Perceptual-Cognitive Skills in Offside Decision Making: Expertise and Training Effects

Peter Catteeuw, Bart Gilis, Johan Wagemans, and Werner Helsen
Katholieke Universiteit Leuven

This two-experiment study aims to investigate the role of expertise in offside decision making (Experiment 1) and the effect of perceptual-cognitive training (Experiment 2). In Experiment 1, a video-based offside decision-making task followed by a frame recognition task demonstrated a bias toward flag errors and a forward memory shift for less-successful elite-standard assistant referees that is in line with the predictions from the flash-lag effect. In Experiment 2, an offside decision-making training program demonstrated a substantial progress from pre- to posttest for response accuracy, but not for accuracy of memory in the frame recognition task. In both experiments, no differences were found for visual scan patterns. First, these results suggest that less-successful elite-standard assistant referees are more affected by the flash-lag effect. Second, an off-field perceptual-cognitive training program can help assistant referees to deal with the perceptual consequences of the flash-lag illusion and to readjust their decision-making process accordingly.

Keywords: visual information processing, flash-lag effect, video training

The flash-lag effect refers to the tendency for observers to misjudge a moving object at a discrete moment defined by a time marker (usually a briefly flashed stimulus). Specifically, the object is perceived further forward in the direction of movement. This phenomenon is well studied in laboratory settings (for an...
Offside Decision Making: Expertise and Training

Offside situations in association football (also known as soccer in North America) offer a unique opportunity to study the flash-lag effect in a real-life setting. Assistant referees have to judge the positions of defenders and attackers (often moving at high speed) at the moment of the pass (i.e., equivalent to the flash in the classical flash-lag effect). The attacker receiving the ball is in an offside position when he is closer to the defender’s goal line than both the ball and the second-last defender at the exact moment of the pass (FIFA, 2009). The assistant referee needs to perceive and assess the exact position of the attacker relative to the second-last defender to make a correct decision. A correct perception at the exact moment of the pass is crucial because a small part of the attacker’s body (except for the arms and hands) nearer to the defender’s goal line means the attacker is in an offside position. In this respect, the assistant referee can make two types of errors, a flag error (i.e., assistant referee raises the flag for an attacker in an onside position) or a non-flag error (i.e., assistant referee keeps the flag down for an attacker in an offside position).

In literature, several hypotheses have been introduced to explain errors in judging offside, like moment in the match (Krustrup, Mohr, & Bangsbo, 2002), movement speed (Oudejans, Bakker, Verheijen, Gerrits, Steinbrückner, & Beek, 2005), shift of gaze (Belda Maruenda, 2004; Sanabria, Cenjor, Marquez, Gutierrez, Martinez, & Prados-Garcia, 1998), optical error (Oudejans, Verheijen, Bakker, Gerrits, Steinbrückner, & Beek, 2000) and flash-lag (Baldo, Ranvaud, & Morya, 2002). This study focuses on the flash-lag effect, which was first introduced by Baldo et al. (2002) to explain flag errors. Gilis and colleagues (Gillis, Helsen, Catteeuw, Van Roie, & Wagemans, 2009; Gilis, Helsen, Catteeuw, & Wagemans, 2008) evidenced the flash-lag effect in on- and off-field offside decision-making tasks. The flash-lag effect also appeared to be a viable explanation for flag errors in the 2006 FIFA World Cup (Catteeuw, Gilis, García-Aranda, Tresaco, Wagemans, & Helsen, 2010) and the 2007–2008 season of the English Premier League (Catteeuw, Gilis, Wagemans, & Helsen, 2010).

The question arises whether an assistant referee can overcome this perceptual problem by specific offside decision-making training. First, opportunities to improve offside decision making are scarce (Catteeuw, Helsen, Gilis, & Wagemans, 2009b; Gilis et al., 2008; MacMahon, Helsen, Starkes, & Weston, 2007). Assistant referees can practice almost only on occasions when they actually have to perform, namely the match, and thus, match experience is related to expertise for assistant referees (Catteeuw et al., 2009b). The creation of training opportunities for offside decision making is not evident. On-field exercises with players simulating offside situations are difficult to organize on a large scale. Off-field training methods with video and computer tools are regarded relevant by assistant referees (Catteeuw et al., 2009b), but efficacy needs to be investigated. Second, assistant referees need to be aware of the perceptual problems associated with offside decision making. Their perception can suffer from an incorrect positioning (Oudejans et al., 2000) and the flash-lag effect (Baldo et al., 2002).

From an expertise-research perspective, the degree of development of perceptual skills of assistant referees is linked with the level of expertise. However, a complex task like judging offside entails many different information-processing components, so the question is which components lead to an expertise effect and which can be influenced by training. Abernethy (1986) distinguished three components
in his visual information-processing model (i.e., the perceptual mechanism, the decision mechanism and the effector mechanism). Catteeuw, Helsen, Gilis, Van Roie, and Wagemans (2009a) and Gilis et al. (2008) argued that decision-making skills discriminate between national-standard assistant referees and international-standard assistant referees based on an off-field offside decision-making task. At the same time, Catteeuw et al. (2009a) observed no expertise differences in visual scan patterns for a limited number of assistant referees. Hence, both expertise groups within elite-standard assistant referees detected and selected the appropriate input. They all seem to scan visual stimuli in the same way. So, perhaps they differ in the next step in visual information processing, the strategy formation and response selection (Abernethy, 1996). This step implies active information pick-up and giving meaning to the appropriate input. It is quite likely that expertise in offside decision-making relates to the cognitive correction of perceptual illusions, like the optical error and the flash-lag effect, during the decision-making stage (Catteeuw et al., 2009a; Gilis et al., 2008, 2009).

In line with this view, the current study aims to investigate skill level among elite-standard assistant referees and the potential profits of an offside decision-making training program. Offside events were simulated and videotaped from the perspective of the assistant referee that is positioned on the offside line. In Experiment 1, the aim was to measure response accuracy and accuracy of memory for visual estimates of the spatial position of the attacker relative to the second-last defender. In Experiment 2, a training study was set up to examine to what extent response accuracy and accuracy of memory for visual estimates can be trained.

Experiment 1

Experiment 1 assessed accuracy of offside decision making and accuracy of visual memory for player positions. In offside decision making, estimating the exact spatial position of attacker and defender is crucial. Videotaped offside simulations from the ego-perspective of the assistant referee were used. Every offside simulation was followed by a screenshot with five probe frames. The assistant referees’ task was first to assess offside at the moment of the pass, and second to remember the exact spatial positions of the players at the moment of the pass. The assistant referee had to compare his perception with the five probe frames and select the identical probe frame, similar to the tasks used by Thornton and Hayes (2004).

Our interest was to examine how accurately assistant referees could judge offside simulations and investigate their perception of the players’ positions in a complex dynamic event. In line with previous work (Catteeuw et al., 2009a; Gilis et al., 2008, 2009), we predicted a forward memory shift associated with the flash-lag effect. In addition, the assistant referees were divided in two expertise groups (successful experts and less-successful experts) using the median-split technique. A forward memory shift was predicted for both groups because assistant referees will all be subject to the flash-lag effect to some extent.

In addition, point-of-gaze data were recorded to investigate whether visual scan behavior could account for higher accuracy in offside decision making. In line with research in badminton and squash (Abernethy, 1990, 1991), Catteeuw et al. (2009a) found no differences for visual scan patterns between a small group of international and national elite-standard assistant referees. In contrast, several
studies in different expertise domains showed clear, but ambiguous differences in visual scan patterns between experts and novices. In association football (Helsen & Starkes, 1999), French boxing (Ripoll, Kerlirzin, Stein, & Reine, 1995) and tennis (Goulet, Bard, & Fleury, 1989), experts showed fewer fixations of longer duration. Other studies in association football (Vaeyens, Lenoir, Williams, & Philippaerts, 2007; Williams & Davids, 1998; Williams, Davids, Burwitz, & Williams, 1994) showed that experts have more fixations of shorter duration. In this study, two larger subsamples of the elite-standard assistant referees were tested. Point-of-gaze data were collected for the potential offside situations, but also for the probe frames. Based on previous results, two outcomes are possible in this study: Either differences in visual scan patterns were found between two levels of elite-standard assistant referees, replicating some earlier expertise effects, or no differences in visual scan patterns were found, implying that differences in response accuracy result from different strategies in dealing with the perceptual information (Catteeuw et al., 2009a).

**Methods**

**Participants**

Participants were 16 Belgian and 1 Portuguese elite-standard assistant referees (mean age 39.39 ± 2.70). Eleven of these assistant referees were on the FIFA list for 3.83 ± 2.01 years.

Written consent was obtained from the Belgian referees’ committee. The nature of the study was explained to the assistant referees before testing. They were free to withdraw from testing at any stage. The study was designed and conducted in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki.

**Task**

The assistant referees observed video simulations of offside situations. Five probe frames of the moment of the pass followed each offside situation. First, the assistant referees judged the clip and marked whether it was offside or not. Second, their task was to remember the exact moment of the pass. They marked the frame that matched the moment of the pass.

For this laboratory offside assessment task, offside situations were simulated by youth elite players of a Belgian second division team in Leuven, Belgium (see Catteeuw et al., 2009a,b). A digital high-definition video camera (Sony, XD CAM PDW-700 24p) was placed at the touch line in a way that these video simulations were recorded from the perspective of the assistant referee. Two attackers (i.e., a passer and a receiving attacker) were playing against one defender and one goalkeeper. The players acted in a position they were used to play in their team. Out of a total of 156 offside actions, 40 high-quality situations were selected. First, the selection was based on the position of the defender relative to the camera position. Only those situations exactly in front of the camera were selected to eliminate an incorrect angle of view. Second, we categorized the situations in four groups of 10 situations according the position of the attacker relative to the defender (i.e., the offside line). First, in situations far behind the offside line, the attacker was more than 1 m behind the defender. Second, in situations slightly behind the offside
line, the attacker was less than 1 m behind the defender. Third, in situations on the offside line, the attacker was at the same level of the defender. And fourth, in situations with the attacker slightly ahead of the offside line, the attacker was less than 1 m ahead of the defender.

The offside situations were digitized and edited with the software program Final Cut Pro 6.0.6 (Apple, Inc., US). On average, the offside situations had a duration of 6.4 ± 1.5 s (range 3.4–9.5). The moment of the pass was on average 2.7 ± 1.1 s (range 0.5–5.2) after the start of the situation. The frame rate of the video simulations was 25 Hz.

The pictures with the five probe frames were created with Adobe Photoshop CS2 9.0.2 (Adobe Systems, Inc., US). First, there were seven probe positions for each offside situation with a one-frame difference. Three probes were taken before the pass (–3, –2 or –1 frame[s] before the pass), one probe at the exact moment of the pass (0), and three after the pass (+1, +2 or +3 frame[s] after the pass). Then, the probe frames were reduced to a selection of five probes (–3 to +1, –2 to +2, or –1 to +3). One of the three selections was randomly shown after each situation for 10 s. These five probe frames of the selection were shown in a random order in one screen capture.

**Apparatus**

The offside situations and probe frames were played on a Tobii T120 Eye Tracker (Tobii Technology AB, Sweden). This eye tracker records visual scan patterns at a constant frame rate of 120 Hz using five infrared lights and a camera mounted below the 17-inch screen. In addition, it allows for gaze angles up to 40 degrees in the semicircle above the camera and up to 10 degrees below the camera. It claims an effective accuracy of approximately 1 degree across the entire screen (Tobii, 2006).

**Procedure**

The assistant referees were seated in front of the eye tracker with their heads fixated on a chinrest precisely 60 cm from the monitor. Standardized instructions were read. A simple five-point eye calibration was performed to verify point of gaze. Subsequently, five familiarization trials were shown. For the test, the offside situations were randomly distributed in two sets of 20 clips. For each situation, the assistant referees were asked to judge as accurately as possible, within a 5 s time window after the final pass, whether the attacker was in an offside position saying “offside” or “no offside.” The test leader wrote down the answers. Subsequently, for the five probe frames, the assistant referees were asked to recognize the probe frame matching the exact moment of the pass as accurately as possible saying clearly the number under the probe frame. The test leader wrote this number down. The probe frames were shown for 10 s.

**Dependent Variables and Analysis**

**Response Accuracy Video Clips.** The number of correct and incorrect decisions was calculated. For the onside situations, a distinction was made between correct
non-flag signals and flag errors. For the offside situations, a distinction was made between correct flag signals and non-flag errors. First, a chi-squared test was used to examine the error distribution. Second, the sensitivity index $d'$ and response bias $c$ were calculated (signal detection theory: Macmillan & Creelman, 2005).

**Accuracy of Memory.** First, the number of correctly recognized frames (hits) was calculated for every probe position. Second, the consistency with the response accuracy of the video clips was determined. “Consistent” means that the assistant referee chose a probe frame with an offside position when he judged the situation as offside or that he chose a probe frame with an onside position when he judged the situation as onside. Third, the weighted mean (cf. Thornton & Hayes, 2004) was calculated by multiplying the proportion of responses at a given probe position by that probe’s frame difference ($-3$, $-2$, $-1$, $0$, $+1$, $+2$, or $+3$) from the correct probe (i.e., $0$). These products are then added and divided by the total number of responses to yield a weighted mean.

**Point-of-Gaze Data.** A fixation was defined as the period of time when the eyes remained stationary within 58 pixels or 1 degree of movement tolerance for a period equal to, or greater than, 120 ms (Catteeuw et al., 2009a; Helsen & Starkes, 1999; Savelsbergh, Van der Kamp, Williams, & Ward, 2005; Williams, Davids, & Williams, 1999). To examine the point-of-gaze data, visual fixations in areas of interest were analyzed for the video clips. Two areas of interest were defined for the video clips. Visual fixations on the passer and the offside line were counted as well as the fixations not on the two areas of interest. The number of visual fixations, mean fixation duration, percentage viewing time and number of areas fixated were calculated. The point-of-gaze data were registered from the start of the video clip until the answer of the assistant referee.

**Data Analysis.** The median-split technique was used to determine between-group differences. This procedure divides the assistant referees in two groups based on their accuracy on the offside decision-making task. The assistant referees were assigned to either successful experts (SE) or less-successful experts (LE) if their score was above or below the median value of the group. To study the differences between the SE and LE group on the different variables, independent $t$ tests were used. Effect sizes were calculated with Cohen’s (1988) $d$ to indicate the meaningfulness of any significant differences. Values of 0.2, 0.5, and more than 0.8 represent a small, moderate, and large difference, respectively. For three variables of the point-of-gaze data (number of visual fixations, mean fixation duration, and percentage viewing time), repeated measures analyses of variance (ANOVA) were used in which group was the between-participants variable and areas of interest was the within-participants factor. We checked the sphericity assumption for repeated-measures ANOVAs with Mauchly’s test of sphericity. The Greenhouse–Geisser correction procedure was applied when violations were apparent. Significant effects were examined using Fisher LSD post hoc procedures. Effect sizes were reported as partial eta squared ($\eta_p^2$). To compare the number of areas fixated between both groups, a $t$ test for independent samples was used and Cohen’s $d$ was calculated as a measure of effect size.

All statistical analyses were executed with Statistica version 8.0 (StatSoft, Inc., USA) and significance level was set at 0.05.
Results

Response Accuracy Video Clips

The median value for all participants was 30 correct answers out of 40. Seven assistant referees (SE group) scored above this value (mean 32.86 ± 1.46) and eight assistant referees (LE group) scored below this value (mean 25.75 ± 2.38). Two assistant referees scored exactly 30, that is, the median value. They were not assigned to a group.

For both groups, the type of errors was studied in more detail. The SE group did not show a bias toward one type of error (flag errors: mean 4.86 ± 2.04; non-flag errors: mean 2.29 ± 2.36; \(\chi^2 = 1.31, p = .253\)). The LE group made significantly more flag errors (mean 13.13 ± 2.36) than non-flag errors (mean 1.13 ± 0.99) (\(\chi^2 = 17.79, p < .001\)).

Sensitivity and response bias showed the expected pattern. The sensitivity index \(d'\) was higher for the SE group (\(d' = 1.88, 95\% \text{ CI} 1.38–2.38\)) than for the LE group (\(d' = 1.38, 95\% \text{ CI} 1.05–1.7\)), although only marginally so, \(t(13) = -2.06, p = .06, d = -1.14\). Both groups discriminated between offside and no offside above chance level, because the \(d'\) values were significantly different from zero in both cases, SE: \(t(6) = 9.14, p < .001, d = 7.47\); and LE: \(t(7) = 9.98, p < .001, d = 7.55\). The response bias differed significantly between the two groups, \(t(13) = -3.12, p = .008, d = -1.73\). The response bias for the SE group was positive (\(c = 0.08\)) but not different from zero, \(t(6) = 0.44, p = .68, d = 0.36\), indicating no clear preference in case of doubt. The \(c\) value was negative (\(c = -0.53\)) and differed from zero for the LE group, \(t(7) = -6.51, p < .001, d = -4.92\), indicating a strong bias toward flagging in case of doubt, in line with the flash-lag effect.

Accuracy of Memory

The overall accuracy of memory in the frame recognition did not differ between the SE (mean 8.43 ± 2.76) and LE group (mean 8.38 ± 2.72), \(t(13) = -0.04, p = .97, d = -0.02\). However, a more detailed look showed clear differences. First, the SE group (mean 0.89 ± 0.04) was clearly more consistent than the LE group (mean 0.80 ± 0.06), \(t(13) = -3.31, p = .006, d = -1.84\). Second, the weighted mean was lower for the SE (mean 0.02 ± 0.29) than for the LE group (mean 0.66 ± 0.20), \(t(13) = 5.00, p < .001, d = 2.77\). Only the weighted mean of the LE group differed from zero, \(t(7) = 9.46, p < .001, d = 7.15\). Third, the SE and LE group showed a different pattern for the seven probe images (see Figure 1). Whereas the SE group showed a more even distribution of errors in both directions (selecting frames before or after the target frame), the LE group showed a clear asymmetry, with forward displacements being chosen more frequently. This asymmetry is congruent with the flash-lag effect.

Point-of-Gaze Data Video Clips

Mauchly’s sphericity tests revealed violation of the sphericity assumption for repeated-measures ANOVA of the areas of interest for the mean number of visual fixations (\(\chi^2 = 13.46, \varepsilon = 0.33, p = .001\)), the mean fixation duration (\(\chi^2 = 14.25, \varepsilon = 0.30, p < .001\)) and the percentage viewing time (\(\chi^2 = 16.77, \varepsilon = 0.25, p < .001\)). Therefore, Greenhouse–Geisser (G-G) correction was applied. For the three eye
movement variables, there was a significant main effect for areas of interest—mean number of visual fixations: $F(2,26) = 74.33, p < .001, \eta^2_p = 0.85$, G-G $\varepsilon = 0.60$, adjusted $p < .001$; mean fixation duration: $F(2,26) = 29.44, p < .001, \eta^2_p = 0.69$, G-G $\varepsilon = 0.59$, adjusted $p < .001$; and percentage viewing time: $F(2,26) = 31.44, p < .001, \eta^2_p = 0.71$, G-G $\varepsilon = 0.57$, adjusted $p < .001$. These results showed that the assistant referees fixated significantly more and longer on the offside line than on the passer (see Table 1).

There was no main effect for group—mean number of visual fixations: $F(1,13) = 0.12, p = .73, \eta^2_p = 0.009$; mean fixation duration: $F(1,13) = 1.39, p = .26, \eta^2_p = 0.10$; and percentage viewing time: $F(1,13) = 1.36, p = .27, \eta^2_p = 0.09$), and no interaction effect for Group × Areas of Interest—mean number of visual fixations: $F(2,26) = 0.09, p = .92, \eta^2_p = 0.007$, adjusted $p = .81$; mean fixation duration: $F(2,26) = 0.27, p = .77, \eta^2_p = 0.02$, adjusted $p = .65$; and percentage viewing time: $F(2,26) = 0.36, p = .70, \eta^2_p = 0.03$, adjusted $p = .58$). The mean number of visual fixations, mean fixation duration, and percentage viewing time were equal for both the SE and LE group (see Table 1).

The SE (mean 1.30 ± 0.18) and LE (mean 1.28 ± 0.07) group fixated on the same number of areas, $t(13) = -0.32, p = .76, d = -0.18$.

**Discussion**

The first aim of this experiment was to measure and compare offside decision-making accuracy within a group of elite-standard assistant referees. Second, the memory test could reveal more directly how assistant referees had perceived and remembered the players’ positions at the moment of the pass.
Based on response accuracy, the median-split technique was used to split the assistant referees in two expertise levels. A detailed analysis of the offside judgments showed that both groups were able to discriminate between onside and offside. The SE group showed no bias toward one type of error. The LE group, on the other hand, showed a clear bias toward flag errors that is in line with the predictions of the flash-lag hypothesis.

Although there was no overall group difference in accuracy of memory for the players’ positions, a closer look showed that the SE group selected frames that were more consistent with their offside judgment than the LE group. In addition, the LE group had a positive weighted mean indicating a clear forward shift in their memory of the spatial position of the players in line with the bias toward flag errors. In case of an erroneous judgment, the LE group selected rather a probe frame ahead (+1, +2 or +3) of the actual situation than behind (–1, –2 or –3). This asymmetry was not found for the SE group.

The combination of a bias toward flag errors and a forward memory shift demonstrated the impact of the flash-lag effect for the less-successful assistant referees. As suggested by Catteeuw et al. (2009a), successful assistant referees have probably learned to better deal with this perceptual illusion. They corrected their visual perception to make an accurate decision.

In line with earlier research (Catteeuw et al., 2009a), the point-of-gaze data showed no differences between both groups for number of visual fixations, fixation duration and percentage viewing time for the video clips. Both groups also fixated on the same number of areas. In other words, both groups look at the same things

### Table 1 Number of Visual Fixations, Fixation Duration, Percentage Viewing Time, and Number of Areas Fixated for the Video Clips of the Offside Situations (Mean ± SD)

<table>
<thead>
<tr>
<th></th>
<th>SE Group (n = 7)</th>
<th>LE Group (n = 8)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of Visual Fixations</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not on area of interest</td>
<td>4.62 ± 1.81</td>
<td>4.41 ± 0.81</td>
</tr>
<tr>
<td>Passer</td>
<td>1.01 ± 0.46</td>
<td>0.89 ± 0.37</td>
</tr>
<tr>
<td>Offside line</td>
<td>1.91 ± 0.39</td>
<td>1.95 ± 0.61</td>
</tr>
<tr>
<td><strong>Fixation Duration</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not on area of interest</td>
<td>1.98 ± 0.56</td>
<td>2.03 ± 0.57</td>
</tr>
<tr>
<td>Passer</td>
<td>0.46 ± 0.26</td>
<td>0.43 ± 0.21</td>
</tr>
<tr>
<td>Offside line</td>
<td>1.48 ± 0.68</td>
<td>1.24 ± 0.40</td>
</tr>
<tr>
<td><strong>Percentage Viewing Time</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not on area of interest</td>
<td>49.46 ± 13.77</td>
<td>53.71 ± 11.75</td>
</tr>
<tr>
<td>Passer</td>
<td>10.11 ± 5.08</td>
<td>10.61 ± 5.03</td>
</tr>
<tr>
<td>Offside line</td>
<td>40.43 ± 18.17</td>
<td>35.68 ± 12.18</td>
</tr>
<tr>
<td><strong>Number of Areas Fixated</strong></td>
<td>1.30 ± 0.18</td>
<td>1.28 ± 0.07</td>
</tr>
</tbody>
</table>

*Note. SE = successful experts; LE = less-successful experts.*
but they perceive different things (Abernethy, 1990, 1991). Our interpretation is that the skill-level differences within elite-standard assistant referees reflect a compensation strategy among the best performers. As indicated by the frame recognition, again, more successful assistant referees seem to have learned to better deal with the flash-lag effect.

**Experiment 2**

The offside decision-making task used in Experiment 1 has already demonstrated skill-level differences within elite-standard assistant referees (Catteeuw et al., 2009a; Gilis et al., 2008). In Experiment 1, the flash-lag effect had a considerable impact on response accuracy and accuracy of memory in less-successful assistant referees. Therefore, a training experiment was set up to improve offside decision-making accuracy of elite-standard assistant referees. A group of international assistant referees received four training sessions with feedback, which allowed them to become aware of the type of error they made (i.e., a forward memory shift due to the flash-lag effect). As in Experiment 1, the assistant referees judged offside simulations and chose the probe frame with the positions of the players at the moment of the pass. Subsequently, they received immediate feedback after each simulation consisting of the frame with the correct positions of the players. A control group of national assistant referees was only pre- and posttested.

After four training sessions, we anticipated higher response accuracy and accuracy of memory for the training group. We predicted no bias toward flag errors and no forward memory shift for the frame recognition. The control group was predicted to demonstrate a bias toward flag errors and a forward memory shift both in pre- and posttest.

In Experiment 1, no differences in visual scan patterns were found between two levels of elite-standard assistant referees. In Experiment 2, point-of-gaze data were only captured for the training group in pre- and posttest. In line with earlier results, we expected no changes in point-of-gaze data from pre- to posttest.

**Methods**

**Participants**

Participants were 24 Belgian elite assistant referees (mean age 39.98 ± 3.16). Ten assistant referees were on the FIFA list for 3.92 ± 2.15 years and formed the training group. The national assistant referees (n = 14) formed the control group.

Written consent was obtained from the Belgian referees’ committee. The nature of the study was explained to the assistant referees before testing. They were free to withdraw from testing at any stage. The study was designed and conducted in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki.

**Task**

The same offside decision-making task as described in Experiment 1 was used for the pre- and posttest. The assistant referees assessed 40 offside situations combined with frame recognition.
The training group received four training interventions in between the pre- and posttest. In each training intervention, the assistant referees participated in an offside decision-making task with immediate feedback, which consisted of 20 video simulations and 20 computer animations of offside situations. In total, the assistant referees assessed 160 offside situations (80 video simulations and 80 computer animations) in the training intervention. Each offside situation was followed with a picture of five probe frames of the potential moment of the pass. Subsequently, feedback was given consisting of the correct probe frame, containing the players’ positions at the exact moment of the pass. The computer animations were shown from a top-view perspective.

The video simulations of the offside situations were created in the same way as the situations of the offside decision-making task used in Experiment 1 and the pre- and posttest. These situations were recorded with Spanish youth elite players of a second division team in Madrid, Spain. The same editing procedures as in Experiment 1 were followed for the video clips and the creation of the probe frames. On average, the offside situations had a duration of 3.6 ± 0.9 s (range 1.8–6.0). The moment of the pass was on average 1.5 ± 0.4 s (range 0.8–3.3) after the start of the situation.

The production of high-quality offside video simulations is very difficult and time consuming. To increase the number of offside situations for each training intervention, computer animations were created which could be easily experimentally manipulated. They were designed with Macromedia Flash MX Professional 2004 version 7.2 (Macromedia, Inc., US) in the same way as in the study of Gilis et al. (2008). Offside situations with three attackers, two defenders, and one goalkeeper were created. The mean duration of the computer animations was 5.4 ± 0.5 s (range 4.3–5.6) and the moment of the pass was on average 2.7 ± 0.3 s (range 2.2–3.0) after the start of the clip. The same procedures were followed for the editing and the creation of probe frames as for the video simulations of the offside situations.

Procedure

Both groups executed the pretest in the first week. The assistant referees of the training group performed the test individually with a Tobii T120 Eye Tracker (Tobii Technology AB, Sweden). The assistant referees of the control group were tested in two groups of seven (same procedure as Gilis et al., 2008). The assistant referees were seated in front of a screen (2.60 × 3.60 m) within the edges of the screen to keep the viewing angle small (12°–17°). The test was projected on the screen with an LCD projector (EIKI Boardroom projector LCX 1100; EIKI International, Inc., US). Subsequently, the training group had one video training session per week from the second to the fifth week, in addition to their regular training routines. The control group continued their regular training routines. Finally, both groups performed the posttest in the sixth week. Again, the assistant referees of the training group were tested individually with the eye tracker and the assistant referees of the control group were tested in two groups. The latter were all seated on the same position in front of the screen as for the pretest.

Dependent Variables and Analysis

The dependent variables and data analyses are similar to the ones used in Experiment 1. Response accuracy of the video clips was calculated for both groups in
pre- and posttest. For a more profound analysis of the accuracy, error distribution (chi-squared test), response sensitivity ($d'$) and response bias ($c$) were analyzed as well. For the probe frames, accuracy of memory, proportions per frame, consistency, and weighted means were calculated.

Point-of-gaze data were only recorded for the training group. As in Experiment 1, the number of visual fixations, mean fixation duration, percentage viewing time, and number of areas fixated were calculated for the video clips in pre- and posttest.

To study differences in response accuracy of the video clips and probe frames between training and control group from pre- to posttest, repeated-measures ANOVAs were used with group as between-participants variable and test as within-participants variable. For the point-of-gaze data analysis of the training group, repeated-measures ANOVAs were used for number of visual fixations, mean fixation duration and percentage viewing time. Test was the between-participants variable and areas of interest was the within-participants variable. Mauchly’s test of sphericity was performed and, in case of violation, Greenhouse-Geisser correction was applied. Significant effects were examined using Fisher LSD post hoc procedures and effect sizes were reported as partial eta squared ($\eta_p^2$). The number of areas fixated was analyzed as a function of Test with a $t$ test for dependent samples. Effect sizes were calculated with Cohen’s $d$.

Results

Response Accuracy Video Clips

The repeated-measures ANOVA for response accuracy showed a Test × Group interaction effect, $F(1,22) = 6.02, p = .02, \eta_p^2 = 0.21$. The training group clearly improved from pretest (mean 28.5 ± 3.3) to posttest (mean 31.2 ± 3.4) ($p = .047$). The control group remained approximately on the same level from pretest (mean 26.4 ± 5.0) to posttest (mean 24.9 ± 2.8) ($p = .20$). Whereas no difference was apparent between both groups in the pretest ($p = .18$), the training group performed clearly better than the control group in the posttest ($p < .001$).

In Table 2, the error distribution, the response sensitivity and bias scores are shown. The chi-squared test showed no bias toward one type of error for the training group. The sensitivity index ($d'$) and response bias ($c$) were calculated using Fisher’s z-transformation.

### Table 2

<table>
<thead>
<tr>
<th>Test and Group</th>
<th>Flag Errors</th>
<th>Non-flag Errors</th>
<th>$d'$</th>
<th>$d' \neq 0$</th>
<th>$c$</th>
<th>$c \neq 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest</td>
<td>9.20 ± 4.52</td>
<td>2.30 ± 2.00</td>
<td>1.41</td>
<td>$p &lt; .001$</td>
<td>-0.13</td>
<td>$p = .47$</td>
</tr>
<tr>
<td>Posttest</td>
<td>7.30 ± 3.83</td>
<td>1.50 ± 1.58</td>
<td>1.85</td>
<td>$p &lt; .001$</td>
<td>-0.17</td>
<td>$p = .22$</td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest</td>
<td>10.57 ± 4.09</td>
<td>3.07 ± 2.13</td>
<td>1.01</td>
<td>$p &lt; .001$</td>
<td>-0.10</td>
<td>$p = .33$</td>
</tr>
<tr>
<td>Posttest</td>
<td>10.64 ± 3.13</td>
<td>4.43 ± 1.09</td>
<td>0.53</td>
<td>$p &lt; .001$</td>
<td>0.12</td>
<td>$p = .10$</td>
</tr>
</tbody>
</table>
(pretest: $\chi^2 = 1.53, p = .22$; posttest: $\chi^2 = 2.97, p = .08$) and control group (pretest: $\chi^2 = 0.63, p = .43$; posttest: $\chi^2 = 2.16, p = .14$). The sensitivity index differed from zero in pre- and posttest for both groups, indicating their ability to discriminate between offside and no offside above chance level. The response bias showed no differences between both groups, $F(1,22) = 3.66, p = .07, \eta^2_p = 0.14$, and no difference from zero indicating no preference toward flagging or not flagging in case of doubt.

### Accuracy of Memory

For accuracy of memory, the Test × Group interaction revealed no differences between the training (pretest: 8.60 ± 2.88; posttest: 10.30 ± 2.00) and control group (pretest: 8.50 ± 2.38; posttest: 9.14 ± 2.54), $F(1,22) = 0.50, p = .49, \eta^2_p = 0.02$. Consistency, $F(1,22) = 15.54, p < .001, \eta^2_p = 0.41$, however, improved for the training group from pretest (mean 0.83 ± 0.05) to posttest (mean 0.91 ± 0.05) ($p = .03$). The control group was less consistent than the training group (pretest: $p = .006$; posttest: $p < .001$) and deteriorated from pretest (0.73 ± 0.13) to posttest (0.63 ± 0.08). The Test × Group interaction for weighted mean, $F(1,22) = 3.48, p = .08, \eta^2_p = 0.14$, showed no differences within the training (pretest: 0.33 ± 0.47; posttest: 0.26 ± 0.40) and control group (pretest: 0.21 ± 0.40; posttest: $-0.11 \pm 0.18$) from pre- to posttest.

### Point-of-Gaze Data Video Clips

Point-of-gaze data were only investigated for the training group from pre- to post-test. Mauchly’s sphericity tests revealed violation of the sphericity assumption for repeated-measures ANOVA of the areas of interest for the mean number of visual fixations ($\chi^2 = 12.91, \epsilon = 0.42, p = .002$), the mean fixation duration ($\chi^2 = 13.16, \epsilon = 0.42, p = .001$) and the percentage viewing time ($\chi^2 = 18.10, \epsilon = 0.30, p < .001$). Therefore, Greenhouse–Geisser correction was applied.

Assistant referees fixated more and longer on the offside line than on the passer (mean number of visual fixations: $F(2,32) = 39.67, p < .001, \eta^2_p = 0.71$, G-G $\epsilon = 0.63$, adjusted $p < .001$; mean fixation duration: $F(2,32) = 26.62, p < .001, \eta^2_p = 0.62$, G-G $\epsilon = 0.63$, adjusted $p < .001$; and percentage viewing time: $F(2,32) = 29.94, p < .001, \eta^2_p = 0.65$, G-G $\epsilon = 0.59$, adjusted $p < .001$) (see Table 3). The mean number of visual fixations, $F(1,16) = 1.84, p = .19, \eta^2_p = 0.10$; mean fixation duration, $F(1,16) = 2.02, p = .17, \eta^2_p = 0.11$; and percentage viewing time, $F(1,16) = 0.10, p = .76, \eta^2_p = 0.01$, were equal from pre- to posttest (see Table 3). From pretest (mean 1.28 ± 0.15) to posttest (mean 1.21 ± 0.20), the training group fixated on the same number of areas, $t(8) = 0.83, p = .43, d = 0.59$.

### Discussion

The training interventions had a positive effect on response accuracy because the training group improved in the posttest, while the control group did not. A more detailed look demonstrated no bias toward a type of error for both groups. They both could discriminate between onside and offside in pre- and posttest and showed no preference to flag or not to flag in case of doubt.
Analysis of accuracy of memory did not concur with anticipated differences. Both groups performed slightly better in the posttest, but we could not speak of a training effect. In addition, the weighted mean was slightly positive for both groups, except in the posttest for the control group. Interestingly, consistency between response accuracy and accuracy of memory was higher for the training group and improved from pre- to posttest. For the control group, however, consistency deteriorated. Although accuracy of memory was not different for both groups, results of the consistency pointed at a higher or lower degree of certainty for the training and control group, respectively.

In the video clips, there were no changes in the point-of-gaze data for the training group from pre- to posttest. In line with Catteeuw et al. (2009a), assistant referees fixated primarily the offside line in the video clips. Again, the lack of expertise differences supports our hypothesis of an optimization of the visual scan pattern (perceptual mechanism). At this standard of refereeing, a higher response accuracy can most probably be attributed to a better developed compensation strategy for the forward memory shift induced by the flash-lag effect (decision mechanism).

### General Discussion

In Experiment 1, the median-split technique was used to create a successful and less-successful expert group of elite-standard assistant referees in an offside decision-making task. The successful expert group was less susceptible to the flash-lag effect. The less-successful expert group, in contrast, demonstrated a bias toward flag errors and a forward memory shift in the frame recognition, two markers of the flash-lag effect. These findings could not be reproduced in Experiment 2. While groups were
based on performance in an offside decision-making task in Experiment 1, here, groups were determined by the ranking of the assistant referees in their national association. The ranking of assistant referees are based on match reports, which comprise marks for application of the FIFA Laws of the game (FIFA, 2009) and offside decision making in particular, physical shape, movement technique and flag handling, and team communication. The best ten assistant referees are nominated as international assistant referees. The international assistant referees train together on a weekly basis and, hence, were suitable to form the training group. The elite national assistant referees train together twice a month and they formed the control group. In our opinion, a classification according to the response accuracy in this off-field offside decision-making task can have an added value to match reports, which are dependent on the number of situations and the coincidental difficulty of the situations.

Although we could not find indications of the flash-lag effect in Experiment 2, the progress of the training group suggested the meaningfulness of the training program with immediate feedback. However, the positive results of the training intervention should be cautiously approached. As testing and training was almost identical, positive results risk being the effect of test familiarity. In contrast to an improvement in response accuracy, accuracy of memory remained at the same level, approximately 20–25%. In Experiment 1, accuracy of memory did also not differentiate. The frame recognition task seemed to be very difficult, leading us to believe that assistant referees need even greater amounts of training possibilities. However, a detailed analysis revealed some differences. Less-successful expert assistant referees made a forward memory shift in line with the flash-lag hypothesis (Baldo et al., 2002; Catteeuw et al., 2009a; Gilis et al., 2008, 2009; Helsen et al., 2006).

Although skill-level differences in response accuracy were recognized within a group of elite-standard assistant referees, no differences were apparent in visual scan patterns between two skill-levels (Experiment 1) (cf. Catteeuw et al., 2009a), and before and after a training intervention (Experiment 2). These data suggest that the perceptual mechanism to detect and select the appropriate visual information is used in an optimal way when the assistant referees are in line with the second-last defender. Although all assistant referees scanned the same sources of information, superior information pick-up and response selection led to more correct decisions. The most successful among the elite-standard assistant referees distinguished themselves through superior decision making (Catteeuw et al., 2009a).

The training interventions consisted of video simulations and computer animations of offside situations. These computer animations were implemented to increase the number of training situations in each intervention. Although the top-view perspective deviates from the sideline perspective in real matches, Gilis et al. (2008, 2009) demonstrated the validity and reliability of these computer animations in offside decision-making. Also similar studies in field hockey (Williams, Ward, & Chapman, 2003) and driving (Devos et al., 2009) showed the benefits of video- or computer-based perceptual training. Still, the effects of video simulations versus computer animations cannot be discriminated in this study. Given that a training intervention with video simulations and computer animations showed its positive added value, future research can concentrate on finding which one is most effective. If a training intervention with only computer animations is as effective as a training intervention with only video simulations, it will be much easier and faster to create larger computer-animation training sets with continually more challenging situations.
An important skeptical remark is the lack of a transfer test. Without a transfer test, it is difficult to generalize the positive effects of this training study to real-life on-field performance. Future research should implement a transfer test, for example on-field offside decision making. In this type of test, researchers can videotape and evaluate assistant referees while judging offside situations simulated by players on the field of play.

In summary, skill-level differences in offside decision making can be detected within elite-standard assistant referees. Only the most successful assistants learned to overcome the flash-lag effect. Bearing in mind the lack of transfer testing, a combination of video- and computer-based training with immediate feedback is efficient to fine-tune the decision-making process and to learn to correct for perceptual illusions like the flash-lag effect.

References


*Manuscript received: January 15, 2010*

*Revision accepted: August 23, 2010*