The Referee's Challenge: A Threshold Process Model for Decision Making in Sport Games

Markus Raab, Simcha Avugos, Michael Bar-Eli, Clare MacMahon

Abstract

Judgment and decision making in sporting officials is a challenging task that involves the use of context. Although process models of decision making describe decision contexts, none of the existing models explains when sports officials use rule-driven decision making, or game management. The basic idea of our work is that referees use a subjective threshold to apply game management, which may explain this decision behavior. We propose a new dynamic threshold model that is based on concepts derived from Decision Field Theory (Busemeyer & Townsend, 1993). The model includes two thresholds of game management (high/low) and two contact situations (foul/no foul) as approaching one of these thresholds. Using the example of soccer refereeing, we argue that if the game hits a subjective threshold of aggressive play, then the referee shifts from applying the rules to managing the game. This new approach changes the scientific discussion from one focused on what referees should decide in one situation or the other, to a dynamic model that explains the basic psychological mechanism underlying the referee's change in behavior during the game, both at the intra-individual as well as inter-individual level.

Key words: decision field theory, decision making, threshold model, referee, officiating

Word count: 9130 (excluding Abstract, Table and Figures)

Introduction

It is generally acknowledged that decision making in sport is complex. Within this domain, the decision making of officials is particularly complex, due to the need to account for the effects of context and given the uncertainty of the choices in many ambiguous situations (Pina et al., 2019; Slack et al., 2013). Officials (judges, referees, umpires) are involved in almost every competitive sport. Depending on the sport, judgments are made using different methods for the evaluation of performance and winners (Stefani, 1998), including objective measurements (e.g., a stopwatch in swimming), objective scores (e.g., the number of goals scored), and subjective judgments (e.g., points awarded to athletes for the aesthetics and difficulty of their performance). In many sports, officials are required to make decisions under time pressure in a dynamic environment (MacMahon et al., 2014 for an overview). The complex task of officiating has been classified based on the amount of interaction with athletes, and the number of cues processed, resulting in broad categories of fficials as either “interactors”, “monitors” or “reactors” (Plessner & MacMahon, 2013). Choices for interactor officials such as soccer and basketball referees are arguably more influenced by context and interpretation compared to monitors such as gymnastics judges and reactors such as linesmen in tennis. We focus here on interactor referees that act in ball games, an accepted and frequently used classification of sport (Carron & Hausenblas, 1998) that is also applied to work focused on referees (Blas et al., 2020).

Statistics from the Euro 2000 Championship show that soccer referees made approximately 137 observable decisions per game, with about 44 made by the head referee alone (Helsen & Bultynck, 2004). Besides these decisions, there are also about 60 non-observable decisions (not strict fouls), when a referee decides not to interfere

with play, and therefore does not make a call (Helsen & Bultynck, 2004). Given an effective playing time of approximately 51 min, a top-class referee makes 3-4 decisions per minute (Helsen & Bultynck, 2004). In terms of perceptual-cognitive demands, this is a huge workload. Neville, Salmon, and Read (2016) also show that in an average Australian Rules football game, there are 6,025 communication instances, and 887 moments when an umpire must decide to intervene in the game. Communication in this work was coded as those from umpires to players, but also between the umpiring team, indicating this is another significant demand and component of officiating.

The problem we want to address

The problem we want to address is to explain how referees actually shift from rules to game management during game play. In essence, the decisions of referees while officiating games reflect a choice between the enforcement of the laws (i.e., to be accurate; Plessner & Betsch, 2001) and game management (i.e., to be adequate, to ensure the flow of the game; Brand & Neß, 2004; Brand, Schmidt, & Schneeloch, 2006; Mascarenhas, Collins, & Mortimer, 2002; Unkelbach & Memmert, 2008). More specifically, the concept of game management is a diverse concept that includes, for instance, effective communication, and establishing players' respect beyond the decision-making component we focus on (e.g. Gomez, Ortega & Jones, 2016; Morris & O'Connor, 2017; Pina et al. 2019, Slack et al., 2013 for examples in rugby and soccer). Within the description of choices and their subcategories, interviews with referees revealed that, for instance "accuracy" of big decisions, consistency of decisions during a game, and correctly applying the "Laws of the Game" have been classified as game management (e.g. Slack et al., 2013, Figure 1, p. 304). In specific situations those choices have been described as trade-off decisions, which some refer

to as a dilemma. For instance, Praschinger et al. (2011) define the dilemma as follows:

> "Clearly, we are faced with a dilemma. On one hand, we have the Laws of the Game which referees are supposed to follow to the letter. On the other hand, we have a highly complex and dynamic situation (a soccer match). Referees seem to solve this dilemma by applying game management. They balance their decisions by being sensitive to various influences (e.g., the minute of play). Game management appears to be a necessary prerequisite for applying the written Laws of the Game to specific situations during a match." (p. 344).

Whether this choice between applying the rules of the game and using judgment (game management) can be conceptualized as a dilemma can be discussed. In some experiments, situations have been created to present a choice between either enforcing the law or game management (e.g. Plessner & Betsch, 2001). In other descriptions, game management adds a communication strategy to the enforcement of the law and cannot be conceptualized as a dilemma (Slack et al., 2013). Neither enforcing the law nor using game management are good or bad per se. Thus, we argue that a sanction can be evaluated based on the challenge in ambiguous situations to either sanction violations by a call or keep the flow of the game by not calling this violation.

MacMahon and Mildenhall (2012) illustrate the problem of how to decide what decision approach to use in ambiguous situations. They provide an extreme example of a basketball referee's choice to make a clearly inaccurate call (an unwarranted foul after a player's simulation), to manage a volatile home crowd in a game with a lopsided score. This choice reflects management of the situation in which the specific call was judged as one which would not adversely affect the outcome. The example clearly illustrates that the referee overwrote application of the rules in favor of game management. While this is an extreme example, and a rare occurrence, Kolbinger and Stöckl (2019) provide evidence of regularly occurring rule violations in

soccer, wherein a trivial offense is perceived by the official, but not enforced. Specifically, they showed that in 96.3% of the 618 penalty kicks assessed from games in four European leagues and one cup event, the referee did not call rule violations. Similarly, referees are often encouraged to apply a preventive refereeing approach (Mascarenhas, Collins, & Mortimer, 2015) – which is a form of game management – expressively verbalizing to players and warning them about the potential of their actions (e.g., "number three, get onside…") to prevent fouls from occurring and maintain the flow of the game. Obviously, referees cannot predict the future consequences of their decisions, in terms of control or lack of control over what happens on the field following one decision or the other; however, their early decisions within the context of game management do have an impact. For example, the first yellow card in the game between Cameroon and Germany during the World Cup match in 2002 was awarded against Cameroon's midfielder Marc Vivien Foé in the eighth minute. Starting with a yellow card early in the game, the referees, we can speculate, felt the need to be consistent; at the end of the game, there was a record of 13 additional yellow cards awarded (MacMahon et al., 2015). This case illustrates the consequence of using consistency and more strictly using rule application.

How is the choice of interactor referees currently explained or approached?

About two decades ago the challenge of law enforcement and game management was discussed in the literature as refereeing being either a craft or an art. For instance, Mascarenhas, Collins, and Mortimer (2002) and Plessner and Betsch (2001) debated whether soccer referees use compensation strategies. In the specific case, referees in experiments showed that if they have given a foul for one team that results in a penalty, there is a lower likelihood that they will give the same team another penalty when faced with an ambiguous situation. Rather, the likelihood of

giving a penalty to the other team was higher, with penalties balancing out between the teams. The conclusion of the studies, the debate, and the rejoinder highlighted that this is indeed a challenge to be addressed: within a game, referees are constantly faced with what guiding principle to use in their decision making – rule application, or game management. Although Brand, Schmidt, and Schneeloch (2006) also provided some laboratory-based empirical evidence for game management as a guiding principle for decision making, 20 years later this discussion has not yet been further or systematically developed to testable predictions or specific recommendations for referees in training or performance. While research on different factors that influence decisions and decision phenomena have been conducted (e.g., gymnastics' country; Damisch, Mussweiler, & Plessner, 2006), they do not explicitly address this challenge.

From the current interviews and task classifications that have been published, the complexity of referee decision making including multiple dimensions of game management are self-reported, observed ,or experimentally studied (MacMahon et al., 2015 for a summary). Our focus on choices will allow operationalization of the challenge using the illustration of a situation that is less ambiguous than a dynamic foul situation: Cursing by soccer players. According to Law 12 of the game (Fédération International de Football Association) a player is to be expelled from the game by a red card for using assaulting language or gestures. Praschinger et al. (2011), asking 113 referees, showed that the sanction varied dramatically between referees and depended on content of the swear word. For instance, if the insulting content of the swear word was attributed to the appearance of the referee about 33% decided on a red card whereas about 74% when the content pertained to sexual orientation. In the appearance content, 31% of the referees decided for a yellow card,

20% for admonition (verbal warning) and about 12% for ‘no reaction’, illustrating the large amount of individual differences in the choice. Thus, our position is that we need to understand how referees solve the challenge between the enforcement of rules and game management to allow better and accepted decisions to develop. The current research in this area is primarily descriptive in nature, as illustrated by two current conceptualizations.

First, the concept of accurate-adequate decisions in refereeing (Brand, Plessner, & Schweizer, 2009) argued that, in essence, referees' choices can be described as a tradeoff between accurate and adequate decision making. It is not always clear, however, which of these two is the better option, or what drives the choice (Schweizer & Plessner, 2016). Accurate decisions are made when a referee can categorize whether an incident was a foul or not according to the criteria in the laws of the game. Adequate decisions consider the dynamics of the respective game. Thus, in officiating games adequately referees should adjust their judgment of each single contact foul (foul/no foul) to the concrete context of the ongoing situation. The accuracy-adequacy model assumes that accuracy is a necessary requirement for making adequate decisions. Still, sport associations (e.g., IFAB – International Football Association Board) do not expect referees to achieve 100 percent accuracy in decisions for every single incident, but rather to avoid clearly incorrect decisions that might influence the outcome of the game ("game-changing" situations), such as in soccer goals, penalty decisions and direct red card incidents (Helsen, MacMahon, & Spitz, 2019).

A second modeling proposes that accurate-adequate choices are in fact the use of Type-1 versus Type-2 decisions in refereeing (Helsen, MacMahon, & Spitz, 2019). Type-1 refers to rapid and intuitive decisions and Type-2 decisions are more

deliberate ones that are used in addition to Type-1 game management type of decisions. In other words, deliberate and slower processing is triggered when contextual cues (e.g., prior decisions) cast doubts on the initially triggered decisions (Helsen, MacMahon, & Spitz, 2019). Biases can occur in high-conflict situations (e.g., second yellow card). The rationale of this model refers to the theoretical view that human judgment and decision making is assumed to arise from the interaction of two different systems of reasoning. Kahneman and Frederick (2002) simply labeled these Type-1 and Type-2 choices coming from "system 1" and "system 2", but other descriptions for the dual process of thinking were proposed, including impulsive versus reflective (Strack & Deutsch, 2004; Strack, Werth, & Deutsch, 2006), symbolic versus associative (Sloman, 1996; Smith & DeCoster, 2000), and intuitive/affective versus rational/deliberative system. System 1 is assumed to be emotional, automatic, fast, and implicit, whereas system 2 is often described as controlled, slow, and explicit (Kahneman, 2011). We can illustrate the pros and cons of applying a very specific dual-process model to referee decisions. Evans' (2008) default-interventionist dual process model (DIDPM) assumes a specific and sequential interaction of heuristic and deliberative thinking such that a referee first uses a default process of intuitive thinking to decide on a foul or not. Before the referee whistles, however, a second and sequential analytical process may or may not intervene and thus the name default-interventionist model. The DIDPM could explain referee choices, but would ignore both context effects of the previous decisions that may influence the current one, and how other cues about the development of aggressive behavior in general influence the choice. Finally, DIDPM ignores individual differences. In contrast, we use individual differences as model parameters to understand the different behaviors in referees even though dual-process models in

principle are able to model multiple interactions of dual-processes, individual differences, and context effects (e.g., Raab, 2015; Furley, Schweitzer, & Bertram, 2015, for applications to athletes' decisions). On a more critical note, it has been argued that, given the limits of the current dual-process models (Melnikov & Bargh, 2018), we can do better in modeling individual differences (Boogaart et al., 2018) and explaining action control in general (Hommel, 2019). Theoretical predictions for a DFT and a DIDPM model in the future could be empirically tested but are out of the realm of the theoretical positions presented here.

Both descriptions of the trade-off in showing accurate and adequate decisions restate the problem in different words but do not explain the decision process to shift from an accurate decision to an adequate one or from type 1 to type 2. Further, both descriptions do not allow any prediction of when each guiding principle is used or when switches take place. Finally, neither description is sensitive to intra- or interindividual differences that would help to personalize and specify the selection and development of referees.

Given the silence in science on explaining the game management-rule application, accurate-adequate trade-off in refereeing decisions is the rationale for developing a testable model. From the above current state of research it seems evident that there are three issues the model needs to address. First, the model should describe the processes of game management and rule application and how choices could be explained. Second, the model should be able to predict when switches occur in the dynamics of a developing game and show violations of predictions and alternative causes driving choices. Third, the model should be sensitive to inter- and intraindividual differences to develop a personalized and context-specific diagnostic and intervention profile for referees. Our intention, then, is to provide a model which

can serve these purposes: explain the process, predict behavior (and identify when each behavior is preferable), and provide guidance for training. A test of our model can be envisioned in empirical studies and we will provide a sample of these at the end of the paper (in Table 1). Further, the model should be able to demonstrate how diagnostics and interventions should be developed, testing better choices against existing typical training of referees.

The threshold model of refereeing

Basic assumptions

Context matters

The primary goal of the proposed model around the challenge between game management versus rule application is to improve the referees' decisions. The function of the model is therefore to provide a practical impact to the field for training and selection of referees. An empirical validation of the model is a future goal that will be described in the discussion. A basic assumption of the model is the need to understand the referees within their environment and its dynamics. Thus, context matters. The necessity of accounting for context in a decision model for refereeing is underscored by the research evidence that decisions change according to context. For example, consider research in referees showing compensation mechanisms, wherein previous decisions may bias a decision in favor of the opponent team (Plessner & Betsch, 2001). Similarly, we can consider effects in refereeing that produce biased decisions in favor of the home-team (Boyko, Boyko, & Boyko, 2007; Poolton, Siu, & Masters, 2011), the effect of reputation of a player or team (Jones, Paull, & Erskine, 2002), and moment-to-moment game encounters (Unkelbach & Memmert, 2008). Perhaps not surprisingly, therefore, high-level decision making has also highlighted the effects of the specific time or score (e.g., pitch counts in baseball; MacMahon & Starkes, 2008).

Framing decisions in different contexts such as searching for similarity or contrast also changes decisions (Damisch, Mussweiler, & Plessner, 2006). Evidence for the effect of context was also found in an experiment by Brand, Schmidt, and Schneeloch (2006) with elite basketball referees, showing that the referees decided on less rigorous sanctions when video clips were presented to them chronologically as was in the game (sequential context) compared to a randomized order. Further, the context of the home crowd behavior was suggested to explain the increase in yellow cards awarded by the referees to the away teams after the introduction of the video system (TMO) in rugby (Dawson, Massey, & Downward, 2019). The TMO is argued to be less influenced by crowed behavior however, rather ironically, it appears to contribute to the previously documented home advantage (e.g., Nevill, Balmer, & Williams, 2002; Nevill et al., 2017). It is clear, from the evidence, that referee decision processes are inextricably linked to the complex and multifaceted context in which they take place – and thus context is an essential component of any decision model.

Individual differences matter

The above-described effects in refereeing have provided evidence that context matters in referee choices and thus it seems valid to consider context in any model that explains shifts from accurate to adequate decisions. The example of the referee giving 13 more yellow cards after an early yellow card illustrates the use of context, but this context was driven by the choice to remain consistent in events that followed of a similar nature. However, consistency cannot account for many of the effects described above, or for the situation of the basketball referee who responded to the crowd by violating consistency in favor of deregulating the current aggressive atmosphere (MacMahon & Mildenhall, 2012). Consistency also does not account for

the minor rule violations, which sometimes are ignored by the referees, as shown in the data of Kolbinger and Stöckl (2019). Our model can be considered a context model in a broader sense that includes the time scale ranges over games or specific short events within a game that allow a shift from accurate to adequate decisions. Moreover, it also considers individual differences in referees for rule application. For example, we might compare two soccer referees, R_1 and R_2. Each of the referees has a subjective threshold level (high/low) for applying game management. R_1 is considered more of a 'law enforcer' referee (Praschinger et al., 2011), and R_2 is 'game manager' (Praschinger et al., 2011). A model of decision processes and shifts from rule administration to game management needs to consider these individuals.

Considering these basic assumptions of context and individuals, there are key components of the model as described below.

Basic description of the threshold model

The difference between referees R_1 (law enforcer) and R_2 (game manager), and the fact that they have different thresholds for when they will use game management, is a key component of the model, hence we call this a threshold model. Thresholds in general decision making are a well-established concept that illustrates that we stop to search for information when a specific threshold is met. Thresholds can vary subjectively by the needs of an individual person to gain information before making a choice and are defined by the situation in which a decision needs to be somewhat fast and the person has limited resources for making a choice. In our use of thresholds, individual referees may always strive to make accurate decisions, however, acknowledge the need to apply game management, similar to the scenario described above (MacMahon & Mildenhall, 2012). Each individual may have a

different level for how much change of context (e.g., change to higher level of aggression) prompts an increase in game management behaviors.

A classical model that assumes choices are made when a threshold is met is the decision field theory (Busemeyer & Townsend, 1993). Decision field theory (DFT) is a dynamic-cognitive approach to human choice behavior. The model describes how a person's preferences evolve across time until a decision is reached. The model is used to predict how humans make decisions under uncertainty, how decisions change under time pressure, and how choice context changes preferences. DFT fits in a general class of sequential sampling models. DFT has multiple parameters but the threshold parameter is the best choice to explain the shifts to game management we describe as a decision to shift after a threshold is met. The basic idea underlying the decision process is illustrated in Fig. 1 and refers to three prototypical situations a referee may experience as documented in the analyses of referee performance and self-reports (Cunningham et al., 2014; Pina et al., 2019; Slack et al., 2011). We acknowledge there can be a large number of response choices in every refereeing situation, but illustrate with a simple example. Suppose the decision maker is initially presented with a choice between three risky prospects, A, B, and C, at time $t = 0$. Each trajectory in the figure represents the preference state for one of the alternatives at each moment in time.

Insert Fig. 1 near here

At each moment in time, the decision maker thinks about various payoffs of each prospect, which produces an affective reaction, or valence, to each prospect. These valences are integrated across time to produce the preference state at each

moment. During the stages of processing, attention can shift towards advantages favoring one prospect over the other. The stopping rule for this process of comparing payoffs and assessing preferences is controlled by a threshold: the first prospect to reach the top threshold is accepted. Choice probability is determined by the first option to win the race and cross the upper threshold, and decision time is equal to the deliberation time required by one of the prospects to reach this threshold.

The threshold is an important parameter for controlling speed-accuracy tradeoffs. High thresholds require a strong preference state to determine the decision, which allows for more sampling of information about the prospects, extending the deliberation process, and thus increasing accuracy. It follows, then, that low thresholds require a weaker preference state and less sampling of information about the prospects, thus shortening the time taken to deliberate processing, but also decreasing the likelihood of choices accuracy. Under high time pressure, decision makers must choose a low threshold, while under low time pressure a higher threshold can be used to increase accuracy. Thus, decisions can be driven by time pressure, with reversed choices depending on the amount of time available.

A very important feature of the concept of thresholds is that it allows individual differences that describe high or low thresholds (inter-individual differences), as well as different thresholds in different contexts (intra-individual differences). Our predictions are based on the tenets of DFT, assuming a threshold model with two thresholds of game management (High/Low) and each choice (foul/no foul) as getting closer to one of these thresholds. The model suggests that referees hold a threshold for applying game management: if game dynamics allows more strict application of the rules, then the referee is under threshold of applying game management. However, as soon as a subjective threshold is met, referees apply game

management to either let the game flow (and thus be looser in applying the rules and not calling), or if aggression starts – make a call earlier (using preventive refereeing and more conversation, or an earlier foul call). Thus, our model provides predictions, for example, when the events in a game mean that it has hit a subjective threshold of becoming rougher, then the referee changes from applying the rules to managing the game by verbalizing to the fouling players and calling minor fouls earlier. Thresholds can be for instance set between to maximum 1 indicating potentially zero switches to game management to 0 indicating for every change in context a shift to game management. Most likely real data showing extreme cases of 13 yellow cards or zero yellow cards in increasing aggression games are the boundary condition of threshold distribution in real games. Threshold can as well vary over time, can have a slope or non-linear curvatures but for simplicity we assume a simple model. Below we will illustrate the details of accurate and adequate decisions based on existing findings and theoretical perspectives that all share the same gap in explaining how referees shift between accurate and adequate choices.

How does the model account for context?

We will describe three possible decision processes across the game's timeline, referring mostly to those ambiguous contact situations where decisions are not clear-cut:

Situation A – the game is played more or less with limited aggression (i.e., assertive, flat game; see Fig. 2). Each referee is likely to whistle according to his/her threshold level. The different thresholds and decision behaviors can be explained by individual differences in personality traits, accumulated experiences and preferences (e.g., Arslanoğlu, Doğan, & Acar, 2018; Guillén & Feltz, 2011; Werger, 2017). Under an assertive scenario the referees, R_1 and R_2, would apply the rules as is, because they

are both under their threshold of applying game management. Context is overlooked, and thus accuracy is increased. In terms of implications when assessing a referee, we argue that fewer mistakes in application of the laws are acceptable for this sort of game, as we would expect more law enforcement. Thus, factors such as the level of the game being played, which may be critical in other ways of conceptualizing decision processes in refereeing, are less critical factors for thresholds for game management than the level of aggression or the severity of the actions from a safety perspective.

It should be emphasized that situation A does not really pose a challenge, as this may be a rare scenario. Nevertheless, a great deal of work that looks at referees' decision making assumes this lack of variation, although previous evidence from game analysis (e.g., Bar-Eli, Tenenbaum, & Geister, 2006; Kirker, Tenenbaum, & Mattson, 2000) shows that most games are *not* flat and unvaried.

Insert Fig. 2 near here

Situation B – the level of aggression fluctuates throughout the game. This is the situation that is most typical of most ball sports (Kavussanu & Tenenbaum, 2014; Russell, 2008), and thus most critical for any model attempting to predict and explain behaviors. Let us assume that the game starts with a low level of violence. This would allow the referees, R_1 and R_2, to apply the rules because they are both under their threshold of applying game management. In time t_1 the level of aggression in the game increases. If a subjective threshold is met, the referee calls a foul or penalty. The referees in our model pay attention to the history of the game when evaluating each event. Thus, the accuracy and adequacy of decisions is influenced by each referee's

threshold, and depends on the specific situation/circumstances at the time that the offense occurs (i.e., what is fair and better for the flow of the game).

Situation C – the game is characterized by a constant high level of aggression, with little fluctuations. According to Unkelbach and Memmert (2008), soccer referees call relatively fewer fouls at the early phase of the game, because they have to calibrate their judgment scale and develop a "feeling for the game" (Brand & Neß, 2004). Then, for the rest of the game referees decide whether to whistle or not and how severely to punish a fouling player or team according to the game's context. Based on this approach to explaining decision choices, the major principle guiding the referees' judgments would be staying consistent with previous decisions in the game. For example, a referee does not call a foul for one of the teams. Later in the game, if the other team commits a similar offense, then he/she would most likely decide not to call a foul in order to maintain consistency, regardless of the context in which the offense occurs.

Relating this to our argument, we expect that a referee with a low game management threshold (R_2 – game manager referee), who shifts to a game management decision control process earlier, would call fewer fouls (e.g., less yellow cards) under possible foul situations in order to stay consistent with his/her previous decisions (a game management behavior). The referee with a higher threshold for game management (R_1 – law enforcement referee) would call more fouls if called a foul early. Thus, the level of accuracy and adequacy in referees' decisions would be influenced by the major principle of consistency.

With what follows, we underline the strengths of the threshold model over the current existing models in explaining and predicting the in-game referees' decisions and actions.

How does the model account for individual differences?

Inter-individual differences

The threshold model assumes that the value for an option at a certain time is determined by the perception of how best that option fits (foul without a yellow card), relative to another option (foul with yellow card) on the single attribute (e.g., bodily contact of the players) under consideration. For simplification, it is illustrated only for two options and one attribute but the model can be extended to multiple options and multiple attribute scenarios (Diederich & Turblond, 2018). This momentary valence is added to a modified trace of the previous preference state (e.g., given the general tendency to award more or fewer yellow cards), resulting in a vector P(t) of preferences for each alternative at each time. An alternative is chosen when the preference for that alternative exceeds some threshold value, denoted θ, that the individual considers ‘sufficient’ for making a decision. Let us assume we have information about the person's preference for intuitive and deliberative processing when making decisions. The knowledge of the preference for intuitive decisions as one of many potential variables to assess a threshold would allow us to individually set a threshold in DFT. The information about the preference for intuitive decisions could be assessed using the Preference for Intuition and Deliberation Scale (PID; Betsch, 2004). It provides questions for the subscale such as "I listen carefully to my deepest feelings" (Intuition) or "When I have a problem I first analyze the facts and details before I decide" (Deliberation). The scales allow categorization of people as high in Preference for Intuition, high in Deliberation or situation-specific high or low

on both scales. The PID scores could be used to transform PID Questionnaire individual differences to Parameter Values of our threshold model. The linear transformations from PID to our threshold parameter could be performed using the Equation:

Threshold, $\theta = (36 - PID)/36$

For example, the two most extreme individuals (in terms of PID score for intuition and thus may consider more game management information taken into account for making a decision) could be characterized by the following parameters of $\theta = .5$ (if a person scores 18 points for PID in favor of intuition resulting in a calculation of ((36-18)/36) for a lower threshold, and $\theta = 1$ ((36-0)/36) for a threshold twice as high, for a relatively deliberative decision maker. The dependent variable to measure a choice (e.g., to call a foul or not) may use the current context of the situation as indicated in our model. Interindividual differences of referees may predict who will use additional information based on a PID score. Also, sensitivity analyses (further predictions generated from around the parameter space) allow us to show if the parameter range affects the results and thus to test the threshold model. A default value to be specified from the data of a sample for θ reflects the mean of the distribution of PID-transformed variables. The distribution of θ is based on the transformation. Finally, the time step parameter (h) was set to 0.01 to closely approximate a continuous (rather than discrete) deliberation process given the time and potential processing underlying dynamic situations such as a foul in soccer. The model would produce a probability for each option (e.g., *p*(foul/yellow) = .27, *p*(foul/no yellow) = .73) given the threshold parameter used from the PID transformation. These values can be compared with referees' real behavior for validation and model fit (see Raab & Johnson, 2004 for a demonstration in athletes'

decision making based on their questionnaire values as action- or state-orientation). The model can as also be individualized or use mean values for a group of people and can be cross-validated to new samples. Outputs can be deterministic or probabilistic depending on the question at hand (see Glöckner et al., 2012 for different kinds of models and validation processes using gaze data to predict choices by athletes). In the case of inter-individual differences the model simply assumes a flat game in which the threshold is stable and is driven by the individuals' baseline threshold for given fouls and yellow cards. Let us now assume a situation in which intra-individual differences need to be explained under varying thresholds given changes of the situation.

Intra-individual differences

In Fig. 2 we describe situations of changing context that require game management. We assume that the threshold parameter is adapted given the current context. For example, if an individual with a threshold of .5 (see above section, where this is based on the PID score) now detects the need for game management by providing more verbal warnings or fouls, the threshold would be reduced by some incremental points (for modeling purposes maybe in steps of .1, depending on the distance between the context n-1 and now). If the change in the game allows for less game management and more rule application, the threshold will increase to .6 or higher. Given the threshold change over time and the current context violations of consistency, changes of choice probabilities for one or the other option will be modeled based on context changes. Again, validation of the model can be achieved by contrasting a model fit of simulated behavior with real referees' behavior.

If we want to account for the combined effects of intra-individual and context effect in referees' choices, we consider both the starting threshold for awarding penalties and the given current context. Russell, Renshaw and Davids (2018) support

the idea that a game is co-created by the interactions of players and officials. For example, a referee's excessive use of yellow cards may change the context and threshold. In response, player behavior continues to change the context, reinforcing and even lowering the threshold.

Comparison of the proposed model to existing ones

Any model of decision making that considers the referee's challenge about when to use game management and when to use rule administration needs to account for the research aimed at understanding influences on both accurate (rule administration) and adequate (game management) decisions. Therefore, in Table 1 we review what the research tells us about referees' decision making based on accuracy and based on adequacy. Fiedler and Bless (2000) is the framework on which much of the research on accuracy is based. This framework looks at components of the decision-making process (see Fig. 3). The multiple cue usage approach similarly examines where errors can be avoided in training (Plessner et al., 2009). In contrast, consistency is the emphasis in the calibration work (e.g., Memmert et al., 2008). What all of these perspectives do not help explain, however, is the switch within the referee's dilemma between adequate and accurate decision making.

Insert Table 1 near here

Insert Fig. 3 near here

Designs of experiments to test the model

In order to test the threshold model we propose the following design: A two-factorial design, with groups of referees reflecting inter-individual differences (R_1 and R_2 grouped by PID) x 3 game contexts that may produce different intra-individual

choices (flat-no aggression, high aggression-constant, high/low aggression -fluctuate). The task would be a classical video test as used in several studies referenced in Table 1: Video-based decisions in ambiguous situations (in soccer) with occlusion points at which the referees (as participants) indicate their decision (e.g., warnings, foul, yellow cards, red cards, check in video-replay, communicate with players, communicate with assistant referees). Our hypotheses predict main effects for inter-individual differences. R_2 (game manager referee) with a high preference for intuition and lower threshold could include more game management as compared to R_1 (law enforcer referee; high on preference for deliberation) over all situations. Further, we predict main effects for game contexts. Specifically, in Situation A with no aggression, both referees will show consistent and few sanctions and low levels of communication. In situation B, with varying aggression and a generally changing context, we would expect more communication and sanctions that are less consistent with previous calls. In addition, we assume an interaction of both factors: interindividual differences and context. This would mean that in Situation B, with variable aggression levels, a referee with a high threshold for switching to game management would make a switch later in a game that has increased in aggression, compared to the referee with a lower threshold. Finally, in situation C, given the high and constant level of aggression, communication and sanctions would systematically go up.

Cross-validation of our model as a computational model would use the threshold parameters derived from individual differences (as explained above in transforming self-reports in model-parameters, e.g., PID score) to predict the choice distribution in the next game. We would use previous exercises of modeling sport choices (e.g., Johnson, 2006; Raab & Johnson, 2004).

Conclusions and future directions

We conclude that the threshold model informs the field by proposing a process of how referees switch between accurate and adequate choices (rule administration vs. game management). Previous models, in contrast, have merely described the challenge or focused on one side to explain either accurate or adequate choices. Our model can be empirically tested against alternative models and be used for computational modeling to predict choices. Whether the value of the model is based on its explained variance is an empirical question, however we do see direct implications for training if our model holds. For instance, typical training and evaluation of referee performance should not purely focus on accurate decisions. Instead performance measures should include when a referee used game management and whether violation of rule application was appropriate. Given the importance of game management for making accurate and adequate decisions, as well as the importance of other dimensions of game management (Slack et al., 2013), we propose that referee selection and referee training can be evaluated in a more holistic manner.

First, we propose that the use of individual video game clips to train decision making should also include context of the game to show sequences of decisions including the sequentially previously made choices. Knowing about thresholds of individual referees and potential indicators of those thresholds may help to personalize the training and decisions of referees, catering to individual needs.

Second, and more generally, we propose that for a good professional (as in other domains such as a good doctor, lawyer) training and education sets the base for good choices but context-specific experience will foster the enforcement of rules through game management.

Referees should thus be evaluated on both their accuracy and game management. The threshold model will have its limits in explaining behavior because referees cannot be trained to zero errors. However, this is precisely why the choices of referees are an exciting part of the game, to which science can contribute some key understanding.

References

Arslanoğlu, C., Doğan, E., & Acar, K. (2018). Investigation of decision making and thinking styles of volleyball referees in terms of some variables. *Journal of Education and Training Studies, 6*, 21-28.

Balmer, N. J., Nevill, A. M., Lane, A. M., Ward, P., Williams, A. M., & Fairclough, S. (2007). Influence of crowd noise on soccer refereeing consistency in soccer. *Journal of Sport Behavior, 30*, 130-145.

Bar-Eli, M., Tenenbaum, G., & Geister, S. (2006). Consequences of Players' dismissal in professional soccer: A crisis-related analysis of group-size effects. *Journal of Sports Sciences, 24*, 1083-1094.

Betsch, C. (2004). Präferenz für Intuition und Deliberation (PID): Inventar zur Erfassung von affekt- und kognitionsbasiertem Entscheiden [Preference for intuition and deliberation (PID): An inventory for assessing affect- and cognition-based decision-making]. *Zeitschrift für Differentielle und Diagnostische Psychologie* [*Journal for Differential and Diagnostic Psychology], 25*, 179-197.

Bloß, N., Schorer, J., Loffing, F. & Büsch, D. (2020). Physical Load and Referees' Decision-Making in Sports Games: A Scoping Review. *Journal of Sports Science and Medicine, 19*, 149-157.

Boogert, N. J., Madden, J. R., Morand-Ferron, J., & Thornton, A. (2018). Measuring and understanding individual differences in cognition. *Philosophical Transactions of the Royal Society, B, 373*, 20170280.

Boyko, R. H., Boyko, A. R., & Boyko, M. G. (2007). Referee bias contributes to the home advantage in English Premiership football. *Journal of Sports Sciences, 25*, 1185-1194.

Brand, R., & Neß, W. (2004). Regelanwendung und Game-Management: Qualifizierende Merkmale von Schiedsrichtern in Sportspielen. [Rule administration and game management: Qualifying characteristics of referees in sport games]. *Zeitschift für Sportpsychologie* [*Journal for Sport Psychology*], *11*, 127-136.

Brand, R., Plessner, H., & Schweizer, G. (2009). Conceptual considerations about the development of a decision-making training method for expert soccer referees. In D. Araújo, H. Ripoll & M. Raab (Eds.), *Perspectives on cognition and action in sport* (pp. 181-190). New York, NY: Nova Science Publishers, Inc.

Brand, R., Schmidt, G., & Schneeloch, Y. (2006). Sequential effects in elite basketball referees' foul decisions: An experimental study on the concept of game management. *Journal of Sport and Exercise Psychology, 28*, 93-99.

Busemeyer, J. R., & Townsend, J.T. (1993). Decision field theory: A dynamic cognition approach to decision making. *Psychological Review, 100*, 432-459.

Carron, A. V., & Hausenblas, H. A. (1998). Group dynamics in sport (2nd ed.). Morgantown, WV: Fitness Information Technology.

Cunningham, I., Simmons, P., Mascarenhas, D., & Redhead, S. (2014). Skilled interaction: Concepts of communication and player management in the

development of sport officials. *International Journal of Sport Communication, 7*, 166-187.

Damisch, L., Mussweiler, T., & Plessner, H. (2006). Olympic medals as fruits of comparison? Assimilation and contrast in sequential judgments. *Journal of Experimental Psychology: Applied, 12*, 166-178.

Dawson, P., Massey, P., & Downward, P. (2019). Television match officials, referees, and home advantage: Evidence from the European Rugby Cup. *Sport Management Review*. In press.

De Oliveira, M. C., Orbetelli, R., & De Barros Neto, T. L. (2011). Call accuracy and distance from the play: A study with Brazilian soccer referees. *International Journal of Exercise Science, 4*, 30-38.

Diederich, A., & Trueblood, J. (2018). A dynamic dual process model of risky decision making. *Psychological Review, 125*, 270-292.

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*, 255-278.

Fiedler, K., & Bless, H. (2000). Social cognition. In M. Hewstone & W. Stroebe (Eds.), *An introduction to social psychology, 3rd Edition*. Cambridge, UK: Blackwell.

Frank, M. G., & Gilovich, T. (1988). The dark side of self- and social perception: Black uniforms and aggression in professional sports. *Journal of Personality and Social Psychology, 54*, 74-85.

Furley, P., Schweizer, G., & Bertrams, A. (2015). The two modes of an athlete: dual-process theories in the field of sport. *International Review of Sport and Exercise Psychology, 8*, 106-124.

Ghasemi, A., Momeni, M., Rezaee, M., & Gholami, A. (2009). The difference in visual skills between experts versus novice soccer referees. *Journal of Human Kinetics, 22*, 15-20.

Glöckner, A., Heinen, T., Johnson, J. G. & Raab, M. (2012). Network approaches for expert decisions in sports. *Human Movement Science, 31*, 318-333.

Gomez, M.-A., Ortega, E. & Jones, G. (2016). Investigation of the impact of 'fouling out' on teams' performance in elite basketball. *International Journal of Performance Analysis in Sport, 16*, 983-994.

Guillén, F, & Feltz, D. L. (2011). A conceptual model of referee efficacy. *Frontiers in Psychology, 2*, Article 25.

Hagemann, N, Strauss, B., & Leißing, J. (2008). When the referee sees red…. *Psychological Science, 19*, 769-771.

Hancock, D. J., & Ste-Marie, D. M. (2013). Gaze behaviors and decision making accuracy of higher- and lower-level ice hockey referees. *Psychology of Sport and Exercise, 14*, 66-71.

Helsen, W., & Bultynck, J. B. (2004). Physical and perceptual-cognitive demands of top-class refereeing in association football. *Journal of Sports Sciences, 22*, 179-189.

Helsen, W., MacMahon, C., & Spitz, J. (2019). Decision making in match officials and judges. In M. Williams & R. Jackson (Eds.), *Anticipation and Decision-Making in Sport* (pp. 250-266). New York, NY: Routledge,

Hommel, B. (2019). Binary theorizing does not account for action control. *Frontiers in Psychology, 10*: 2542. doi: 10.3389/fpsyg.2019.02542

Johnson, J. G. (2006). Cognitive modeling of decision making in sports. *Psychology of Sport and Exercise, 7*, 631-652.

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

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment: In T. Gilovich, D. Griffin & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49-81). New York, NY: Cambridge University Press.

Kavussanu, M., & Tenenbaum, G. (2014). Aggression. In R. C. Eklund & G. Tenenbaum (Eds.), *Encyclopedia of Sport and Exercise Psychology*. Thousand Oaks, CA: Sage Publication.

Kirker, B., Tenenbaum, G., & Mattson, J. (2000). An investigation of the dynamics of aggression: Direct observations in ice hockey and basketball. *Research Quarterly for Exercise and Sport*, *71*, 373-386.

Kolbinger, O., & Stöckl, M. (2019). Misbehavior during penalty kicks and goalkeepers holding the ball too long as trivial offenses in football. *Frontiers in Psychology, 10,* Article 844.

Levitt, E. E., & Tockman, R. S. (1991). Impact of the sideline behavior of coaches on the decisions of game officials. In W. K. Simpson, A. D. LeUnes & J. S. Picou (Eds.), *Applied research in coaching and athletics* (pp. 185-194). Boston, MA: American Press.

MacMahon, C., Mascarenhas, D., Plessner, H., Pizzera, A., Oudejans, R. R. D., & Raab, M. (2014). *Sports officials and officiating: Science and practice.* Taylor and Francis Inc.

MacMahon, C., & Mildenhall, B. (2012). A practical perspective on decision making influences in sports officiating. *International Journal of Sports Sciences & Coaching, 7*, 153-166.

MacMahon, C., & Starkes, J. (2008). Contextual influences on baseball ball-strike decisions in umpires, players, and controllers. *Journal of Sport Sciences, 26*, 751-760.

Mallo, J., Frutosa, P. J., Juáreza, D., & Navarroa, E. (2012). Effect of positioning on the accuracy of decision making of association football top-class referees and assistant referees during competitive matches. *Journal of Sport Sciences, 30*, 1437-1445.

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

Mascarenhas, D. R. D., Collins, D., & Mortimer, P. (2015). Elite refereeing performance: Developing a model for sport science support. *The Sport Psychologist, 19*, 364-379.

Melnikoff, D. E., & Bargh, J. A. (2018). The mythical number two. *Trends in Cognitive Sciences, 22*, 280-293.

Memmert, D., Unkelbach, C., Ertmer, J., & Rechner, M. (2008). Gelb oder kein Gelb? Persönliche Verwarnungen im Fußball als Kalibrierungsproblem [To award or not to award a yellow card? Personal warnings in soccer as a calibration problem]. *Zeitschift für Sportpsychologie* [*Journal for Sport Psychology*]*, 15*, 1-11.

Morris, G., & O'Connor, D. (2017). Key attributes of expert NRL referees. *Journal of Sports Sciences, 35*, 852-857.

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

Nevill, A. M., Hemingway, A., Greaves, R., Dallaway, A., & Devonport, T. J. (2017). Inconsistency of decision-making, the Achilles heel of referees. *Journal of Sports Sciences, 35*, 2257-2261.

Neville, T. J., Salmon, P. M., & Read, G. J. M. (2016). Analysis of in-game communication as an indicator of recognition primed decision making in elite Australian Rules football umpires. *Journal of Cognitive Engineering and decision making, 11*, 81-96.

Oudejans, R. R. D., Bakker, F. C., Verheijen, R., Gerrits, J. C., Steinbrückner, M., & Beek, P. J. (2005). How position and motion of expert assistant referees in soccer relate to the quality of their offside judgments during actual match play. *International Journal of Sport Psychology, 36*, 3-21.

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

Page, K., & Page, L. (2010). Alone against the crowd: Individual differences in referees' ability to cope under pressure. *Journal of Economic Psychology*, *31*, 192-199.

Pina, J., Passos, A. M., Carvalho, H., & Maynard, M. T. (2019). To be or not to be an excellent football referee: different experts' viewpoints. *Journal of Sports Sciences, 37*, 692-700.

Pizzera, A., & Raab, M. (2012). Perceptual judgments of sports officials are influenced by their motor and visual experience. *Journal of Applied Sport Psychology, 24*, 59-72.

Plessner, H. (1997). *Urteilsverzerrungen bei Kampfrichtern im Kunstturnen – Der Einfluß von Erwartungen* [*Judgement distortions of judges in artistic gymnastics – The influence of expectations*]. Aachen: Shaker.

Plessner, H. (1999). Expectation biases in gymnastics judging. *Journal of Sport and Exercise Psychology, 21*, 131-144.

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

Plessner, H., & MacMahon, C. (2013). The sports official in research and practice. In D. Farrow, J. Baker & C. MacMahon (Eds.), *Developing sport expertise: Researchers and coaches put theory into practice, 2nd Edition* (pp. 71-95), London, UK: Routledge.

Plessner, H., & Raab, M. (1999). Kampf- und Schiedsrichterurteile als Produkte sozialer Informationsverarbeitung [Umpire and referee judgments as products of social information processing]. *Psychologie und Sport* [*Psychology and Sport*], *Schorndorf 6, Heft 4*, 130-145.

Plessner, H., Schweizer, G., Brand, R., & O'Hare, D. (2009). A multiple-cue learning approach as the basis for understanding and improving soccer referees' decision-Making. In M. Raab, J. Johnson & H. Heekeren (Eds.), *Progress in brain research, Mind and motion: The bidirectional link between thought and action* (pp. 151-158). Amsterdam: Elsevier Press.

Poolton, J. M. Siu, C. M., & Masters, R. (2011). The home team advantage gives football referees something to ruminate about. *International Journal of Sports Science & Coaching, 6*, 545-552.

Praschinger, A., Pomikal, C., & Stieger, S. (2011). May I curse a referee? Swear words and consequences. *Journal of Sports Science and Medicine, 10*, 341-345.

Raab, M. (2015). SMART-ER: A Situation Model of Anticipated Response consequences in Tactical decisions in skill acquisition — Extended and Revised. *Frontiers in Psychology, 5*:1533. https://doi.org/10.3389/fpsyg.2014.01533

Raab, M., & Johnson, J. G. (2004). Individual differences of action-orientation for risk-taking in sports. *Research Quarterly for Exercise and Sport, 75*, 326-336.

Russell, G. W. (2008). *Aggression in the sports world: A social psychological perspective.* New York, NY: Oxford University Press.

Russell, S., Renshaw, I., & Davids, K. (2018). How interacting constraints shape emergent decision-making of national-level football referees. *Qualitative Research in Sport, Exercise and Health, 11*, 573-588.

Scheer, J. K., Ansorge, C. J., & Howard, J. (1983). Judging bias induced by viewing contrived videotapes: a function of selected psychological variables. *Journal of Sport Psychology, 5*, 427-437.

Schwarz, W. (2011). Compensating tendencies in penalty kick decisions of referees in professional football: Evidence from the German Bundesliga 1963-2006. *Journal of Sports Sciences, 29*, 441-447.

Schweizer, G., & Plessner, H. (2016). The accuracy-adequacy model: A theoretical perspective for understanding referees' decisions. *Research Quarterly for Exercise and Sport, 87*(S1), S82.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P. et al. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results [Online version published by the Department of Economics, University of Maryland, under the title: Crowdsourcing data analysis: Do soccer

referees give more red cards to dark skin toned players?]. *Advances in Methods and Practices in Psychological Science, 1*, 337-356.

Slack, L. A., Maynard, I. W., Butt, J., & Olusoga, P. (2013). Factors underpinning football officiating excellence: Perceptions of English Premier League referees. *Journal of Applied Sport Psychology, 25*, 298-315.

Sloman, S. A. (1996). The empirical case of two systems of reasoning. *Psychological Bulletin, 119*, 3-22.

Smith, E. R., & DeCoster, J. (2000). Dual-Process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review, 4*, 108-131.

Souchon, N., Cabagno, G., Rascle, O., Traclet, A., Dosseville, F., & Maio, G. R. (2009). Referees' decision making about transgressions: The influence of player gender at the highest national level. *Psychology of Women Quarterly, 33*, 445-452.

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

Ste-Marie, D. M. (2003). Memory biases in gymnastic judging: Differential effects of surface feature changes. *Applied Cognitive Psychology, 17*, 733-751.

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

Ste-Marie, D. M., & Valiquette, S. M. (1996). Enduring memory-influenced biases in gymnastic judging. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1498-1502.

Ste-Marie, D. M., Valiquette, S. M., & Taylor, G. (2001). Memory-influenced biases in gymnastic judging occur across different prior processing conditions. *Research Quarterly for Exercise and Sport, 72*, 420-426.

Stone, J., Perry, Z. W., & Darley, J. M. (1997). "White men can't jump": Evidence for the perceptual confirmation of racial stereotypes following a basketball game. *Basic and Applied Social Psychology, 19*, 291-306.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review, 8*, 220-247.

Strack, F., Werth, L., & Deutsch, R. (2006). Reflective and impulsive determinants of *Journal of Consumer Psychology, 16*, 205-216.

Sutter, M., & Kocher, M.G. (2004). Favoritism of agents – the case of referees' home bias. *Journal of Economic Psychology, 25*, 461-469.

Unkelbach, C., & Memmert, D. (2008). Game-management, context-effects and calibration: The case of yellow cards in soccer. *Journal of Sport and Exercise Psychology, 30*, 95-109.

Unkelbach, C., & Memmert, D. (2010). Crowed noise as a cue in referee decisions contributes to the home advantage. *Journal of Sport and Exercise Psychology, 32*, 483-498.

Van Quaquebeke, N., & Giessner, S. R. (2010). How embodied cognitions affect judgments: Height-related attribution bias in football foul calls. *Journal of Sport and Exercise Psychology, 32*, 3-22.

Werger, J. (2017). Decision making in ice hockey referees: Officiating style and accurate detection of penalties. Thesis submitted for the master degree of Arts. University of British Columbia, Vancouver, Canada.

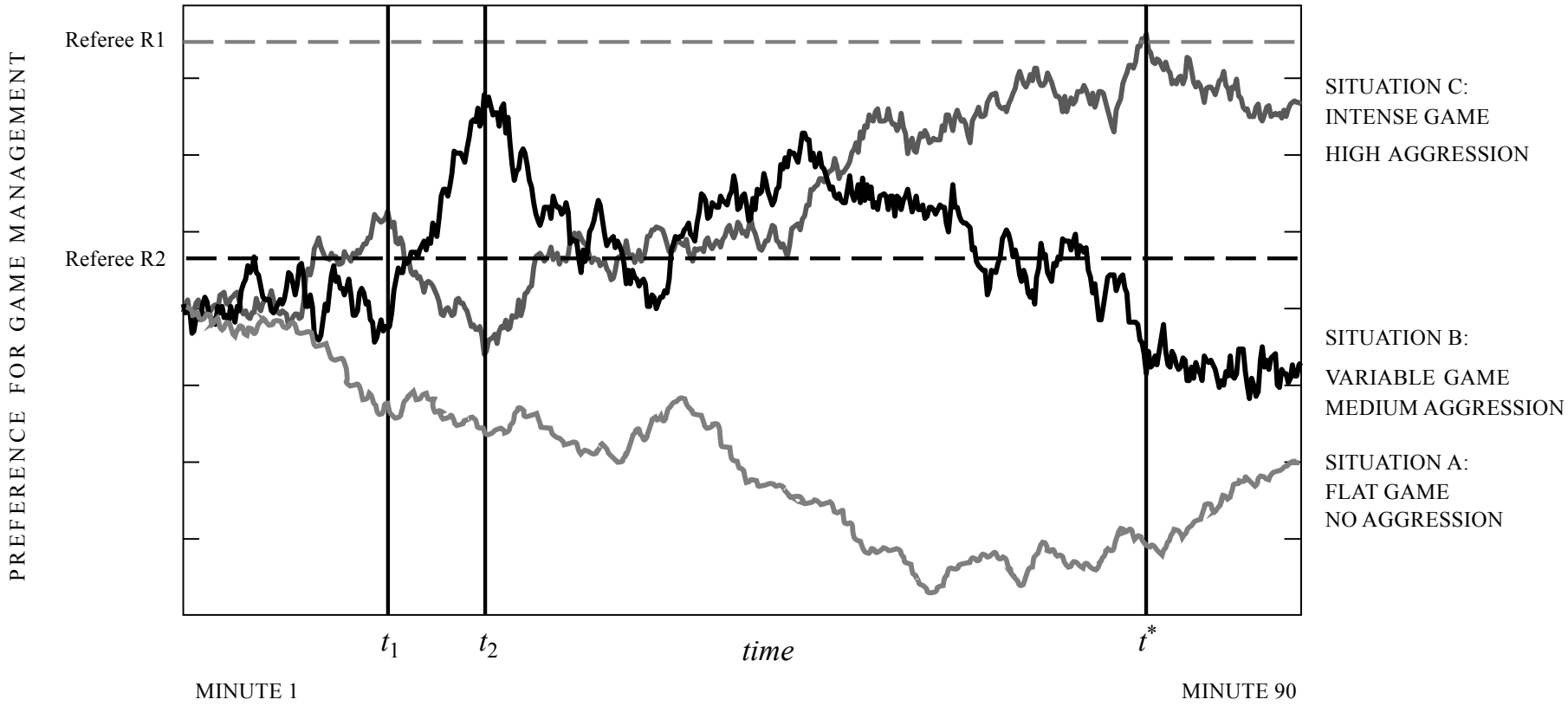
Table 1. List of empirical studies on referees' decision making based on accuracy and adequacy

Goal	Approach	Premise	Components	Example effects shown in literature	Example references
Accurate actions should be called according to the laws	Fiedler and Bless' (2000) social information processing model of decision making (multiple domains)	Mental processes follow a standard sequence The goal is accuracy, errors are due to break down at one stage of the process of information processing	Perception*	Eye movement and visual processing (e.g., peripheral vision, speed of shape recognition) differences by expertise level Influence of positioning on decisions	Ghasemi et al. (2009) –visual skills differences between experts versus novices Hancock & Ste-Marie (2013) – gaze behaviors among higher- and lower-level referees Pizzera & Raab (2012) – influence of motor and visual experiences De Oliveira et al. (2011); Mallo et al. (2012) – distance from the foul situations Oudejans et al. (2000); Oudejans et al. (2005) – positioning relative to the offside line and viewing angle
			Categorization	Perceived information assigned a meaning A number of factors have been shown to influence judgments that may or may not be associated with performance	Plessner (1997, 1999) – order effects and expectations Kolbinger & Stöckl (2019) – reaction to trivial offences Frank & Gilovich (1988); Hagemann et al. (2008) – color of jerseys Unkelbach & Memmert (2008) – time in the game Dawson et al. (2020) – TMO study and influence of crowd
			Memory organization	Systematic memory distortions can influence the accuracy of decisions. Previous decisions can influence subsequent decisions	Ste-Marie & Lee (1991) – perceptual fluency effects: memory for warm up Ste-Marie (2003); Ste-Marie & Lee (1991); Ste-Marie & Valiquette (1996); Ste-Marie et al. (2001) – memory influences of prior information Plessner & Betsch (2001); Schwarz (2011) – previous own decisions
			Judging and deciding	Judgments can be influenced by accompanying information (e.g., stereotypes, competition order, decisions by other judges, social sources of information) Relevant available information is not systematically considered	Jones et al. (2002); Plessner (1999) – reputation of an athlete or a team Boyko et al. (2007); Sutter & Kocher (2004) – home bias Stone et al. (1997) – stereotypes about race Van Quaquebeke & Giessner (2010) – players' height Silberzahn et al. (2018) – skin tone Souchon et al. (2009) – stereotypes about gender Scheer et al. (1983) – conformity effect Levitt & Tockman (1991) – coaching staff behavior Balmer et al. (2007); Nevill et al. (2002); Page & Page (2010); Unkelbach & Memmert (2010) – effect of crowd Plessner (1997, 1999) – heuristic judgment based on overall impression

	Multiple cue probabilistic model	The goal of decisions is accuracy: learn weighting of cues; design of decision making training tool	Identification of environmental cues, weighting of cues and exclusion of underweighted cues to reach solution	Choosing between a limited number of cues according to their degree of validity	**Brand et al. (2009) – cue-based learning
Adequate actions (not necessarily accurate)	Calibration effects in soccer decision making	Referee foul decision making behavior is driven by a desire for consistency across a game; each game has its own scale Consistency is the main goal. "Because absolute judgments are concerned with subjective impressions only, there can be no right or wrong answers. The only criterion judges can use is the internal consistency of their own responses" (Haubensak, 1992, p. 304)	Calibration of judgments from early in the game to achieve consistent decision making as a whole across a game	Early fewer yellow and red cards	***Memmert et al. (2008); Unkelbach & Memmert (2008) – game-management, context-effects and calibration ***Bar-Eli et al. (2006) – red cards issued and scoring goals

Note:
*Unclear which stage drives errors in accuracy; e.g., reputation influences categorization and judging. Not specific to refereeing or to game management versus rule administration choice – does not help explain differences in choices between Fig. 2 situations A (flat), B (fluctuating aggression), C (high aggression, constant).
** Time pressure of referee decision making makes it difficult to give weights to all of the various cues recognized in the environment
*** Does not explain re-calibration or shifts and what prompts shifts. Does not account for inter-individual differences and inter-individual on different time scale

1


PREFERENCE FOR GAME MANAGEMENT
Referee R1
Referee R2
SITUATION C:
INTENSE GAME
HIGH AGGRESSION
SITUATION B:
VARIABLE GAME
MEDIUM AGGRESSION
SITUATION A:
FLAT GAME
NO AGGRESSION
t_1
t_2
time
t^*
MINUTE 1
MINUTE 90

1