Game Management, Context Effects, and Calibration: The Case of Yellow Cards in Soccer

Christian Unkelbach1 and Daniel Memmert2
1Ruprecht-Karls-Universität and 2University of Heidelberg

Referees in German first-league soccer games do not award as many yellow cards in the beginning of a game as should be statistically expected. One explanation for this effect is the concept of game management (Mascarenhas, Collins, & Mortimer, 2002). Alternatively, the consistency model (Haubensak, 1992) explains the effect as a necessity of the judgment situation: Referees need to calibrate a judgment scale, and, to preserve degrees of freedom in that scale, they need to avoid extreme category judgments in the beginning (i.e., yellow cards). Experiment 1 shows that referees who judge scenes in the context of a game award fewer yellow cards than referees who see the same scenes in random order. Experiment 2 shows the combined influence of game management (by explicitly providing information about the game situation) and calibration (early vs. late scenes in the time course of a game). Theoretical implications for expert refereeing and referee training are discussed.

Keywords: refereeing, decision processes, decision making, judgmental biases

Imagine a referee in a soccer game who observes Team A’s striker falling in a duel with a defender in Team B’s penalty area; however, the referee could not observe whether it was a foul or the striker just lost his footing. Thus, she does not call the foul and does not award a penalty kick. Five minutes later she observes a foul committed against Team B’s striker in the penalty zone of Team A. Does she call the foul and thereby award a penalty kick for Team B? The answer to this question is at the heart of the discussion about whether referees’ judgments and the following decisions should be made context-free or within the context of the game (Bar-Eli & Raab, 2006; Mascarenhas, O’Hare, & Plessner, 2006). These two positions can be labeled rule administration and game management (Brand & Neß, 2004; Mascarenhas, Collins, & Mortimer, 2002). If rule administration is the correct approach to refereeing, referees in the described situation must award the penalty.
If, however, game management is the correct approach, they should consider what is best for the game’s smooth flow and what might be considered fair. Because the referee did not award a penalty in an ambiguous situation for Team A, she might not award the penalty for Team B either.

Underlying both these situations are the cognitive and psychophysical demands of referees’ judgment. As a general rule in psychology, it is necessary to analyze these basic processes (e.g., perception, memory retrieval, categorization) before higher-order processes are considered (cf. Plessner & Haar, 2006). There are elaborate models in general psychology for judgment tasks like the one described in our example that specify the cognitive processes (e.g., Parducci, 1965, 1968). One such process is calibration: Any observed judgment situation must be assessed according to an internal classification scale or category system (Parducci, 1965). Because this category system or scale needs to be calibrated in relation to a given context (Parducci, 1968), systematic deviations from rule administration occur. Phenomenologically, these deviations mimic what is predicted by a game management approach; however, the calibration effects are more basic and thereby provide a possible alternative explanation for effects ascribed to game management processes (e.g., Brand, Schmidt, & Schneeloch, 2006). Indeed, calibration is a prerequisite for game management, and they are not mutually exclusive. The aim of the current study is to investigate the contribution of both calibration and game management on the observed effects (here: on yellow card decisions; cf. Plessner & Haar, 2006).

Hence, we present a specific model for calibration effects (Haubensak, 1992) and test the predictions of this model in two experiments. The critical investigated behavior is a referee’s decision to award a yellow card in soccer.

First, let us discuss a referee’s judgment situation a little further. Again, if referees follow a rule administration approach, they have to make their decisions in a context-free manner. They have to evaluate each individual situation in isolation, separately from the current match, score, or playing time, and then make a decision according to the rules and regulations in place (Plessner & Haar, 2006). For instance, whether a foul in the penalty area is called and punished with a penalty kick should not depend on whether this team or the opponent team has already been awarded a penalty or not. Previous studies have shown that referees are rarely able to follow this guideline; instead, they make their decisions within the context of the match. For example, Plessner and Betsch (2001) showed that identical foul scenarios in the penalty area are judged differently (penalty kick vs. no foul and no penalty kick), depending on whether the team in question has already been awarded a penalty kick or not. In the first case, a second penalty kick is hardly ever granted. If, however, the opponent team has been awarded a penalty kick, the frequency of the awarded penalties increased enormously. Such context effects have also been reported for ratings in gymnastics (Ste-Marie & Lee, 1991; Damisch, Mussweiler, & Plessner, 2006), for referee judgments in Australian football (Mohr & Larsen, 1998), and for foul decisions in basketball (Brand, Schmidt, & Schneeloch, 2006).

In light of this evidence, the game management approach regarding referee decisions is currently discussed for sport games, which evaluates referees’ decisions in the context of the game (Mascarenhas, Collins & Mortimer, 2002). Accordingly, sequential effects like the one described are not seen as biases, but as deliberate attempts to dynamically manage a game. For example, a referee might know that a foul has been committed and then actively decide to call the foul or not. This decision depends on the context of the game, the specifics of the situation, or what-
ever is deemed best for an optimal flow of the game. Hence, there is a trade-off between the strict administration of rules and regulations and what is considered best for a fair game. For instance, referees could decide not to show a player in Team A the yellow card after committing a hard foul because they were not able to see a foul’s severity accurately by a player in Team B a few minutes earlier and therefore did not penalize it. However, before such deliberate and controlled behavior is possible, referees need to develop and calibrate an internal judgment scale that measures whether a foul actually falls into the category “yellow card.” This scale development and calibration is a rather automatic process, which needs time and the possibility to apply the scale (i.e., to observe fouls).\(^1\) Only after the assessment against an internal scale can a referee decide whether awarding or not awarding a yellow card will be best for the flow of the game. If the scale is fuzzy to begin with or the classification of a foul is not clear, a referee would probably not award a yellow card, just because the prerequisite is not fulfilled. The present claim is that this is precisely what happens in the beginning of a game: referees need to calibrate their internal scale (i.e., develop a feeling for the game) anew each time they umpire a new match.

In the meantime, Brand, Schmidt, and Schneeloch (2006) presented empirical evidence that basketball referees categorize fouls in line with the game management approach. In an experiment, 121 top-level basketball referees were shown series of videotaped contact situations. Half of the participants saw the scenes in the original game sequence, as they actually occurred in the match. The other half saw the same scenes in random succession. Referees who watched the scenes without the sequential context awarded significantly more rigorous sanctions (e.g., flagrant instead of personal fouls) than their colleagues who evaluated the same scenes in the original order. This result was in line with the game management approach. However, the concept of scale calibration offers a somewhat different explanation for the harder punishment of fouls presented in random order, as in the Brand et al. (2006) report, and the sequential effects for penalty decisions in soccer in the experiment by Plessner and Betsch (2001).

The calibration argument is as follows. Referees’ decisions can be construed as categorization tasks—for a given game situation, they classify scenes into categories of “no foul,” “foul,” “severe foul punishable with a yellow card,” and so forth. Such categorization tasks have received much attention in general psychology (Parducci, 1965; Parducci & Wedell, 1986). A simpler substitute to these elaborate models is the consistency model by Haubensak (1992), which we apply to the problem of awarding a yellow card in soccer. This model makes two assumptions. First, for a given stimulus series, judges develop and calibrate a judgment scale and, second, they apply this scale consistently: “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).

Let us illustrate this model with a standard setup of a psychophysical experiment, in which participants’ task is to classify rectangles according to their size into the categories large, medium, and small. When participants are shown the first rectangle, they have no clue yet how to apply the category system—will the following rectangles all be smaller or larger? If participants place the first rectangle into the category “large,” all same-size or larger rectangles need to be categorized as “large” as well. If the first rectangle, however, was the smallest of the series, the
usage frequency of the categories will be heavily distorted. Thus, judges preserve their degrees of freedom until they feel that their scale is calibrated. Naturally, one might have a rule-based scale about what constitutes a “large” or “small” rectangle. For example, when the first rectangle is presented on a computer screen and it almost fills the entire screen, it can be classified as “large.” Nevertheless, within the limits of the screen, the category system needs to be calibrated. As a further illustration of this necessity, imagine a height check in a school class: An examiner measures all students one by one and might have a perfect scale in centimeters or feet; but to categorize a taken measure as small, medium, or large, the examiner must wait until a given proportion of the class has been assessed, that is, until the category system is calibrated.

Applying this model to a referee’s task of awarding yellow cards in a soccer game is straightforward: Instead of judging the size of rectangles, a referee must judge the severity of a foul. And because referees are not calibrated in the beginning of the stimulus series (i.e., the game), they avoid the extreme categories (i.e., the yellow card). One might argue now that the two described judgment situations differ enormously. First, whereas participants have no idea about the sizes of the upcoming rectangles, referees most likely have a rule-based scale available. This scale could be a summary of previous experiences as a referee or the rules and regulations of a given sports game. However, as discussed in the previous paragraph, the existence of a scale does not preclude the necessity of calibrating the category system (e.g., no foul—foul—yellow card—red card). Second, a referee is not in the luxurious situation of a participant in a psychological laboratory. A game situation contains noisy and unpredictable manifestations. Yet the necessity for calibration and the resulting effects should be stronger the less controlled the environment is and the noisier the stimulus input (e.g., Wedell & Parducci, 1988). Thus, what appears on a phenomenological level as game management might actually be caused by more basic calibration processes. And as mentioned above, even if willful game management is the explanation proper, calibration is a necessary prerequisite: Before referees can actively decide to award a yellow card or not, they have to categorize a foul as a “yellow card” foul.

An additional assumption to fit the model to the game situation is that yellow cards or severe punishments are used sparsely and moderately. This is the case empirically (for example, we found a mean of 4.11 cards across 1,836 German first-league games; Memmert, Unkelbach, Rechner, & Ertmer, submitted), and is advised theoretically (personal communication with Egon Striegl, May 23rd, 2007, first training referee of the German Football Association; German Football Association, 2005, p. 147). According to the consistency model, judges should avoid extreme categories to both sides, which also entails the classification of a rectangle (i.e., a foul) as “small.” Yet there is an asymmetry of omission and commission in decisions (Spranca, Minsk, & Baron, 1991). In a psychophysical experiment on size judgments, participants should avoid both the “small” and “large” categories in the beginning. Referees, however, should avoid only the “large” category because in the context of the game it is not experienced as a consistency violation if a referee does not punish a foul with a yellow card (an omission), and later a comparable foul is punished. However, it is experienced as a consistency violation if a foul is punished with a yellow card (a commission), and later a comparable foul is not (Spranca, Minsk, & Baron, 1991). Again, referees must try to preserve
their judgmental degrees of freedom, and omissions preserve them whereas the commissions reduce them. Without this additional assumption, the model simply predicts that the decisions of a referee during the beginning of a game should deviate more from a normative correct standard than during the end phase of a game (see also Memmert et al., 2007).

Construing game situations as a series of categorization tasks in this way implies that there should be fewer yellow cards in the beginning of a soccer game compared with the rest of the game. This is exactly what we found in a statistical examination of 6 years of German first-league soccer games (1,836 games). Although there are many pragmatic reasons why there should be fewer yellow cards at the beginning of the game, a comparison across different playing times strongly suggests that fewer yellow cards are awarded during the beginning of a game than is objectively warranted or statistically expected (Memmert, Unkelbach, Rechner, & Ertmer, 2007). Across 1,836 games, only 606 yellow cards were awarded in the first 1–15 min, whereas for the 16–30, 31–45, 46–60, 61–75, and 75–90 min blocks, 1,175, 1,397, 1,250, 1,453, and 1,505 yellow cards were awarded, respectively. Possible alternative explanations for this pattern and more evidence for a calibration interpretation of this pattern are presented elsewhere (Memmert et al., 2007).

Applied to referees’ decisions in basketball or soccer, the consistency model predicts the results found by Brand et al. (2006) when scenes are judged in random order in comparison with the chronological order. Referees in the random order condition do not need to preserve their degrees of freedom, but judge according to the rules. However, if shown in chronological order, the scenes are construed as a stimulus series and hence, owing to calibration effects, less severe punishments are awarded in the beginning of a sequence than toward the end. Again, it is important to note that although the calibration idea presents a different explanation, it is not a process that competes against game management, but is instead a prerequisite. Thus, some effects ascribed to referees’ willful and controlled game management might be due to basic and automatic calibration processes.

Two experiments were designed to test the contributions of calibration and game management on referees’ yellow card decisions. The first experiment presents a conceptual replication of the study by Brand and colleagues (2006), with the design of random vs. chronological order of scenes applied to yellow card judgments in soccer. In the second experiment, we try to disentangle the two explanations, building on the notion that game management is a conscious and controlled strategy, whereas calibration is a more automatic process (Haubensak, 1992; Parducci & Wedell, 1986; cf. also Note 1 herein). Hence, referees were given explicit information about the playing time, which should influence a deliberate game management process. To test calibration, we manipulated whether foul scenes stemmed from the end or the beginning of a game. In the beginning, the actual referee should not be calibrated. As a result, decisions in the experiment should deviate more from actual decisions in the beginning of a game compared with the end of a game.

**Experiment 1**

In Experiment 1, the issuing of personal cautions was embedded in a game management design, as presented by Brand et al. (2006). Referees saw foul scenes either in chronological order, as they occurred in a match, or in a random sequence. Analo-
gous to the results obtained by Brand and colleagues (2006), we expected fewer yellow cards for the chronological sequence compared with the random sequence; this would replicate the game management effect for awarding yellow cards in soccer. The consistency model, however, also predicts the effect. In the beginning of a game referees must calibrate their category system to the sequence and, to preserve their judgmental degrees of freedom, they avoid the yellow card category. To test the calibration hypothesis over and above the game management hypothesis, referees were shown 10 scenes from two matches. The calibration hypothesis would be corroborated if referees awarded fewer yellow cards in the chronological condition in the first match only (i.e., first 10 scenes), compared with referees who had seen the same scenes in random order. This pattern should disappear in the second match (i.e., scenes 11–20) because the referees were already calibrated to the stimulus series by the first 10 scenes in the chronological condition.

Method

Participants and Design. Participants were 24 male referees of the German Football Association (DFB), who were recruited during a workshop at a DFB training center. The mean age was 22.9 years and ranged from 16 to 45, and they had refereed across various levels of DFB leagues (depending on their age). They were randomly assigned to one of the two between-participants conditions, namely, the random presentation of the scenes or the chronological presentation. All referees gave their consent to participate in the experiment at the beginning of the session.

Materials. We had an active soccer player select 20 scenes from two German first-league games (10 from each) that showed fouls that could be punished with a yellow card. None of the scenes called directly for a yellow card (e.g., no holding of the shirt, no tackling from behind). In these 20 scenes, one yellow card was awarded in the first 10 scenes and three yellow cards were awarded in the second set of 10 scenes. Note, however, that the actual decisions did not matter because only the variations in decisions between the two experimental conditions were of interest. The 20 scenes were digitalized and cut in a way that the actual decision was not included. Then they were implemented in a computer program that could present them in either chronological or random order. For the chronological condition, the program displayed the actual playing time before each scene. Parallel with this program, we constructed a 20-page questionnaire. On each page, the number of the scene was given together with two check boxes, of which one was labeled “yellow card” and the other “no yellow card.”

Procedure. The experiment was conducted in a session room of the DFB training center. After a brief introduction, a randomly selected half of the participants were sent outside. First, the random condition was conducted. Participants were seated in front of a large projector screen and handed the questionnaire. The experimenter informed them that the research project was concerned with factors that lead to yellow card decisions in soccer. They were told that they would see a random sequence of scenes and were asked to tick the respective box for their decision on the provided questionnaire. Then the computer program was started. Following each scene, the referees decided whether to award a yellow card or not. The session lasted...
16 min. After finishing the 20 scenes, they were told to wait outside and the second group was called into the room. The instructions for the chronological group were largely the same, with three minor changes: First, the referees were told that they would see 20 scenes in chronological order from two games. Second, the actual playing time was displayed previous to each scene. Finally, as a manipulation check, they were instructed to write down the minute of the game when the scene took place before making a decision. The session lasted 17 min.

After rating the 20 scenes, the random condition participants were called in and the experimenter thoroughly debriefed both groups about the hypothesized calibration and game management effects and thanked them for their participation.

**Results.** For all following analyses, neither age nor experience showed up as significant covariates. Because only 11 referees left the room at the start of the random condition, there were 11 referees in the chronological and 13 in the random condition. Prior to analysis, we deleted the data from the scenes where all judges in both conditions decided not to award a yellow card; in other words, scenes in which there was no uncertainty involved. This was the case for three scenes. We computed the mean of yellow card decisions for both games, with a yellow card decision coded as 1 and a no card decision as 0. These means are displayed in Table 1 and can be interpreted as the average probability of awarding a yellow card for a given scene. Note that participants in the random condition saw the scenes not as a coherent game; the means are only organized in that fashion for ease of presentation. These data were analyzed using a 2 (condition: chronological vs. random) × 2 (game: first vs. second) mixed ANOVA with repeated measures on the second factor. Foremost, this analysis yields a main effect for condition, \( F(1, 22) = 4.45, p < .05, d = 0.90 \). Referees in the random condition awarded a yellow card with higher probability (\( M = .444, SD = .085 \)) than referees in the chronological condition (\( M = .377, SD = .067 \)). The game factor was not significant, \( F < 1, ns \). The interaction was also not significant, \( F(1, 22) = 2.62, p < .12 \). However, our prediction was concerned with an existing difference for the first 10 scenes and no difference for the last 10 scenes. The interaction parameter in a 2 × 2 ANOVA design tests a symmetrical effect for all four cells (Rosnow & Rosenthal, 1995), which would indicate an unpredicted reversal of the difference in the first 10 scenes for the second game. Thus, we tested the mean probabilities separately for the two games. This analysis shows that the condition effect in the overall ANOVA is largely due to the difference in the first game, \( t(22) = -2.83, p < .01, d = 1.21, \) and not to the second game, \( t(22) = -0.56, ns \).

**Table 1** Mean Probability to Award a Yellow Card in the Chronological and the Random Condition from Experiment 1, Separately for the Two Games (SD in Parentheses)

<table>
<thead>
<tr>
<th>Condition</th>
<th>First game</th>
<th>Second game</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chronological (( n = 11 ))</td>
<td>.364 (.075)</td>
<td>.391 (.104)</td>
</tr>
<tr>
<td>Random (( n = 13 ))</td>
<td>.473 (.107)</td>
<td>.415 (.107)</td>
</tr>
</tbody>
</table>

*Note.* The labels *first* and *second* game are only of relevance for the chronological condition because participants in the random condition saw all scenes in a random order.
Discussion

First, we replicated the result by Brand and colleagues (2006). Referees were more rigorous in their decisions when scenes from a game were presented in random order compared with scenes presented in chronological order. Whereas Brand and colleagues (2006) explained this effect with the concept of game management, we propose a more basic calibration process. Indeed, the data suggest that the game management explanation is not sufficient: We only found significantly less yellow cards in the chronological condition for the first 10 scenes. A pure game management approach predicts equal differences for both the first and the second set of 10 scenes. Following the consistency model, however, we can assume that referees in the chronological condition were calibrated after 10 scenes and then showed almost the identical pattern as the referees in the random condition. Yet, in hindsight, we have to acknowledge that the former data pattern would have been possible, depending on whether the referees construed the 20 scenes as one set or as two independent sets. Because we did not check for this differentiation, we can only infer from the observed pattern that the referees construed the 20 scenes as one set, thereby showing no calibration effects within the second game.

This problem exemplifies again that on a phenomenological level, game management and calibration predict the same results. Moreover, even with the present differentiation, one might still explain the observed pattern as a game management effect. Referees in the chronological condition may have felt that if they were in the suggested game situation, they would have not awarded a yellow card, although they consciously knew that a yellow card would have been in order. It just went against their idea of a smoothly flowing game. The question of whether game management or calibration is responsible for the observed effects—and not only the present effects, but also Brand et al.’s (2006) as well as Plessner and Betsch’s (2001)—comes down to this: Game management is a controlled process that requires deliberate attention, whereas calibration is a more automatic process within a given series of decisions. To test the interplay of these explanations, we conducted a second experiment.

Experiment 2

Are the observed effects due to a consciously controlled game management process or due to the necessity to calibrate a judgment scale or category system? To answer this question, the setup of the experiment was as follows: First, it is probably impossible to perfectly simulate the proposed calibration process in an artificial situation. Thus, we planned to take referees’ actual yellow card decisions (yellow card vs. no yellow card) in soccer games as comparison standards. To compare an ongoing calibration with a completed calibration, we selected scenes from the first 15 min (early scenes) of soccer games and the 60th to 75th minutes (late scenes). Without any knowledge about which game is taking place or the minute in which a particular scene is happening, referees in an experiment should agree more with the actual decisions in the late scenes compared with the early scenes. Referees in the actual game are still in their calibration phase and need to preserve their degrees of freedoms, whereas referees in an experiment can base their decision on the rules and regulations proper, without concerns for judgmental degrees of freedom for
following decisions. Thus, referees who watch the scenes without context should deviate more from the actual decisions in the beginning of the game (although they do not know that it is the beginning) and converge more with the actual decisions toward the end of the game.

To test conscious game management, we planned to display the playing time of the game ahead of each scene. In contrast to the previous experiment, this time was not veridical, but was selected to indicate that a scene was early or late from a game. Thus, half of the actual early scenes were randomly assigned to the status of late scenes, and the other half to the status of early scenes. The game management concept would be strongly supported if playing time, as a mere propositional fact, exerts an influence on the decision pattern: Referees who see the scenes without context should award more yellow cards when the nonveridical time indicates a late scene compared with nonveridical early scenes. Such an influence would allow no other interpretation but that referees consciously and strategically avoid awarding yellow cards early in the game.

Accordingly, the present experiment varies three factors, all within participants. First, whether a yellow card was actually awarded or not in a game; second, whether a scene is actually early or late from a game; and third, whether a scene is assigned the status of an early or late scene by means of the displayed playing time. The dependent variable is again the decision to award a yellow card or not when a scene is watched without the context.

Method

Participants and Design. Participants were 19 referees who participated in a 2-day DFB referee-training program. Their mean age was 23.2 years and age ranged from 18 to 31. All of them had at least 2 years of experience in officiating DFB league soccer matches, although at varying levels. All three factors (as described above) were realized as within-participants manipulations. We used seven scenes for each of the eight categories resulting from the factors’ orthogonal combination, yielding a total of 56 scenes. All referees gave their consent to participate in the experiment at the beginning of the session.

Materials. Obviously, the selection of the scenes for the present design is crucial. Hence, we did not select the scenes ourselves. First, a preselection by a group of active soccer players was made. They selected 120 foul scenes from more than 300 games of the Bundesliga and the Third German League, which they deemed hard enough to deserve a yellow card and were either early or late from a game (1–15 min or 60–75 min). This set included no scenes that immediately called for a yellow card (holding/pulling the opponent’s shirt, arguing with the referee, etc.). These 120 scenes were then given to a former first-league referee to select scenes in which it would have been possible not to award a card if a card was actually awarded, and vice versa, to award a card if none was actually awarded. In other words, he was instructed to select ambiguous scenes. This expert was blind toward the actual purpose of the scene selection. Once 15 scenes were selected for each category (early / late by yellow card / no yellow card), the process was stopped. One additional scene was deleted afterward from each category to allow for an equal number of scenes in each of the eight categories. The final set of 56 scenes
was digitalized and implemented in a computer program that could display them in a random order under the constraint that first single scene from all eight categories needed to be shown before there could be another scene from a given category. This program also randomly assigned half of the scenes from each category to the status of “early” scenes and displayed a random playing time between the 5th and 15th minute before these scenes. The other half were assigned the status of “late” scenes and a random playing time between the 75th and 85th minute was shown before these scenes.

To ensure that none of the 56 scenes included cues for the actual playing time, we had 13 sport students judge the scenes. They saw the scenes in a random order, and they were explicitly asked whether they thought a scene to be from the beginning or the end of a game. We analyzed their responses using signal-detection theory (SDT), which provides a very strong test against the hypothesis that the scenes contained no cues. Signal-detection theory delivers a discrimination parameter $d'$, and a $d'$ of zero indicates no discrimination ability at all. The estimated $d'$ (for early vs. late) was $d' = 0.144$ ($SD = 0.377$), which was not significantly different from zero. Hence, we can conclude that the scenes contained no cues about playing time. The same sport students rated the severity of the fouls in the 56 scenes on a scale from 1 (not severe) to 5 (very severe). Again, there was no difference between early ($M = 3.02$, $SD = 0.37$) and late scenes ($M = 2.86$, $SD = 0.39$), $t(12) = 1.65$, ns. If anything, the early scenes show the more severe fouls.

Lastly, a questionnaire similar to the one in Experiment 1 was constructed, which included 56 pages, one for each scene, on which the play time should be noted and one of two boxes should be checked, namely, whether to award a yellow card for a given foul scene or not.

**Procedure.** Participants were received and greeted as a group. They were seated around a huge screen on which the scenes would be displayed by means of a data projector. The experimenter explained that they would take part in a research project to investigate factors that lead referees to award yellow cards. Their task would be to decide for an extended series of foul scenes whether they would award a yellow card or not. The point was stressed that they should make their decisions “as if in the game” and without looking to the other participants and without pondering too much. Then the experimenter handed out the questionnaires and started the computer program. After judging the 56 scenes, the experimenter collected the questionnaires and extensively debriefed the participating referees about the hypothesized game management and calibration effects. The session lasted approximately 28 min.

**Results and Discussion**

As in Experiment 1, neither age nor experience as covariates influenced the results. All referees noted the appropriate displayed playing time on their questionnaire for all 56 scenes. For the decisions, the mean of yellow card decisions within each category was computed, with “yellow card” coded as 1 and “no yellow card” as 0. The resulting means across the 19 referees in the eight categories are displayed in Table 2. Again, these means can be understood as the mean probability to award a yellow card for a foul scene from that category.

We analyzed the data using a 2 (actual decision: yellow card vs. no yellow card) × 2 (actual playing time: early vs. late) × 2 (displayed playing time: 5–15 min vs.
This analysis shows a strong effect for the actual decision, \( F(1, 18) = 143.14, p < .001, d = 5.64 \). The referees awarded a yellow card with greater probability if a card had been awarded in the actual game (\( M = .690, SD = .189 \)) compared with when no card had been awarded (\( M = .314, SD = .198 \)). This effect shows that although we instructed the former first-league referee to select ambiguous scenes, the scenes still contained enough cues that the participating referees overall concurred with the decisions made on the pitch.

Of greater interest is a main effect of displayed playing time on the decisions, \( F(1, 18) = 9.61, p < .01, d = 1.46 \). If a scene was supposedly from the beginning of a game (displayed minute between 5 and 15), participants were less likely to award a yellow card (\( M = .464, SD = .273 \)) compared with scenes that were supposedly from the end phase of a game (displayed minute between 75 and 85; \( M = .539, SD = .262 \)). This effect is clear evidence for conscious and willful game management: Scenes were judged differently based on the mere propositional fact that a scene stemmed supposedly from the beginning or the end of a game. The effect is qualified by an interaction of displayed playing time and actual decision, \( F(1, 18) = 10.83, p < .01, d = 1.55 \). If a foul was actually punished with a yellow card, there was not much impact of the playing time factor (\( M = .684, SD = .153 \) vs. \( M = .695, SD = .221 \); for supposed beginning and end of a game, respectively). However, if a foul was actually not punished with a yellow card, referees deviated from that judgment a lot more when they thought that a scene stemmed from the end of the game (\( M = .383, SD = .202 \)) than when they thought that a scene was from the beginning (\( M = .244, SD = .169 \)). One interpretation of this result is that the game management influence is stronger when the ambiguity in a given scene is higher, as in the “no yellow card” scenes, compared with less ambiguity for the fouls in the “yellow card” scenes. In other words, referees show game management effects only if situations provide the necessary freedom for their decisions. This stronger impact for more ambiguous scenes is well in line with previous research; for example, Hill and Barton (2005) only found an impact of wrestler’s shirt colors (red vs. blue) when the opponents were comparable in their skill, and thus, the bouts were ambiguous. In fact, room for variation is a requirement for every psychological test item (here, a given scene): To be diagnostic, some need to solve it (awarded a yellow card) whereas others do not. If an item does not allow for variation, it cannot discriminate and hence, it is not diagnostic (cf. Amelang & Schmidt-Atzert, 2002, p. 110ff.). Thus, the probability for a yellow card increases

<table>
<thead>
<tr>
<th>Actual yellow card</th>
<th>No actual yellow card</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actually early (1–15)</td>
<td>.677 (.133)</td>
</tr>
<tr>
<td>Actually late (60–75)</td>
<td>.692 (.173)</td>
</tr>
</tbody>
</table>

Note. Higher numbers represent a higher probability of awarding a yellow card within a given category.
the more clearly a foul is punishable with a yellow card, independent from previous decisions or game management considerations.

Yet there was also an interaction of actual decision and whether a scene was actually early or late from the game, $F(1, 18) = 5.67, p < .05, d = 1.12$. This effect is easier to interpret as a deviation probability; we computed this score in a way that it became large if our referees did not concur with the actual decision in the game. As expected, the referees deviated more from the actual decisions in the beginning of a game ($M_{dev} = .344, SD = .081$) than for decisions toward the end of a game ($M_{dev} = .280, SD = .098$). This is exactly the pattern that is predicted by the calibration hypothesis and presents a perfect replication of an experiment reported in Memmert et al. (2007). Note that for the referees in this experiment, there was no indication whether a scene was actually from the beginning or the end, which precludes the idea of willful control. Rather, the effect shows up despite the direct manipulation of the supposed playing time.

Finally, the combined contributions of conscious game management and the need for calibration is evinced by a strong three-way interaction, $F(1, 18) = 12.64, p < .01, d = 1.68$. It is not easy to see the clear pattern of this interaction from Table 2, but to point out the main cells should suffice. Referees agreed with the actual decisions most when the factors acted in accordance; that is, most yellow cards were given when actual late scenes with actual yellow cards were labeled as late (.789) and the least cards were awarded when actual early scenes with no actual yellow cards were labeled as early (.211). If we compute these means as deviation probabilities, as above, and rearrange them as suggested by Shaffer (1977), the three-way interaction becomes visible as a main effect: Referees in the experiment deviated much more from the factual decisions when displayed and actual playing time differed ($M_{dev} = .357, SD = .193$) than when they concurred ($M_{dev} = .267, SD = .183$). Thus, conscious game management and calibration both contribute to the effect that fewer yellow cards are awarded in the beginning of a soccer match compared with the rest of the game.

General Discussion

We started out by conceptualizing the situation of a referee in a soccer game as a categorization task and showed how a model from general psychology, the consistency model by Haubensak (1992), can be applied to the game situation. This model is built to explain fundamental perceptual and cognitive processes in categorization tasks and, accordingly, should be applicable to a wide range of decision situations in sports. We applied it to the case of yellow card decisions in soccer, offering an alternative explanation to the game management interpretation of previous results (Brand et al., 2006; Plessner & Betsch, 2001). The difference between these two approaches is that, according to game management (here, applied to yellow card decisions), referees can consciously and willfully decide not to award a yellow card in the beginning of a game (i.e., a series of stimuli), whereas calibration is a more automatic process. Our argument is that it is the necessity of the judgment situation (i.e., the need for calibration) that forces referees to award not as many yellow cards in the beginning as normatively justified. Thus, what might appear on a phenomenological level as a game management effect could actually be due to a more basic calibration process. And as we have discussed above, theoretically, calibration is a prerequisite for game management processes.
The first experiment replicated the results by Brand and colleagues (2006), with the additional effect that for the end of the chronological stimulus presentations, referees’ judgments converged with the judgments of referees in the random condition. It is interesting to speculate whether the same results would have been obtained in Brand and colleagues’ (2006) experiment if their whole sequence of 18 scenes had been presented as one stimulus series and not clearly as four different series by the switch of the focal player. Similarly, it is questionable whether Plessner and Betsch (2001) would have obtained their strong sequential effects for the penalty decision if the scenes had been presented in random as from different games, that is, with new degrees of freedom for each decision. The manipulation of context or coherence of stimulus presentations is a promising variable for future research because it tackles the degrees of freedom and calibration problem.

The manipulation of random vs. chronological presentation, however, cannot differentiate between game management and calibration effects. Yet the second experiment shows that it is no either/or decision, but calibration and game management both play a role for the decision to award a yellow card in soccer. In some sense, this is good news for both positions. The result that willful control plays an important role grants more importance to referees as free individuals who try to direct a game to the best of their competence. On the other hand, the calibration idea offers a venue for intervention if one is to conclude that referees should not calibrate their judgment scale to each specific game. It is noteworthy that the referees in our experiments reported this calibration experience in postexperimental interviews as “learning to read a game,” “getting attuned to a game,” or “developing a feel for the game.” If, as requested by Plessner and Betsch (2002), refereeing should be a craft and not an art, then one might also request that normative, rule-based standards are applied from the first minute of a game and consistently across games, not only within a game, as we have postulated here. Partly, this is already done with the clear definition of offenses that immediately lead to a yellow or red card (tackling from behind, holding the jersey, verbal offenses, etc.). Thus, it should be possible to minimize the need for calibration in each particular game, as it is probably the case with experienced and novice referees (although we found no effect for experience level). To vary referees’ levels of experience systematically is another promising venue for further research and we would expect that expert referees should be less prone to calibration effects. However, we would also argue that it is a good thing that each game receives its own calibrated judgment scale. The vagaries and imperfection of referees’ judgments and decisions in soccer are usually part of what makes soccer such an attractive and fun sport.

Notes

1. One might challenge that this process is indeed automatic; however, if one considers the five production rules of Haubensak (1992, pp. 304–305), no one would assume that people willfully or actively follow these steps. The same is true for Parducci’s original range-frequency theory (1965), in which an observable reaction is based on the range value and the frequency value of a stimulus, of which the computation is also complex. Yet both models predict observable behavior. Furthermore, we would refrain from a strict dichotomy of automatic and controlled, or conscious and unconscious, following the idea of a continuum of mental control (cf. Bargh & Chartrand, 1999; Kruglanski et al., 2003). For the present article, however, we will use the terms automatic and controlled.
2. For the first game, the displayed playing times were Minutes 6, 9, 13, 15, 59, 75, 80, 87, 88, and 90. For the second game, the playing times were Minutes 6, 8, 20, 23, 36, 37, 39, 50, 56, and 80.

3. The pattern does not change if these scenes remain in the analysis; however, this is the analytic strategy we used for the data in Memmert & Unkelbach (2007).

4. A problem of the presented research is surely the ecological validity; referees’ judgments were based on video sequences presented via a data projector. Furthermore, these were compared with judgments of the referees who had led the original match. This procedure meets with objections already described by Mascarenhas, Collins, & Mortimer (2002) in their comment on the work by Plessner and Betsch. Specifically, the perspectives of the referees in the experiments were certainly not the same as those of the referees in the original game. The problem of various observation perspectives associated with decision making in sports has been discussed intensively (Mascarenhas, Collins, & Mortimer, 1999; Omodei, Wearing, & McLennan, 1997; Stokes, Kemper & Kite, 1997). It will, however, probably not be possible to solve this issue to a satisfactory extent in future studies investigating referees’ decisions in current games in connection with experimental manipulations.

References


