 1 and time 2, however, the differences approached significant levels, $F(1, 32) = 3.93$, $p < .056$, with all groups performing this task faster at time 1 ($M = 1.46$ s, sd =1.07) than at time 2 ($M = 1.80$, sd = 1.06). With regards to accuracy of decisions for the refereeing task, there were no main effects of Group or Time. As well, when time 1 data are examined alone, a one-way ANOVA with Group as the three level factor shows no significant differences, $F(2, 33) = 2.04$, $p = 0.15$. On the other hand, the interaction between Group and Time approached significant levels, $F(2, 32) = 2.88$, $p = 0.071$. Tukey's Post Hoc analyses on the interaction show that at time 1, the referee group ($M = 66.5\%$, sd = 7.2%%) is more accurate than the coach or player groups ($M =$ 59.9%, sd = 12.5%; $M = 58.5\%$, sd = 13.8%, respectively). As well, the referee group is more accurate at time 1 ($M = 66.5\%$, sd = 7.2%) than at time 2 ($M = 61.3\%$, sd = 10.4%). In contrast, coaches show the opposite pattern with superior accuracy at time 2 ($M = 62.5\%$, sd = 10.5%) as compared to time 1 ($M = 58.0\%$, sd = 12.5%), with these differences approaching significant levels (critical value = .048, test value = .045) These findings are displayed in figure 6.

Insert figure 5 and figure 6 about here

Signal detection statistics were used to compare different facets of performance on this detection task. 3 Group (player, coach, referee) x 2 Time (1, 2) mixed ANOVAs were used to test for any differences in the hit rate, false alarm rate,

and PĀ. Analysis of the hit rate showed no differences between groups or time (Group: F (2, 33) = 1.29, p = 0.29; Time: F (1, 33) = 1.55, p = 0.22), with hit rates averaged across the two sessions at 68.5% (sd = 17.6%), 74.6% (sd = 14.2%) and 80.7% (sd = 14.3%) for players, coaches and referees, in order. The false alarm rate, however, showed no differences by Group, but a main effect of Time, F (1. 33) = 5.14, p = .03, with a greater number of false alarms at time 1 (M = 23.9%, sd = 19.6%) than at time 2 (M = 18.3%, sd = 18.7%) for all participants. These differences are shown in figure 7. Although the decreases in false alarm rates indicate improvements in sensitivity, these differences may be due primarily to the large variability in this measure. This may thus explain the finding that changes in false alarm rates did not have an impact on PĀ, as shown through the lack of significant ANOVA effects for Group, F (2, 33) = 0.21, p = 0.81, or Time, F (1, 33) = 0.80, p = 0.38. Averaged across both sessions, PĀ values were 0.78 (sd = 0.1), 0.76 (sd = 0.1) and 0.78 (sd = 0.1) for players, coaches, and referees, respectively.

Insert figure 7 about here

The ability to correctly name infractions, indicated by the "name" measure did not differ between groups, F (1, 33) = 0.42, p = 1.00, or testing session, F (1, 33) = 0.37, p = 0.55, with average naming accuracies of 69.3% (sd = 17.4%), 67.9% (sd = 13.9%) and 68.8% (sd = 12.4%) for players, coaches and referees, respectively. Similarly, ratings of confidence in task performance were once again high, and did not differ by Group or Time, with players, coaches and referees indicating average ratings of 4.2 (sd = 0.6), 4.0 (sd = 0.9) and 4.5 (sd = 0.4), respectively. For ratings of

task difficulty, the difference between time 1 and time 2 ratings approached significant levels, F (1, 33) = 3.36, p = .076), with all participants rating infraction detection more difficult at time 1 (M= 2.2, sd = 0.9) than at time 2 (M = 2.0, sd = 0.8). For recognition of referee clips, the one-way ANOVA comparing the three groups showed no differences, F (2, 33) = 6.37, p = 0.7 with correct clip classification at 77.1% (sd =10.4%) for players, 77.8% (sd = 10.3%) for coaches, and 80.6% (sd = 10.9%) for referees. There were similarly no group differences in recall of decisions made for these clips ("same", "different"), F (2, 32) = 0.74, p = 0.48, with accuracy rates of 71.6% (sd = 13.6%) for players, 73.3% (sd = 15.0%) for coaches, and 77.3% (sd = 11.0%) for referees. Groups were also similar in the consistency of their decisions between testing session, F (2, 33) = 0.61, p = 0.55, with the proportion of consistent decisions 73.7 % (sd = 11.7%), 68.3% (sd = 12.9%) and 70.5% (sd = 11.6%) for the player, coach and referee groups, respectively.

Pearson correlations and regression analyses showed no relationship between experience variables and performance indicators for the refereeing task.

Discussion

This study compared basketball referees, coaches and players in tasks crafted to match the demands specific to each role. A warm-up (DT) task was also included to familiarize participants with the general use of video clips for time-pressured decision-making. A second testing session (time 2) was completed to measure the consistency of responses, and allow a test of data reliability. During time 2, recognition of clips and recall of decisions from time 1 were measured. An important point to note is that these additional time 2 tasks may have had an influence on performance of the main target decisions, regardless of the fact that participants were

instructed to treat them as secondary. This addition of tasks and subsequent creation of a dual-task situation at time 2 is an important factor in interpreting our findings. The results from each task are discussed in the following sections.

Tasks 1 and 2: Warm-up/DT and Player Task

As expected, there were no differences in performance on the warm-up/DT task either between the groups of players, coaches and referees, or between testing sessions. Recognition of clips at time 2 also did not differ between the groups.

In the player task, wherein participants indicated the next best offensive move, there was a main effect of Group for the accuracy measure, with referees providing more accurate responses than players. Referees were not more accurate than coaches, however, and coaches did not differ from players.

A closer examination of the player task helps explain and interpret these findings. Recall that this task was adapted from a decision-making training study involving semi-skilled female basketball players (Starkes & Lindley, 1994). The mean age of the participants used in the original study was 15.5 years, and they had accumulated an average of only 3-4 years of playing experience. In the original task, participants were asked to indicate the optimal offensive move from three choices: dribble, pass, or shoot. In an attempt to add greater complexity to the task and reduce the possibility of correct responses due to chance, we chose to elaborate the number of choices by adding directional options to the dribble and pass choices (e.g., dribble left, dribble right, or dribble ahead). It is possible that this alteration created a mismatch between the relative simplicity of the stimuli and the complexity of response options.

An additional possibility in accounting for the accuracy results in the player task is that players, coaches and referees differed in their interpretation of the question: "what is the next best offensive move?" Indeed, a fundamental difference between referees and players in deciding on the next best offensive move is consideration of the motor components of the choice. That is, from a player's point of view, deciding on the next move relies on knowing what one is capable of doing ("what can I do?") as well as what one's teammates are capable of doing ("what can they do?"). Indeed, Nevett and French (1997) have shown that baseball players' strategic game plans are influenced by their motor skills. As well, high-skill tennis players make modifications of their action plans based on situational information (McPherson & Kernodle, 2003). In the case of the player, then, a response choice is complicated by consideration of experience-based knowledge specific to the motor components of various actions. This in turn may result in players overlooking the "correct response" due to its lack of complexity outside the context of their experience.

In comparison to players, referees are not concerned with actually performing the next best offensive move, but rather with anticipating the upcoming action. Indeed, refereeing requires anticipation of the next play in order to move to an appropriate viewing position. Given this demand, referees may reinterpret the question as: "*what is most likely to occur next*?", a comparatively less complex decision uninfluenced by considerations of the motor components of the choice and the ability to carry them out. In effect, then, this reinterpretation may result in referees performing a movement anticipation task. In fact, within the clips used in the player task, the next best offensive move was always the next move performed by the

athletes after the clips were occluded. Correctly choosing the optimal offensive move can thus be equated with correctly anticipating the next movement. Seen in this light, accuracy as it was measured in this task may represent the ability to anticipate movements, a skill that is essential for refereeing. In this context, it is not surprising that referees were superior to players in this task.

Coach accuracy in indicating the next best offensive move was not significantly different from that of either players or referees: coaches were "as good" as referees, but were also "as bad" as players. Again, this finding can be explained by examining the features of the task relative to the role-related demands of coaching. Like referees, coaches do not have to carry out the motor components of an action choice, eliminating what was discussed as a potentially problematic mismatch between the stimuli and response choices of this task. In this sense, then, it is not surprising for coaches to perform similarly to referees. On the other hand, it may be far less important for coaches than referees to anticipate upcoming movements. As well, while the *most current and recent experience* of the coaches in this study was in a coaching role, this group had accumulated an average of 16.4 years' playing experience (as compared to the 10.6 years of the player group). Coaches had thus accumulated extensive experience carrying out motor choices and in doing so, making choices based on their own capabilities. This experience may have been essential in creating coach performances similar to those of players. In summary, it is possible that in the coach group, the combined influence of coach demands and extensive playing experience influenced performance, resulting in no significant differences from either the player or referee groups.

Another interesting finding for the player task is that coaches were superior to players in recognition of clips at time 2. When we consider that in order to indicate the next best offensive move, participants were required to follow the development of plays, this finding is consistent with previous work. Indeed, Allard, Parker, Deakin and Rodgers (1993) found that coaches showed the biggest gain when recalling real plays as compared to recalling random plays. Similarly, expert coaches were shown by Garland and Barry (1991) to have superior recall and recognition of game plays as compared to their novice counterparts. There are therefore indications that clip recognition is a task that taps in to coaching-specific skill. The lack of differences between referees and coaches in clip recognition may reflect that this ability also contributes somewhat to referee skill. Future work to explore these possibilities is needed.

Task 3: Coach Task: Offensive Formation Identification

The coach task required participants to identify the offensive formation shown in game clips. At time 2, participants again performed clip recognition and decision recall. Although the clips used in this task had the same temporal and spatial features as those used in the player (optimal offensive move) task, changing the decision-making component and focus of processing altered the results. Indeed, at time 1, the lack of significant group differences in identifying offensive formation indicates that this task is not coach-specific, but rather a basketball-general task. That is, individuals with experience in basketball, regardless of role, are relatively proficient at this task, due to the fact that this skill contributes to some degree to the demands of each role.

At time 2, however, when coaches were asked to both identify the offensive formation and recognize clips, the accuracy of their performance declined

significantly, both relative to time 1 and to a level below that of the referee group. Recall the proposition made in discussing the player task (task 2: optimal offensive move) that recognition of clips is a coaching-specific skill. If this is the case, then at time 2 of the coaching task, participants were required to complete both a basketball-general (offensive formation identification) and a coach-specific (clip recognition) skill. For coaches, then, performance at time 2 can be explained in the context that performing both a general basketball task (identify offensive formation) and a coach-specific task (clip recognition) puts a strain on attentional and processing resources, resulting in inferior performance on both tasks. As players and referees have limited coaching experience and coach-specific processing, the same strain is not shown. Thus, in the time 2 dual-task condition, coaches suffered performance disruptions of both tasks, performing only equally as well as the other groups in clip recognition, and showing worse offensive formation identification relative to time 1.

Taken together, then, although the player and coaching tasks incorporated the same style of game clips, changing the decisions being made may change performance when participants occupy different sport roles.

Task 4: Referee Task: Foul/Violation Detection

The referee task incorporated video clips that were spatially and temporally different to those used in the previous three tasks. Recall that these clips were played in slow motion and often included freeze frames with restarts of the action. One of the most interesting findings from this task comes from the accuracy of detecting and naming infractions where the Group x Time interaction approached significant levels. Post hoc analyses of the interaction showed that referees were superior to both other groups at time 1, and then suffered performance decrements at time 2, where group

differences disappeared, and referees were significantly less accurate than at time 1. In contrast, coach performance showed a pattern of improvement at time 2 relative to time 1. To provide a possible explanation of these findings, we have once again targeted the clip recognition task at time 2.

Given the slow motion and close visual angle of the camera in the referee-task clips, it is possible that performing a clip recognition task at time 2 prompted a feature-based analysis of the display. That is, in order to decide whether the clip was old or new, participants may have analysed individual clip features such as the players and teams in the clips (as indicated by the school name and player name on the shirt), as well as the particular movements performed by each player in the clip. To determine whether the clip is old or new, these features would then be compared to what was stored in memory from time 1. Given that the clips are brief, and can only be processed once, this feature-based visual analysis would then also be used for the infraction detection component of the task.

The literature within medical diagnosis in visual domains can be used to explain our findings. Indeed, this literature indicates that performers at different stages of expertise use different processing styles. Changing, or emphasizing these styles may have an impact on task performance. Schmidt, Norman, and Boshuizen (1991) propose four stages that learners pass through in acquiring skill for medical diagnosis. In the early stages of development, learners make diagnoses by using elaborate propositional networks of information. As they gain expertise, however, diagnosis becomes more efficient, with clinicians moving from an analytic, feature-based processing style to a more holistic style of processing, with diagnosis based on similarity to previous experiences (Kulatunga-Moruzi, Brooks, & Norman, 2001). In

sum, novices use feature-based analysis, processing the separate components of the display, while experts use a non-analytic, holistic processing style in which "the whole picture" is taken in to account.

As mentioned, we propose that, at time 2 of the referee task, the clip recognition component, in concert with the close angle and slow motion of the clips may have elicited a novice, feature-based style of analysis, which was then also applied to the infraction detection task component. With an average of 1.5 years of refereeing experience, this feature-based analysis improved performance for the coach group, allowing them to access and apply the refereeing knowledge that they had developed to that point. Indeed, there was a pattern of decreased false alarms at time 2 for the coach group, indicating that the way they performed this task changed. No similar benefits were seen for the player group, given that they had no refereeing experience or training at all, and thus no knowledge base to access.

In contrast, referees showed performance decrements from time 1 to time 2. For the referee group, then, the feature-based processing style, prompted by the recognition task, not only reverted them to a novice analysis of the display, but also resulted in novice levels of performance.

Not only are changes in processing *style* related to the accuracy for the "expert" (referee) and "novice" (coach) groups across both testing sessions, but the changes in the *speed* of processing may have compounded this effect. Recall that performance at time 2 was slower than at time 1, likely due to the extra processing demands. Research in visual domains of medical diagnosis has found that, in experts, longer viewing times are associated with more errors (e.g., Kundel, Nodine, & Carmody, 1978; Norman, Brooks, Allen, & Rosenthal, 1989). Furthermore, this line

of research has shown that tasks with no time restrictions do not produce expertise effects, whereas those incorporating time pressure through only brief exposures to stimuli result in performance by experts superior to that of novices (Norman, Coblentz, Brooks, & Babcock, 1992). Although participants were always reminded to perform the task "as quickly and accurately as possible", the addition of a clip recognition component may have resulted in slower performance at time 2 than at time 1, interfering in expert/referee infraction detection.

If, in fact, participants in the referee task are prompted to use a feature-based analytic style at time 2, which then influences novice and expert performers in different ways, this has important implications for the acquisition and performance of cognitive skills. Within expert-novice research targeting movement skills, experts show performance decrements when they focus on the task they are performing (task-focus). In contrast, novice performers in the same task-focused condition show performance enhancements (Beilock, Carr, MacMahon & Starkes, 2002). The borderline findings in this study open the door to future investigations of the acquisition of cognitive skill and the attentional and processing styles at different levels of expertise.

An alternative explanation for the interaction findings in the referee infraction detection and naming task is that the use of video clips fails to elicit the processing characteristics or problem representations that referees typically use in game situations. This laboratory task may thus have created a new type of expertise in solving the task. This possibility highlights the necessity of future research using verbal protocol data to examine problem representations and knowledge base.

Conclusions

Using video clips of game play as stimuli, study 1 tested players, coaches, and referees in tasks designed to replicate the respective role-related demands of each group. In addition, a warm-up task was used to provide experience with the basic perceptual demands of visually processing the clips. At time 2, the same tasks were repeated, with participants asked to decide if the clips were old or new, and if the decisions for the old clips were the same or different as at time 1.

The groups expected to excel at the given tasks did not always do so. For example, referees were more accurate than players in the "player" task. These findings helped us to more accurately identify role-related skills. For instance, we proposed that game action anticipation is a skill central to refereeing demands, and play recognition is a skill central to coaching demands.

Two important findings highlight the complexity of role-related differences in visual processing of sport information. First, the player and coach tasks incorporated the same type of clips. Yet, asking participants to process this information in different ways (i.e., to perform different tasks) changed the relative performance strengths of each group, and thus the results. Second, the results of the referee task indicate that both the display features of the clips (e.g., speed, angle of view) and the inclusion of a second visual processing task (i.e., clip recognition) have differential effects on expert and novice decision-making.

The findings in study 1 contribute to our understanding of role-based differences in sport performers. Keeping in mind the more specific goal of advancing our understanding of expertise in refereeing, however, we were surprised at the absence of stronger evidence for referee superiority in infraction detection, in light of

the results of Allard et al. (1993). We conducted a follow-up study to further investigate this particular finding. Study 2 thus focused on a more in-depth exploration of potential factors influencing task performance in infraction detection.

Study 2

The goal of study 1 was to investigate refereeing expertise from the perspective of role-based differences as a component of skill. Given this overarching goal, it was surprising that we found only weak evidence for referee superiority at a referee-specific task. This surprise is heightened given that we used a task from Allard et al. (1993) who found stronger evidence of referee superiority. Specifically, Allard et al. found no differences in the ability to detect infractions (a measure we have defined as hit rate), but superiority in naming infractions (our naming measure). While we also found no differences in detection (hit rate), in contrast to Allard et al., we did not find any differences in naming. More specifically, in study 1, accuracy, a relatively conservative measure that combined the ability to both detect and name infractions, revealed a borderline interaction between Group and Time, but no main effects. Furthermore, when this measure was examined for only time 1, referees were not superior to coaches and players. Given these results, study 2 focused upon the referee infraction detection task and the potential impact of task design in eliciting expert-novice, or role-based differences.

When examining the method used in Allard et al. (1993) more closely, we identified one potential influence on performance in the infraction detection task. That is, in Allard et al., participants completed a basketball rules test, and a referee signals test prior to completing the infraction detection task. This order of task completion may have resulted in a priming of referee knowledge before completion of the

infraction detection task. Given that the referee group has a more in-depth refereeing knowledge base than the coach or player groups, this may have facilitated superior referee performance. This study thus tested whether completion of the rules and signals tasks prior to completing the infraction detection task has an impact on performance.

The results of Cuypers (2003) highlight another potential source of priming for performance in laboratory tests of perceptual-cognitive referee decision-making. Cuypers investigated the assessment of soccer tackles using film clips of soccer game-action. Specifically, groups of referees and players viewed video clips of soccer tackles and decided on the appropriate call (i.e., no foul, free kick, free kick followed by a yellow card, free kick followed by a red card). The referee group showed superior accuracy in this task. Whereas the participants in both study 1 and Allard et al. indicate that infraction assessment based on video film clips is a perceptually difficult task, Cuypers' participants had the advantage of knowing what they were looking for. That is, participants knew that every clip would show a tackle situation that would then be evaluated. This knowledge served to direct attention and potentially prepare participants' search strategies. With referees possessing superior knowledge of relevant cues for this decision, this "infraction priming" may have helped performance. In the current study 1, referees were not directed to the same extent. A second manipulation in study 2 was thus to assess the impact of directing participants' attention. To do this, we instructed one group of participants to focus on one particular type of infraction, and then evaluated whether this manipulation results in superior detection for that type of infraction. A real-world priming analogy would

occur when one team's coach asks an official to "keep an eye out for" holding by the other team.

This study investigated two conditions for infraction detection: priming through a rules and a signals test, and priming through instructions to focus on one particular type of infraction. In order to investigate the impact of these priming conditions, we created two sets of clips from the original 50 clips of the Allard et al. study (i.e., 1-25 = set 1, 26=50 = set 2).

There are practical implications of investigating the two priming conditions in this study. For instance, if completing rules and signals tests prior to infraction detection improves performance, this implies that referee game-performances will benefit from not only physical, but also perceptual-cognitive warm-up. If instructions have an impact on infraction detection performance, this has implications for pre-game referee meetings, as well as coach and captain behaviour during games (e.g., asking a referee to focus on a particular type of infraction).

Method

Participants

Two separate groups of 11 basketball officials were recruited for a total of 22 officials. The groups were similar in average accumulated years refereeing experience, with one group averaging 14.9 years (sd = 10.3) and the second averaging 14.6 years (sd = 11.1). A t-test comparing number of years of experience confirmed that the groups did not differ in this measure, t (20) = 0.06, p = 0.95. The first group was designated the knowledge primed (KP) group, and the second was designated the infraction primed (IP) group. The different manipulations for these groups are explained below.

Materials

We used three of the referee tasks from Allard, Parker, Deakin and Rodgers (1993), having obtained materials from the authors. The first test assessed knowledge of the rules of basketball using short answer, fill-in-the-blank and true/false questions. The second test assessed knowledge of basketball refereeing signals by asking participants to identify the signal being shown in a series of drawings. These two tests were paper and pencil tests. The third task was the infraction detection and naming task used in the previous study. For this study, however, all of the 50 original clips from Allard et al. were used, divided into two different sets of clips. Set 1 consisted of the first 25 clips from the original tape, and set 2 of the second 25 clips. Clips were displayed on a large screen using an LCD projector. Responses were recorded on answer sheets.

Procedure

Both referee groups completed all 3 tasks (rules, signals, infraction detection). The rules and general knowledge tasks were pencil and paper tests in which participants answered written questions in short answer, true/false, or yes/no format. In the case of the signals test, participants identified the picture in the space below it. The method used in the foul/violation detection task differs slightly from that used in study 1. In a procedure closer to that used in Allard et al., participants were required to indicate whether there was a foul or violation in the clip, or no call warranted. Where a foul or violation was detected, participants were asked to name the offence (e.g., violation: goal tending; defensive foul: push). There was no time pressure in performing this task, and answers were only recorded on the answer sheet. For each set of clips, the first 3 clips were treated as lead-in items wherein participants became

familiarized with the procedure. This resulted in 22 clips within each set used for testing purposes.

Testing took place in two groups. The first group was designated the knowledge (KP) primed group, and the second group the infraction primed (IP) group. The two testing procedures used for these groups are depicted in Table 2. In the KP group, participants filled out a biographic information sheet indicating amounts of refereeing experience and then performed the foul/violation detection task using the first half of clips (set 1). After set 1, the rules, and signals tests were completed. Finally, the KP group completed the infraction detection task using set 2 of the clips, and the exact same procedure as for set 1.

In the infraction primed (IP) group, participants first completed biographic information sheets. Following this, the foul/violation detection task was completed using the set 2 clips. After a brief pause, the task resumed, using the set 1 clips. For this set of clips, however, the IP group was instructed to perform the task, but to pay particular attention to *defensive fouls,* making sure to detect all instances of defensive fouls. Finally, the IP group completed the rules and signals tests.

Insert Table 2 about here

The KP and IP groups were tested in their respective groups. They were instructed to refrain from consulting other group members, or calling out an answer. Both portions of the video-based infraction detection tasks were group-paced, whereas the rules and signals tests were self-paced.

Analysis

The rules and signals tests were scored by comparing participant answers with those provided by experienced referees operating in the absence of time constraints, with access to basketball rulebooks and colleagues. This scoring resulted in a maximum of 25 points for the rules test and a maximum of 10 points for the signals test. Scores were expressed as the percentage of correct answers.

Responses in the infraction detection task were coded in the same manner as in study 1, wherein a response was given one of four designations: 1) hit, 2) miss, 3) false alarm, or 4) correct rejection. Once again, then, the hit and false alarm rates were used to calculate PĀ as a measure of perceptual sensitivity. The ability to name infractions was also calculated in the same manner as in study1. An overall accuracy measure was calculated using the hits that were correctly named added to the correct rejections. Finally, the ability to detect defensive fouls was evaluated using hit and false alarm rates for defensive fouls. The defensive hit rate was calculated as the number of defensive fouls detected, whereas the defensive false alarm rate was based on the number of false alarms indicating defensive fouls as a proportion of the total possible false alarm responses. These two rates were combined to provide a PĀ specific to defensive fouls. This analysis represents a conservative measure of defensive foul detection.

We first compared the experimental measures between each group's control condition performances in order to verify that the level of expertise of the two groups was comparable. We then compared performances between conditions (experimental versus control) within each group (knowledge primed, infraction primed) to assess the impact of the different priming conditions.

If priming using the rules and signals tests improves performance (knowledge priming) then the total accuracy or sensitivity of performance in the KP group was expected to be superior for the experimental clips as compared to the control clips. If instructions to focus on one type of infraction has an effect (infraction priming) then the IP group was expected to show an improved PĀ only for defensive fouls. In fact, if priming is occurring, it was expected that performance on the experimental clips would show an increased hit rate, but also an increased false alarm rate, as participants shift their bias and criteria for indicating the existence of an infraction.

Results

Rules and Signals Tests

While the two referee groups were equated in number of years' experience, we compared performance on the rules and the signals tests, to ensure equal levels of basic rule-based basketball knowledge. The t-test results show no differences between the groups in performance on either the rules test, $t\ (20) = 0.46$, $p = 0.65$, or the signals test, $t\ (20) = 1.49$, $p = 0.15$. Furthermore, performance for both groups was highly similar to that reported by Allard et al. For instance, the KP and IP groups averaged 73.7% (sd = 6.6%) and 73.1% (sd = 10.5%) respectively on the rules test where Allard et al. report 77% correct for their referee group. The signals test showed even better performance with 90% (sd = 5.7%) and 92.9% (sd = 7.6%) of the pictures identified correctly for the KP and IP groups, in order. The officials in Allard et al. averaged 97% on this task. While the Allard et al. officials showed superior performance, it must be highlighted that Allard et al.'s group were national and international level officials, whereas the two groups in this study referee primarily high school level play. Considering these findings, the two referee groups do not

differ significantly in skill or experience, and in fact perform comparably to Allard et al.'s officials.

Filtering of Clips

In the initial design of this study, we assumed equivalent difficulty level across the set of video clips. Indeed, study 1 incorporated only 18 of the original 50 clips from Allard et al. (1993). In addition, consideration of balancing the clips designated as experimental versus control for both manipulations (KP, IP) necessitates the use of four groups rather than the two used here. Including two more groups, however, would require double the number of participants, a very difficult prospect given the limited regional population of experienced basketball referees. In order to address this potential methodological flaw, we examined performance for each set of clips during control conditions and filtered them in what we assume is an appropriate way.

Using accuracy, the most conservative measure, we examined control performance for each set of clips. We identified clips in each set that were either significantly "too easy" or significantly "too hard", and thus did not adequately discriminate skill. This process of eliminating the "easy" and "hard" clips from each sequence was expected to create testing sets more sensitive to experimental manipulations and resulting performance changes. That is, performances on "easy" clips and "hard" clips are not likely to alter. Within each control group, then, clips on which fewer than 3 out of 11 participants (27%) were accurate and on which greater than 9 out of 11 (81%) participants were accurate were identified and eliminated from the experimental analyses. This resulted in 16 clips in set 1 and 14 clips in set 2.

Having ensured that the groups were not different in experience or knowledge, and eliminated "outlier video clips", performance within each group was then compared between the control and experimental conditions. Control performance on the two sets of clips did not differ significantly on any measure (hit rate: t (20) = 0.77, p = 0.45; false alarm rate: t (20) = 0.12, p = 0.91; PĀ: t (20) = 0.46, p = 0.65; naming: t (20) = 0.92, p = 0.37; accuracy: t (20) = 1.61, p = 0.12).

Knowledge Primed Group

For the knowledge primed (KP) group, the clips in set 1 were used for the control condition and the clips in set 2 were used as the experimental condition. A one-way repeated measures ANOVA comparing hit rate between the two clip sets showed no differences, F (1, 10) = 2.27, p <. 16, with hit rates of 74.2% and 82.8% for the control and experimental clips, respectively. Similarly, the false alarm rate did not differ, F (1, 10) = 1.07, p <. 32, with a rate of 43.2% in the control condition and 31.8% in the experimental condition. Not surprisingly, then, the PĀ between the conditions also did not differ, F (1, 10) = 1.79, p <. 21 with a PĀ of 0.66 and 0.75 for the control and experimental clips.

The ability to name the detected infractions does not appear to have been altered through our knowledge priming manipulation, as shown in the lack of significant differences between the two sets of clips, F (1, 10) = 0.06, p <. 82, with 66.5% of the control clips named correctly and 67.9% of the primed clips. With accuracy based on the combination of detection and naming, it is also not surprising that there were no differences in this measure either, F (1, 10) = 2.04, p <. 18, with a rate of 51.1% accuracy for set 1 (control) and 60.4% for set 2.

Infraction Primed Group (IP)

Comparisons of performance in the control and experimental conditions show a significant decline in overall hit rate, $F(1, 10) = 5.3$, $p = 0.04$, from a rate of 79.1% for control clips to 67.3% for the experimental clips. This effect is illustrated in figure 8. There was no corresponding difference in false alarm rate, however, $F(1, 10) = 1.16$, $p < .31$, with rates of 41.8% and 30.3% for the control and experimental conditions, respectively. This lack of significantly different false alarm rates in the face of a difference of 11.5% is likely due to the large variability in these rates for both sets of clips (sd control = 23.4%, sd experimental = 29.2%). Although the hit rate differs between conditions, the variability in the false alarm rate may be having an effect in that there are no differences in PĀ, $F(1,10) < .001$, $p < .98$, with a PĀ of 0.69 for the control clips and 0.68 for the experimental clips. Though not significant, there is a pattern of declining accuracy, from the control to experimental conditions, $F(1, 10) = 3.96$, $p = 0.07$, with a 61.6% accuracy rate in the control clips dropping to 50.9% for the experimental condition.

Insert Figure 8 about here

In terms of the measures specific to detection of defensive fouls, once again, there was a significant decline in the hit rate, $F(1, 10) = 7.83$, $p = 0.02$, with a 59.1% rate of defensive foul detection for the control set (sd = 12.6%), and 49.4% detection for the infraction primed clips (sd = 13.0%). This difference is illustrated in figure 9. The false alarm rate specific to defensive fouls did not change, however, $F(1, 10) = 0.02$, $p < .90$, with rates of 24.8% (sd = 22.9%) for the control condition and 23.5%

(sd = 26.3%) for the experimental. Once again, standard deviations on this measure indicate that participants varied widely in their percentage of defensive false alarms. This variation likely led to the lack of significant differences between conditions in perceptual sensitivity (PĀ) for defensive fouls, $F(1, 10) = 0.54$, $p < .48$, with values of 0.67 (sd = 0.10) and 0.63 (sd = 0.17), for the control and experimental conditions, respectively.

Insert Figure 9 about here

Discussion

This study addressed factors with a potential influence on performance in a laboratory refereeing infraction detection task. We specifically tested the impact of two different types of "priming" on referee infraction detection and naming performance. These two manipulations did not have as large an impact as we anticipated. The first manipulation was based on exploring the possible confound in Allard et al. (1993) of completing a rules test and a signals test prior to completing the infraction detection task. To explore whether this procedure results in priming of declarative knowledge and thus superior performance in infraction detection for referees, we tested participants in a control condition and a knowledge primed (KP) condition. Analyses comparing performance in the two conditions showed no significant differences on any of the performance measures. In this test, then, it does not appear that knowledge priming took place or had an impact on infraction detection.

A second group completed infraction detection in a control and an "infraction primed" (IP) condition. For performance on the IP trials, the group was instructed to pay particular attention to detecting all defensive fouls. Comparisons between performance for the control and primed sets of clips indicated that priming might indeed have an impact on performance. However, the impact of the priming was not in the expected direction, with significant decreases for the primed set of clips in both the overall hit rate and the hit rate specific to defensive fouls. As well, there was a pattern of decreased accuracy in the experimental condition relative to the control condition. Thus, rather than improving performance, infraction priming appears to have interfered with performance.

That infraction priming might interfere with performance seems likely in the context of the perceptual difficulty and demands of the infraction detection task. That is, it was felt that if participants expected a particular type of foul (defensive foul) this would facilitate its detection. Indeed, in Cuypers (2003), all of the video clips showed soccer tackles: one type of action. Participants thus knew that clips would always show tackles, which would then be evaluated. In the infraction detection task used here, participants were faced with various actions and any number of potential infractions, including defensive fouls. This did not change in the infraction-primed condition. Thus, rather than directing attention or preparing participants for where to look, infraction priming instructions may have served to *distract* attention. When instructed to pay particular attention to defensive fouls, participants may have felt an increase in attentional demands, resulting in performance decrements.

Another possible explanation for performance decrements in the IP condition is also linked to attention. As discussed in study 1, studies in golf and soccer show

that expert performance is disrupted when a task focus is elicited (Beilock, Carr, MacMahon & Starkes, 2002). It is possible that instructing referees to pay particular attention to defensive fouls creates similar task-focused disruption. As we currently know very little about the impact of attentional focus on the performance of cognitive tasks, this idea is worth further investigation.

While attention demands and shifts may have resulted in performance decrements in the IP condition, there are also possible methodological explanations for the results of both the KP and IP groups, which cannot yet be ruled out. Indeed, more thorough control testing is warranted to explore the potential differences between the sets of clips used. While we attempted to minimize the influence of "bad testing items" by eliminating the least discriminating clips, there is still the possibility that the particular clips used as well as the sequence of clips within each set influenced performance. For example, we do not know at this point whether presenting four "no call" clips in a row influences decision-making, or if this factors in to referees' expectations. Thus, the fact that study 1 used only a subset of the clips from the original Allard et al. (1993) study may be the reason for its weaker evidence of referee superiority in this task relative to players and coaches. Similarly, the use of two different sets of clips for the primed and experimental conditions in study 2 may have influenced the results. Finally, when comparing study 1 to Allard et al., it is possible that our use of signal detection statistics and accuracy measures resulted in the use of more conservative measures of performance. The complexity of issues surrounding these studies certainly indicates this as an area worthy of continued investigation.

Overall, the results of this study indicate that referee performance in an infraction detection task using video clips may be very sensitive to task features such as the particular clips/items used and the task instructions. The potential of these factors to influence performance needs to be investigated more rigorously.

General Conclusions

The first study of this paper investigated whether referees display skills distinct to those of other roles within a sport. Our comparison of coaches, players and referees shows that there are, in fact, role-related differences in performance based on the task. Specifically, when the instructions and task change for film clips with the same format and perceptual features (e.g., angle, speed) there are different outcomes, with different groups showing strengths. For example, the findings indicate that anticipation of action is a referee-specific skill, whereas play/clip recognition is a coach-specific skill. This indicates that there may be role-related differences in the information extracted from a game display.

An especially interesting finding from study 1 was the pattern of referee superiority in a referee task only when the task was performed in isolation. When the same task was repeated with an additional demand (clip recognition), performance for the referee group declined, while coach performance improved. Although these differences are only borderline significant, we have proposed that the additional task created a change in processing style that facilitated novice (coach) performance and interfered with expert (referee) performance. Whether this is the case needs to be explored in more depth in future research.

The follow-up to study 1 focused upon the lack of strong evidence of referee superiority in the referee task. We proposed that priming participants before infraction

detection and naming would have an impact on performance. The results showed, however, that this priming did not have as large an impact as we anticipated. Specifically, study 2 shows that knowledge priming, or performing a rules and signals test prior to infraction detection has no significant impact on subsequent infraction detection performance. On the other hand, instructing participants to focus on detecting one particular type of infraction (defensive fouls) did appear to have an impact on subsequent performance. Unfortunately, we cannot rule out that these findings are due to differences between the two sets of clips.

In infraction detection performance for both study 1 and study 2, participants showed both large false alarm rates, and large variation in this measure. This may be due to the low number of "no call" clips in the stimuli used for both studies, resulting in an inflation of the mean false alarm rate for groups, with the possibility of no false alarms occurring in some participants. Regardless, participants in these two studies show varied strategies in making the no call versus foul/violation decision, with some making no false alarms at all. As well, participants in these studies as well as those in Allard et al. have indicated the perceptual difficulty of the task. That performance appears to change according to instruction (e.g., player versus coach task in study 1; IP group in study 2) and task demands (e.g., difference in performance at time 1 versus time 2 of study 1) indicates that not only do we need to investigate the visual features of the stimuli used in laboratory tasks to elicit expert refereeing skills (e.g., clip length, size of display, speed, angle of view), but we also need to devote consideration to the design and task that accompanies the stimuli.

Future research should investigate whether referees, players and coaches focus on different cues in a display and if information pick-up in each group changes

according to task demands. As well, controlled studies investigating the effects of priming and task design (e.g., using one category or type of infraction, specific instructions) will advance the study of laboratory skills in refereeing, and ultimately training tools.

References

Abernethy, B., Neal, R. J., & Koning, P. (1994). Visual-perceptual and cognitive differences between expert, intermediate, and novice snooker players. Applied Cognitive Psychology, 8, 185-211.

Allard, F., Parker, S., Deakin, J., & Rodgers, W. (1993. Declarative knowledge in skilled motor performance: Byproduct or constituent? In J.L. Starkes & F. Allard (Eds.), Cognitive issues in motor expertise. (pp. 95-108). Amsterdam: Elsevier.

Beilock, S.L., Carr, T.H., MacMahon, C., & Starkes, J.L. (2002). When paying attention becomes counterproductive: Impact of divided versus skill-focused attention on novice and experienced performance of sensorimotor skills. Journal of Experimental Psychology: Applied, 6, 6-16.

Chase, W.G., & Simon, H.A. (1973). Perception in chess. Cognitive Psychology, 4, 55-81.

Côté, J., Salmela, J.H., & Russell, S. (1995). The knowledge of high-performance gymnastic coaches: Competition and training considerations. The Sport Psychologist, 9, 65-75.

Ericsson, K. A., & Smith, J. (1991). Prospects and limits in the empirical study of expertise: An introduction. In K. A. Ericsson and J. Smith (Eds.), Toward a general theory of expertise: Prospects and limits. (pp. 1-38). Cambridge: Cambridge University Press.

Farrow, D., & Abernethy, B. (2003). Do expertise and the degree of perception-action coupling affect natural anticipatory performance? Perception, 32, 1127-1139.

Garland, D. J., Barry, J. R. (1991). Cognitive advantage in sport: The nature of perceptual structures. American Journal of Psychology, 104 (2), 211-228.

Janelle, C. M, & Hillman, C.H (2003). Expert performance in sport: Current perspectives and critical issues. In J.L. Starkes and K.A. Ericsson (Eds.), Expert performance in sports: Advances in research on sport expertise. (pp.19-48). Champaign, Ill.: Human Kinetics.

Kundel, H.L., Nodine, C.F., & Carmody, D. (1978). Visual scanning, pattern recognition and decision making in pulmonary nodule detection. Investigations in Radiology, 13, 175-181.

Kulatunga-Moruzi, C., Brooks, L.R., & Norman. G.R. (2001). Coordination of analytic and similarity-based processing strategies and expertise in dermatological diagnosis. Teaching and Learning in Medicine, 13(2), 110-116.

Cuypers, K. (2003). Decision-making and skill training in top-class soccer referees: Tackling the perceptual-cognitive dimension. Unpublished master's thesis, Katholieke Universiteit Leuven, Leuven, Belgium.

Macmillan, N. A., & Creelman, C. D. (1991). Detection theory: A user's guide. Melbourne Aus.: Cambridge.

McPherson, S.L., & Kernodle, M.W. (2003). Tactics, the neglected attribute of expertise: Problem representations and performance skills in tennis. In J.L. Starkes and K.A. Ericsson (Eds.), Expert performance in sports: Advances in research on sport expertise. (pp. 137-168). Champaign, Ill.: Human Kinetics.

Norman, G.R., Brooks, L.R., Allen, S.W., & Rosenthal, D. (1989). The development of expertise in dermatology. Archives of Dermatology, 125, 1063-1068.

Norman, G.R., Coblentz, C.L., Brooks, L.R., & Babcock, C.J. (1992). Expertise in visual diagnosis: A review of the literature. Academic Medicine, 67(10) October Supplement, s78-s83.

Nevett, M.E., & French, K.E. (1997). The development of sport-specific planning, rehearsal and updating of plans during defensive youth baseball game performance. Research Quarterly for Exercise and Sport, 68, 203-214.

Salmela, J.H., Draper, S. & LaPlante, D. (1993). Development of expert coaches of team sports. In S. Serpa, J. Alves, V. Ferreira, & A. Paul-Brito (Eds.) Sport psychology: An integrated approach. (pp. 296-300). Lisbon, FMH.

Salmela, J.H., & Moraes, L.C. (2003). Development of expertise: The role of coaching, families, and cultural contexts. . In J.L. Starkes and K.A. Ericsson (Eds.), Expert performance in sports: Advances in research on sport expertise. (pp. 275-294). Champaign, Ill.: Human Kinetics.

Schmidt, H.G., Norman, G.R., & Boshuizen, H.P.A. (1990). A cognitive perspective on medical expertise: Theory and implications. Academic Medicine, 65, 10, 611-621.

Starkes, J.L. (2000). The road to expertise: Is practice the only determinant? International Journal of Sport Psychology, 31(4), 431-451.

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

Trudel, P., Dionne, J.-P., & Bernard, D. (2000) Differences between assessments of penalties in ice hockey referees, coaches, players, and parents. In A. B. Ashare, (Ed.) Safety in ice hockey: Third Volume. (pp. 274-289). ASTM, PA.

Williams, A.M. & Davids, K. (1995). Declarative knowledge in sport: a byproduct of experience or a characteristic of expertise? Journal of Sport and Exercise Psychology, 7 (3), 259-275.

Acknowledgments

This research was partially funded by a SSHRC grant awarded to Janet Starkes, and supported by a doctoral fellowship awarded to Clare MacMahon. The authors would like to thank John Moroz for his technical support, as well as Guy Cipriani for his invaluable help recruiting participants.

Table 1.
Biographic Profile of Groups in Study 1.

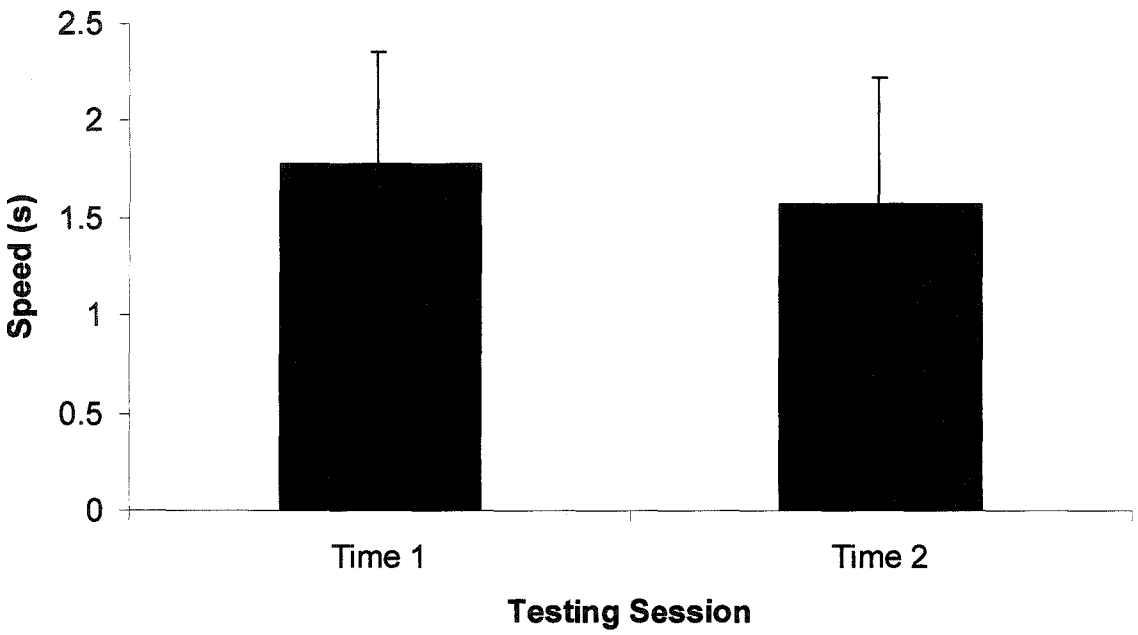
Group	Age		Years Played		Years Coached		Years Refereed	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Players	21.8	1.9	10.6	3.4	2.1	3.5	0.3	1.2
Coaches	36.7	9.5	16.4	9.5	10.8	5.3	1.5	2.3
Referees	44.3	6.3	14.0	8.9	2.1	4.4	20.3	7.9

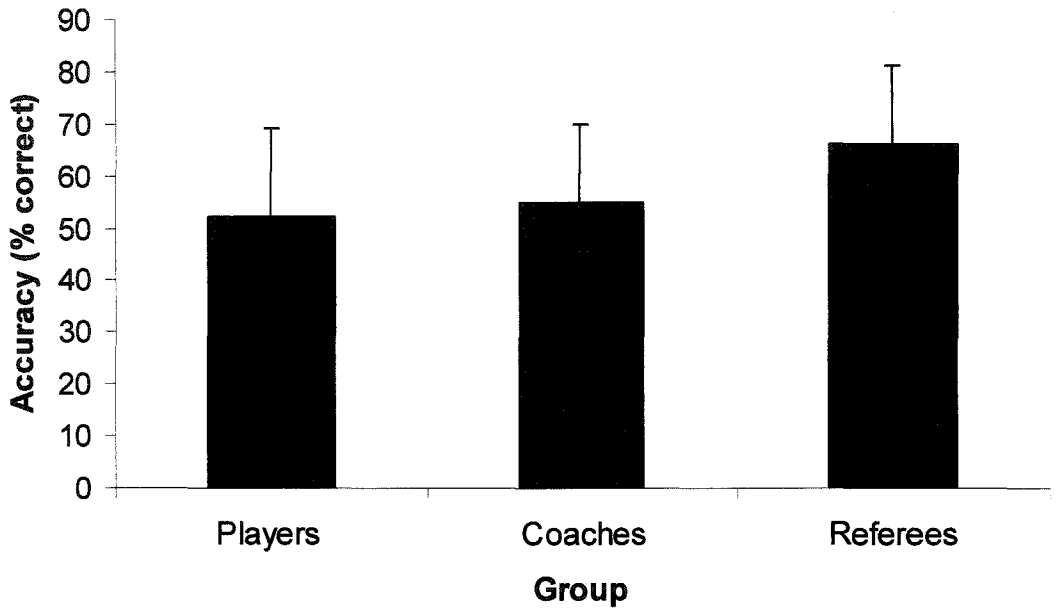
Table 2.
Study 2 Testing Procedure.

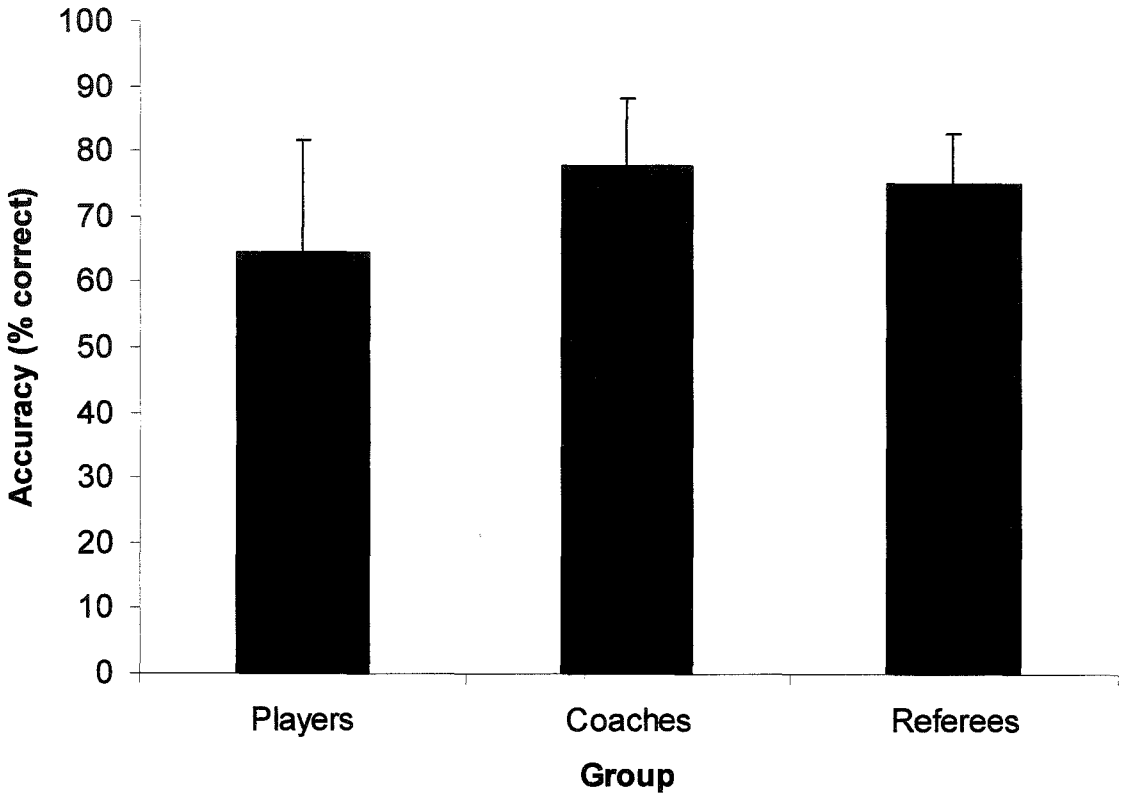
	Task order			
Group	1	2	3	4
Knowledge Primed	ID: set 1	Rules test	Signals test	ID: set 2
Infraction Primed	ID: set 2	ID: set 1*	Rules test	Signals test

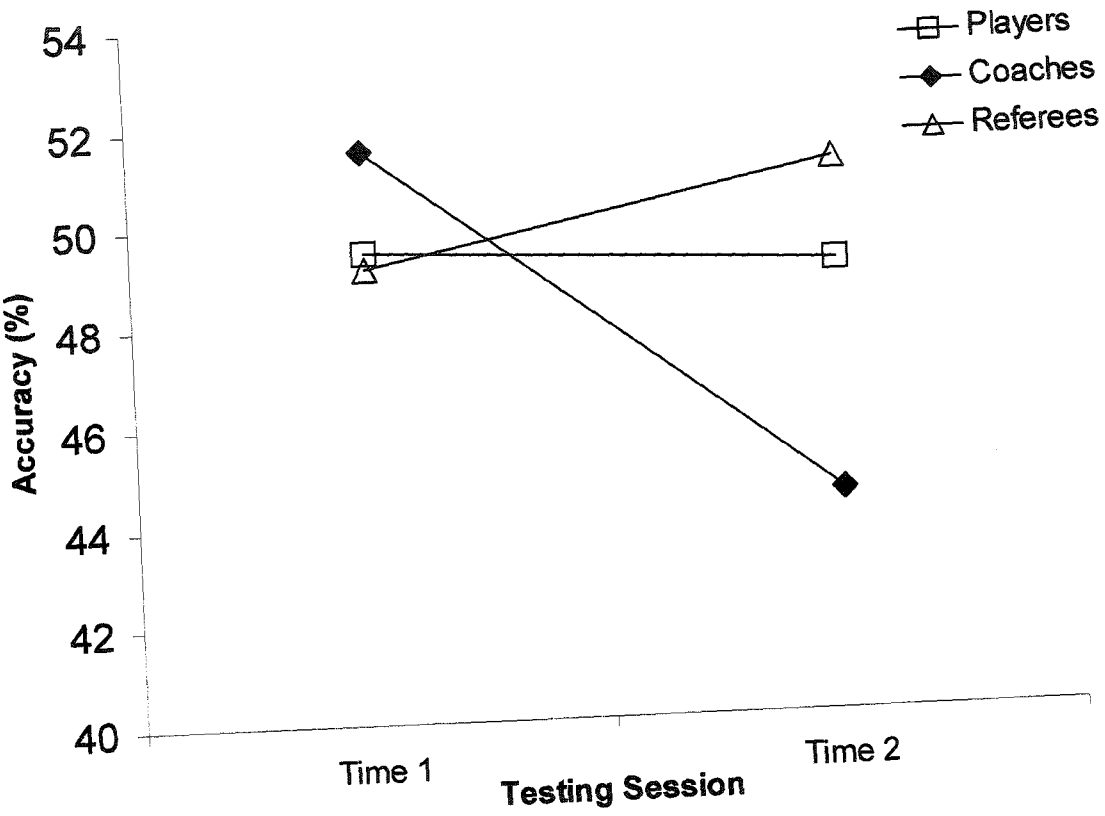
Note: ID = infraction detection
* instructed to focus upon defensive fouls

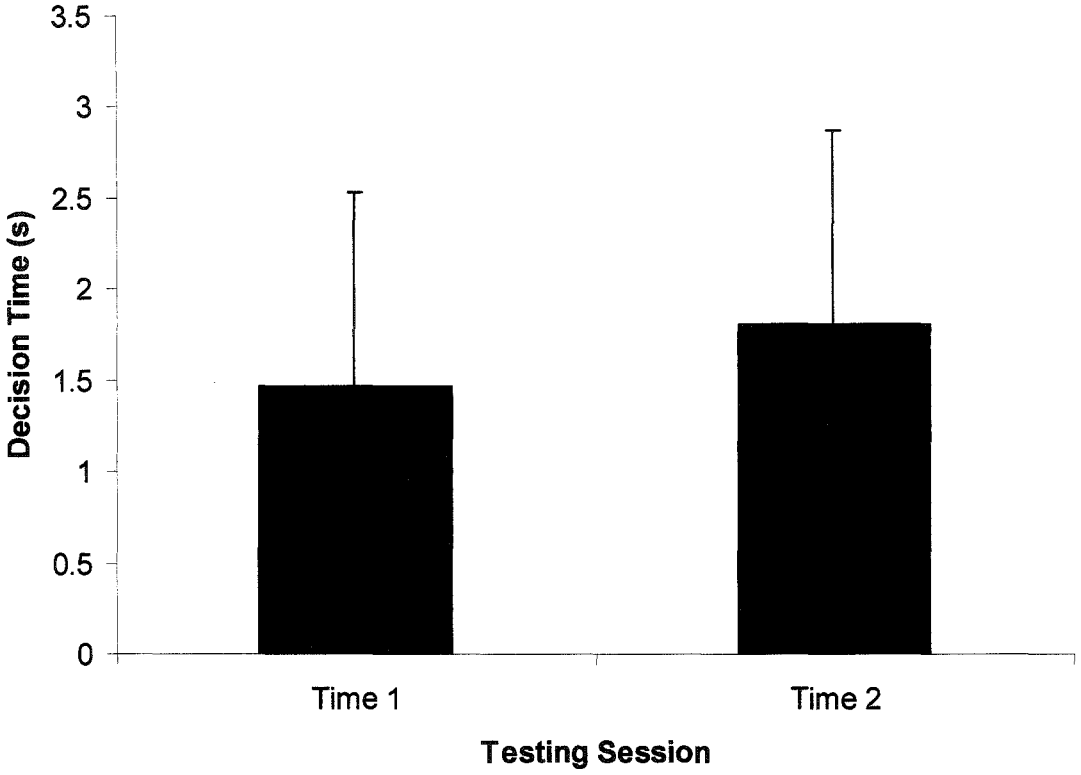
List of Figures

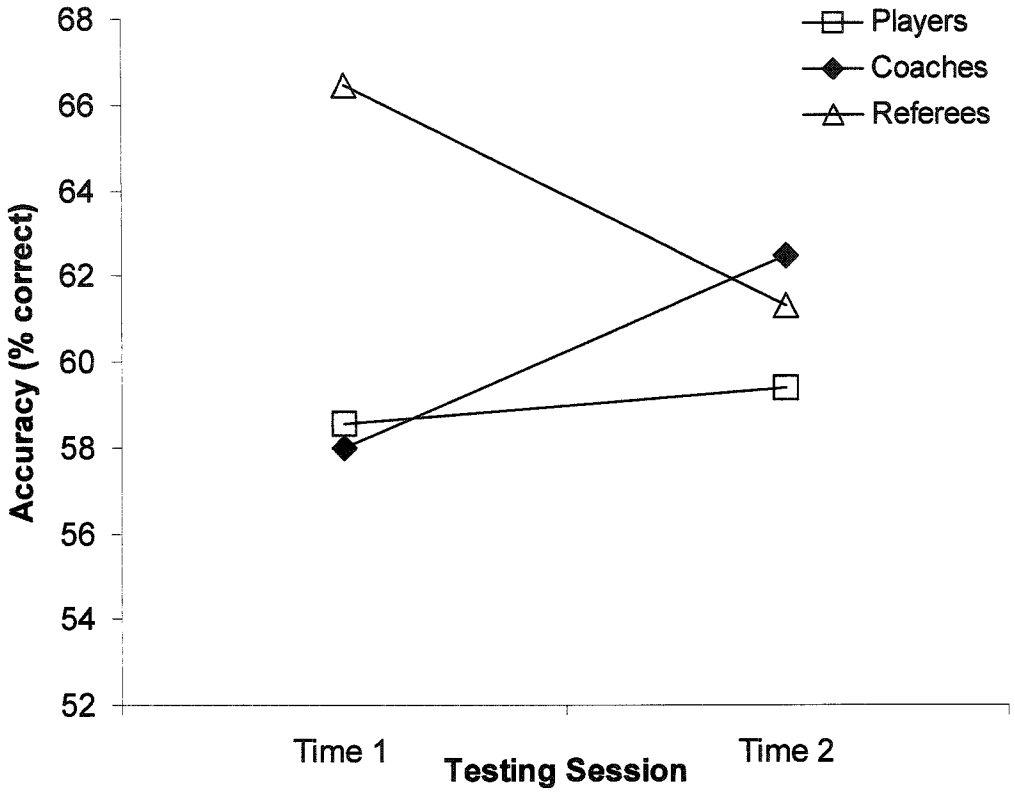

2.5
2
1.5
1
0.5
0
Speed (s)
Time 1
Time 2
Testing Session

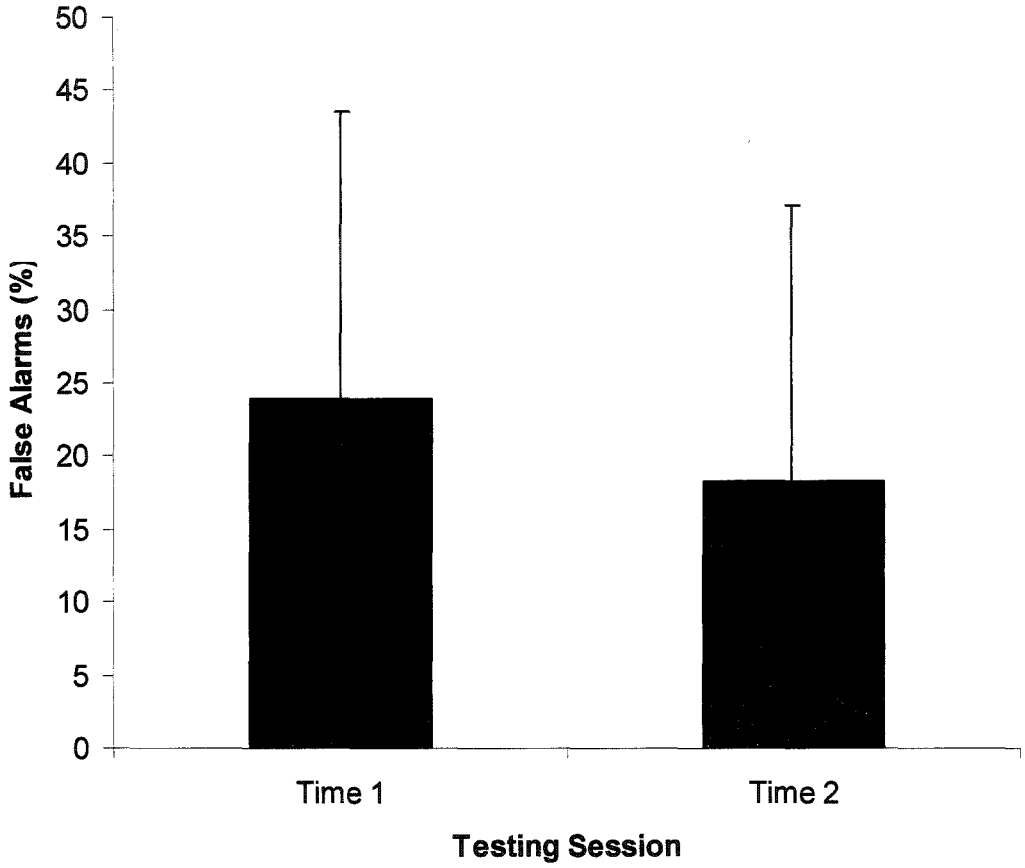

90
80
70
60
50
40
30
20
10
0
Accuracy (% correct)
Players
Coaches
Referees
Group

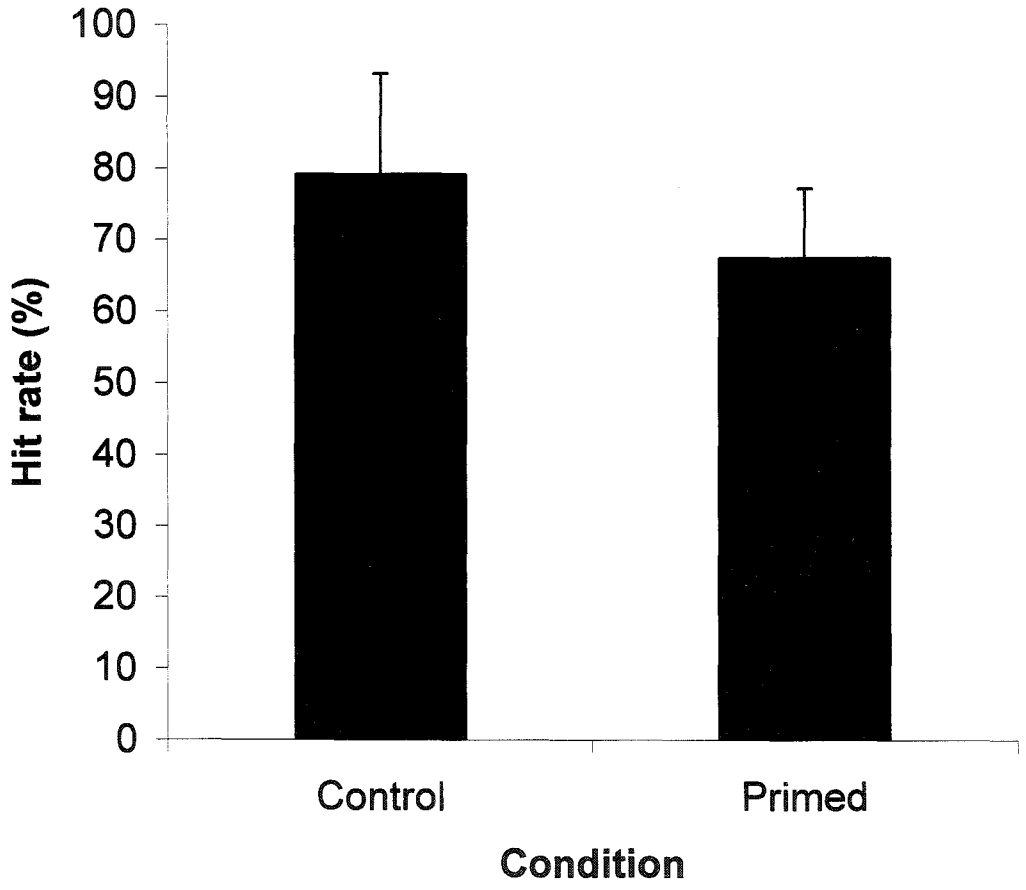

100
90
80
70
60
50
40
30
20
10
0
Accuracy (% correct)
Players
Coaches
Referees
Group

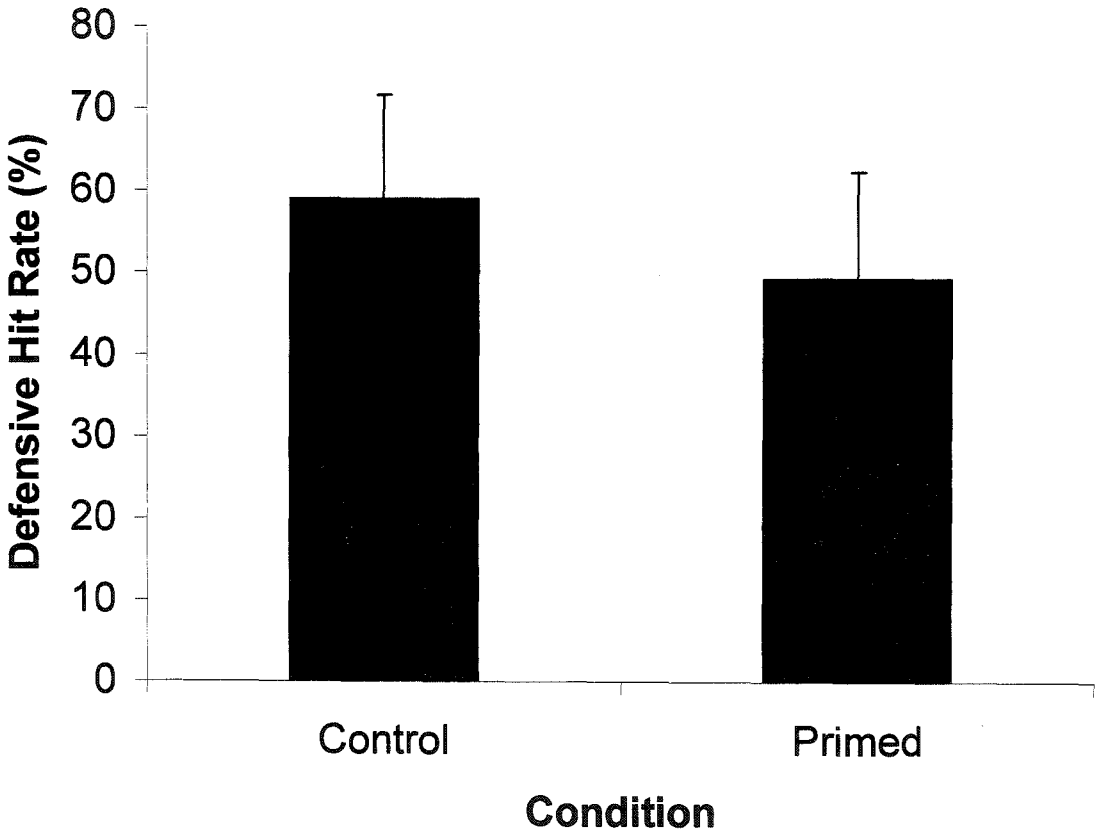

54
52
50
48
46
44
42
40
Accuracy (%)
Time 1
Time 2
Testing Session
Players
Coaches
Referees


3.5
3
2.5
2
1.5
1
0.5
0
Decision Time (s)
Time 1
Time 2
Testing Session


68
66
64
62
60
58
56
54
52
Accuracy (% correct)
Players
Coaches
Referees
Time 1
Time 2
Testing Session


50
45
40
35
30
25
20
15
10
5
0
False Alarms (%)
Time 1
Time 2
Testing Session


100
90
80
70
60
50
40
30
20
10
0
Hit rate (%)
Control
Primed
Condition


80
70
60
50
40
30
20
10
0
Defensive Hit Rate (%)
Control
Primed
Condition

GENERAL CONCLUSION

The purpose of the studies in this dissertation was to explore expertise in open sport referees. To this end, a series of four studies applied the expert performance approach (Ericsson & Smith, 1991), with two foci: 1) practice activities, and 2) skills. Since there is currently very little knowledge related to expertise in sport referees, the findings in this dissertation create a valuable base on which to build future work.

The first paper reported on the first dissertation study. This study examined the training activities of world-class soccer referees by applying the theory of deliberate practice. Similar to findings in athletes (e.g., Helsen, Starkes, & Hodges, 1998), this study found that referees commit a great deal of time to training, with yearly increases in the volume of training as they gain expertise. While training becomes more varied over time, with the addition of different types of activities, there is always a large emphasis on the physical demands of the game, and very little focus on the perceptual-cognitive demands.

A robust finding in previous applications of the deliberate practice theory is that 10 years of practice are necessary to achieve expert performance levels (e.g., Ericsson, Krampe, & Tesch-Römer, 1993). While the first study of this thesis may indicate that it takes longer for referees, with an average of 16 years' experience before reaching the FIFA list, this may be a conservative estimate given the criteria used to classify experts. Furthermore, the lack of structure and feedback from referee training may qualify as *structured practice* more so than deliberate practice as it is defined in the literature (Coté, Baker & Abernethy, 2003). If it is the case that referees are engaging in structured

practice rather than deliberate practice, it is not surprising that performance improvements may take longer.

The second paper took a relatively unique approach to exploring expertise in referees by considering the role-related specificity of skills. A video-based laboratory task identified tackle assessment as a representative skill in refereeing, with referees outperforming players. In contrast, performance similarities in "reading game plays" showed this to be a task representative of more general soccer skill, not specific to role. In general, these findings support previous research showing that skill is specific to role (Allard, Parker, Deakin & Rodgers, 1993; Williams & Davids, 1995). More specifically, it identified role as a vital consideration in task design for sport expertise research, and provides a potential training tool for referee skill.

The final paper in this series moved to the sport of basketball in order to further explore the role-specific nature of referee skill. This exploration was first addressed through a replication and extension of Allard et al. (1993). One of the most interesting findings in the first study of this paper was that, in general, for video clips with identical features (e.g., speed, camera angle), and even for the exact same video clips viewed a second time, changing the task demands changes performance. For example, the findings indicate that anticipation of action is a referee-specific skill, whereas play/clip recognition is a coach-specific skill. Coinciding with Ericsson and Smith's (1991) second step of the expert performance approach, this study provides the hypothesis that information pick-up varies by role, and is thus a mediating mechanism in role-related skill. Moreover, findings using the referee task were used to present the hypothesis that the manner in which information is picked up can be altered, and thereby impact

performance. This finding was linked to previous studies targeting visual diagnoses in medicine (e.g., Kulatunga-Moruzi, Brooks, & Norman, 2001).

The second study in paper 3 focused on the referee task by investigating the lack of strong referee superiority on this task, as compared to the findings in Allard et al. (1993). The influences of knowledge priming and infraction priming were investigated. While it appeared that knowledge priming has no impact on performance, whereas infraction priming interferes, further data collection is needed to eliminate the possibility that these findings are due to the specific clips used in each condition.

Altogether, paper three concludes that further study is needed to explore representative tasks for basketball refereeing skill. One specific area for future investigation is the impact of the visual features of the stimuli used in laboratory tasks (e.g., clip length, size of display, speed, angle of view).

If we compare the soccer officials and their development in paper 1 with the basketball officials and their development in paper 3, we discover that there are inherent differences. European soccer officials begin their training early and do not necessarily play or coach in the early stages of refereeing. Basketball officials and coaches, however, tend to have played for many years both prior to and along with officiating. This makes the goal of finding laboratory tasks that specifically tap referee skills in basketball very difficult. At this point, then, the transfer of knowledge/skills across role is not well understood.

As presented in the general introduction section, the overarching framework for this thesis is the Starkes, Cullen, & MacMahon (2004) developmental model of perceptual-motor expertise. As a conclusion, then, it is worthwhile to compare the results

found in this thesis with the characteristics of performers presented in the model. The perspective taken in this research was perceptual-cognitive in nature, with a focus on expert performers. It is thus the characteristics of routine expertise in the perceptual-cognitive stream that are the most important points for comparison (figure 1, pp. 17).

The first three characteristics used to describe perceptual-cognitive performance at the routine expertise level are: 1) the emergence of causal types of knowledge organized in to "game scripts", 2) the inclusion of temporal and spatial ordering of events in game scripts, and 3) attention to context and unique features. These characteristics, put simply, indicate that the routine expert reads the situation and then makes use of detailed "if-then-do" knowledge to choose and carry out a response. While it is difficult to determine whether these characteristics apply to expert referees given the relatively limited scope of this research, one possibility is that the equivocal findings related to referee superiority in infraction detection reflect the complexity of referee game scripts.

Whereas the use of context improves as an athlete gains expertise, this may be a skill that is much more central to refereeing performance, even early on. For example, in MacMahon (1999), one rugby referee indicated the influence of previous experience on decision-making with the comment ""...usually, if you know the teams, you know if the ball is going to be well-taken or not."" (p. 55). Moreover, the use of game scripts may differ in referees, as opposed to players, who do not *enact* game scripts ("if-then-do"), but rather, *watch* for game scripts to play out ("if he does, then he will do, and I will call"). In study 1 of paper 3, the referee superiority to players in action anticipation speaks to an ability to correctly generate "if he does, then he will do" statements. Perhaps the brevity of the clips used for infraction detection did not allow referees to access the

longer, observer-based scripts used for making referee calls. It also seems likely that, if referees use game scripts, these also incorporate spatial and temporal features. For instance, knowing when the basketball defender's feet are planted relative to contact with an offensive player may determine if the call is an offensive or defensive foul. Changing these features may interfere with performance. This hypothesis based on the model does not, however, account for Allard et al.'s (1993) findings.

The fourth characteristic presented in the routine expertise phase of perceptual-cognitive skill is the use of probabilities and knowledge of one's own skill level in making certain responses more likely. Once again, the findings in the third paper indicate that this characteristic may also be present in referees, with role-specific features. That is, referees' ability to anticipate action may translate into the ability to use probabilities related to knowledge of the physical skills of the *players* being observed. For example, a referee may know that a rugby player has the ability to kick the ball a long distance. This action choice may thus become more likely for this player in a particular situation than for a player who does not have this ability. It follows, then, that, where an athlete's knowledge of his/her own skill level makes certain responses more likely, in referees, knowledge of the skill level of the players helps a referee anticipate likely infractions, in some cases, perhaps also making certain responses (calls) more likely.

The last characteristic of perceptual-cognitive functioning in routine experts as presented by Starkes et al. (2004) is addressed in the first paper of this thesis. Starkes et al. indicate that it is necessary, to achieve routine expertise, to have highly variable practice within competitive environments. It is clear, from paper one, that expert referees spend very little practice time in competitive environments. This raises a number of

questions: If the expert referees examined in paper one, some of the best in the world, were to engage in competitive practice, would their performance improve beyond current levels? Was this group a sample of routine experts, or are referees as a population underdeveloped relative to athletes? Should we use the model to determine whether referees display skills distinct to athletes, or to prescribe training and skill development? Either way, these questions are paradoxically both too broad and too specific to be addressed by the studies in this thesis.

In sum, although we still know very little about expertise in open sport referees, the present work has created a solid base on which to build. It has also shown the efficacy of applications of the expert performance approach to this population.

References

Allard, F., Parker, S., Deakin, J., & Rodgers, W. (1993). Declarative knowledge in skilled motor performance: Byproduct or constituent? In J.L. Starkes, & F. Allard (Eds.), (1991). Cognitive issues in motor expertise. (pp. 95-108). Amsterdam: Elsevier.

Coté, J., Baker, J., & Abernethy, B. (2003). From play to practice: A developmental framework for the acquisition of expertise in team sports. In J.L. Starkes and K.A. Ericsson (Eds.), Expert performance in sports: Advances in research on sport expertise. (pp. 115-136). Champaign, Ill.: Human Kinetics.

Ericsson, K. A., Krampe, R. T., & Tesch- Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. Psychological Review, 100, 363-406.

Ericsson, K. A., & Smith, J. (1991). Prospects and limits in the empirical study of expertise: An introduction. In K. A. Ericsson and J. Smith (Eds.), Toward a general theory of expertise: Prospects and limits. (pp. 1-38). Cambridge: Cambridge University Press.

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.

Kulatunga-Moruzi, C., Brooks, L.R., & Norman. G.R. (2001). Coordination of analytic and similarity-based processing strategies and expertise in dermatological diagnosis. Teaching and Learning in Medicine, 13(2), 110-116.

MacMahon, C. (1999). Making sense of chaos: Decision-making by high and low experience rugby referees. Unpublished master's thesis , University of Ottawa, Ottawa, Ontario, Canada.

Starkes, J.L., Cullen, J.D., & MacMahon, C. (2004). A model of the acquisition and retention of expert perceptual-motor performance. In A.M. Williams, N.J. Hodges, M.A Scott, & M.L.J. Court (Eds.) Skill acquisition in sport: Research, theory and practice. (pp. 259-281). London: Routledge.

Williams, A.M. & Davids, K. (1995). Declarative knowledge in sport: a byproduct of experience or a characteristic of expertise? Journal of Sport and Exercise Psychology, 7 (3), 259-275.