Judging the Judges' Performance in Rhythmic Gymnastics

KONSTANTINOS FLESSAS[1], DIMITRIS MYLONAS[2], GEORGIA PANAGIOTAROPOULOU[3], DESPINA TSOPANI[4], ALEXANDREA KORDA[3], CONSTANTINOS SIETTOS[2], ALESSANDRA DI CAGNO[5], IOANNIS EVDOKIMIDIS[6], and NIKOLAOS SMYRNIS[6,7,8]

[1]Sports Medicine Department, Faculty of Physical Education and Sports Science, National and Kapodistrian University of Athens, Athens, GREECE; [2]School of Applied Mathematics and Physical Sciences, National Technical University of Athens, Athens, GREECE; [3]School of Electrical and Computer Engineering, National Technical University of Athens, Athens, GREECE; [4]Gymnastics Department, Faculty of Physical Education and Sports Science, National and Kapodistrian University of Athens, Athens, GREECE; [5]Department of Health Sciences, "Foro Italico" University of Rome, Rome, ITALY; [6]Neurology Department, National and Kapodistrian University of Athens, Aeginition Hospital, Athens, GREECE; [7]Laboratory of Sensorimotor Control, University Mental Health Research Institute, Athens, GREECE; and [8]Psychiatry Department, National and Kapodistrian University of Athens, Aeginition Hospital, Athens, GREECE

ABSTRACT

FLESSAS, K., D. MYLONAS, G. PANAGIOTAROPOULOU, D. TSOPANI, A. KORDA, C. SIETTOS, A. DI CAGNO, I. EVDOKIMIDIS, and N. SMYRNIS. Judging the Judges' Performance in Rhythmic Gymnastics. *Med. Sci. Sports Exerc.*, Vol. 47, No. 3, pp. 640–648, 2015. **Introduction**: Rhythmic gymnastics (RG) is an aesthetic event balancing between art and sport that also has a performance rating system (Code of Points) given by the International Gymnastics Federation. It is one of the sports in which competition results greatly depend on the judges' evaluation. In the current study, we explored the judges' performance in a five-gymnast ensemble routine. **Methods**: An expert–novice paradigm (10 international-level, 10 national-level, and 10 novice-level judges) was implemented under a fully simulated procedure of judgment in a five-gymnast ensemble routine of RG using two videos of routines performed by the Greek national team of RG. Simultaneous recordings of two-dimensional eye movements were taken during the judgment procedure to assess the percentage of time spent by each judge viewing the videos and fixation performance of each judge when an error in gymnast performance had occurred. **Results**: All judge level groups had very modest performance of error recognition on gymnasts' routines, and the best international judges reported approximately 40% of true errors. Novice judges spent significantly more time viewing the videos compared with national and international judges and spent significantly more time fixating detected errors than the other two groups. National judges were the only group that made efficient use of fixation to detect errors. **Conclusions**: The fact that international-level judges outperformed both other groups, while not relying on visual fixation to detect errors, suggests that these experienced judges probably make use of other cognitive strategies, increasing their overall error detection efficiency, which was, however, still far below optimum. **Key Words:** EXPERT–NOVICE JUDGES, FIXATION, DECISION MAKING, EYE TRACKING

Rhythmic gymnastics (RG) is an aesthetic event balancing between art and sport that also has a performance rating system (Code of Points) given by the International Gymnastics Federation (FIG), the official governing body for gymnastics in the world. RG is one of the sports (e.g., artistic gymnastics, diving, figure skating, and synchronized swimming) in which competition results (scoring and ranking of gymnasts' performance) greatly depend on the judges' evaluation. A third of all sports recognized by the International Olympic Committee have a performance rating system in which judging plays a significant role (27).

In RG at the world-class level, extremely difficult bodily movements are performed in combination with skillful handling of the apparatus, resulting in a fascinating spectacle. The particular task that a judge of RG is faced with when officiating a five-gymnast ensemble is to attend simultaneously to all gymnasts, following their performance in real time and detecting errors in performance on the basis of specific scoring rules. A basic mechanism for focusing attention on a particular object of interest is to direct foveal vision to it. Using foveal vision in this case would result in attending to only one athlete at a time, losing execution information from the other athletes. It is known that humans can simultaneously track more than one moving object

Address for correspondence: Despina Tsopani, Gymnastics Department, Faculty of Physical Education and Sports Science, National and Kapodistrian University of Athens, Ethnikis Antistaseos 41, 17237 Dafne, Greece; E-mail: tsopanid@hol.gr.
Submitted for publication January 2014.
Accepted for publication June 2014.
Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.acsm-msse.org).

0195-9131/15/4703-0640/0
MEDICINE & SCIENCE IN SPORTS & EXERCISE®
Copyright © 2014 by the American College of Sports Medicine
DOI: 10.1249/MSS.0000000000000425

without visually fixating at all objects simultaneously. This capacity reaches exceptional levels in highly trained professional athletes (10). Pylyshyn and Storm (23) formally demonstrated that humans can track up to 4–5 identical items that move for a few seconds in a random fashion along with identical distractor items in a two-dimensional computer display (see also Cavanagh and Alvarez (10)). In a more recent study, Oksama and Hyönä (21) confirmed that human observers can track up to four different objects that move independently and fewer objects in the presence of distractors, although this capacity limit varies considerably from subject to subject. In a subsequent study, Oksama and Hyönä (22) showed that this capacity for tracking multiple objects depends on several stimulus characteristics. An increase in the number of features of each object to be attended to results in dramatic decrease of capacity. An increase in the speed of the moving objects also results in such decrease. Oksama and Hyönä (22) also showed that the tracking of familiar objects in comparison with novel objects resulted in increase in capacity. The density of targets and the hemifield of view impose additional limitations to human capacity for attending to multiple moving objects. When targets and distractors are too close, it becomes difficult to individuate the targets and maintain tracking (15). In addition, it has been shown that humans can simultaneously track two separate items per hemifield at most (10).

Judges of RG have to track the performance of five gymnasts simultaneously for 2 min 15 s to 2 min 30 s. These gymnasts move within a large competitive field of size 13 × 13 m, coming close to each other and then spreading out running or jumping at high speed. Each of them also holds an apparatus (rope, ribbon, and hoop) that at certain time points becomes separated from the gymnast after a complex moving trajectory of its own. Thus, there are certain moments when the judge might have to trace up to 10 individual objects (five gymnasts plus five apparatuses) and detect several possible errors in body shape, position and coordination, apparatus trajectory and contact with the gymnast, as well as combinatorial errors in the coordination of all gymnasts and apparatuses. This task also has to be performed in the presence of a variety of distractors during an international competition event. Experience and training of judges certainly play an important role in their ability to perform this task. Nevertheless, they are required to perform a formidable cognitive task that lies beyond the normal limits of human attentional capacity, even taking into consideration the fact that attention can be maintained at objects outside the focus of visual fixation.

Besides capacity limitations of the attentional system, other factors are thought to affect judgment of RG judges. One such factor is *a priori* bias. Bias in sports officiating is a known phenomenon in competitive sports, examples being social biases, crowd noise, etc. (6,8,11,20,28,30,31). The effects of biased officiating are potentially more dramatic in sports, like RG, in which the officials actually determine the outcome of the competition with a score of points resulting from the evaluation of the performances of gymnasts.

In the current study, we explored the capacity in judging performance in a five-gymnast ensemble routine of international-level, national-level, and novice-level RG judges using an expert–novice paradigm. In previous studies, the professional versus amateur classification of the individuals was based on the level of expertise measured by the league status (national, local) or the status of the games (Olympic, national) (1–4, 24,25,32). Especially, the studies that investigated the difference between expert and novice sports observers at visual behavior (e.g., coaches, judges, and teachers) have used classifications based on qualification rather than number of years of experience (e.g., 7,19). In this study, we used official qualification for the classification of the judges into international, national, and novice groups while our grouping also followed the level of expertise based on league status. The study design used a fully simulated procedure of judgment (using videos of routines performed by the Greek national team of RG). In this design, the bias and other context-dependent factors influencing judgment were minimized and the performance of judges was compared with perfect performance on the basis of a frame-by-frame examination of the routine videos by five of the international-level judges.

In the current study, we also used the recordings of two-dimensional eye movements of all judges performing the judgment task to gain more insight into their performance. Because eye movements provide insight into problem solving, reasoning, mental imagery, and search strategies (16), several researchers have used objective methods to investigate how eye movements are related to cognitive processes during visual tasks. The eye tracking technology provides objective and quantitative evidence of the user's visual perception and (overt) attention processes (12). In relation to sport science literature, eye tracking recorders have been used widely to measure visual fixations. Usually, a fixation refers to a period of 100 ms or longer when an individual is focusing on one location of the visual scene (33).

Whereas athletes' perceptual–cognitive skills have been frequently examined using eye movement recordings, the skills and expertise of sport officials did not receive as much attention. Bard et al. (7) measured the visual search patterns of gymnastics judges and noted that experts had fewer fixations of longer duration and detected more errors than novices, although neither result was statistically significant. Moreno et al. (19) studied gymnastic coaches with different levels of expertise, who evaluated gymnastics routines, and showed that expert participants showed longer and fewer visual fixations than the novice group. This observation was not supported by other studies on sport officials. Catteeuw et al. (9) studied international and national assistant soccer referees, showing that international assistant referees made more accurate decisions than national assistant referees but the groups did not differ on visual search patterns. Recently, Hancock and Ste-Marie (14) studied the gaze behavior of higher-level and lower-level ice hockey referees. They observed that the higher-level ice hockey referees were superior to lower-level referees in decision making but they did not differ in gaze behaviors. In this case, again, the total number of fixations was not a sensitive measure for detecting differences in decision making that are related to the level of expertise of judges.

In this study, we used eye movement measurements to study fixation performance of international-level, national-level, and novice-level RG judges when an error in gymnast performance had occurred. The basic question we addressed was whether judges used visual fixation on errors efficiently to aid them in their decision making process.

METHODS

Participants. The study population included 30 women judges of Greek nationality. All participants were recruited by invitation from the Hellenic Gymnastics Federation. The first group of 10 novice judges was recruited from a pool of 14 judges that had obtained an official license from the Hellenic Gymnastics Federation at the time of the testing but had not yet participated as judges in an official event (mean age, 24.60 yr; SD, 5.64 yr). Another 10 participants (mean age, 29.50 yr; SD, 12.5 yr) were Greek national judges for RG accredited by the Hellenic Gymnastics Federation. The national judges were recruited from a pool of 20 judges that had obtained an official license from the Hellenic Gymnastics Federation and had already participated in official events, as judges at the time of testing, organized by the Hellenic Gymnastics Federation. Finally, the third group of 10 participants was international judges for RG (mean age, 45.50 yr; SD, 8.10 yr) accredited by the FIG. The international judges were recruited from a pool of 20 national judges that had obtained an international license from the FIG at the time of testing and had already participated in international official events as judges. All participants gave an informed consent for their participation in this study, and the study protocol was approved by the ethics committee of the Aeginition University Hospital.

Procedure. For the purposes of the study, we received permission from the Hellenic Gymnastics Federation to videotape two performances of the Greek national team in RG during a special evaluation event, in the presence of four international judges, at the National Gymnastics Training Centre. The first routine that lasted 2 min 29 s was a performance of the five-gymnast ensemble using hoops (H). The second routine that lasted 2 min 30 s was a performance of the same five-gymnast ensemble using ribbons and ropes (RR). Both routines were videotaped using a video camera (AG-HPX170P; Panasonic, with recording rate of 25 frames per second) placed at a specific stable position to simulate, approximately, the visual scene of the judge seated at the left side of the judges' panel. In the determination of this position, the distance of the judges' panel from the competitive field and the average height of the judges' eye position were taken into account, as standardized for all World Championships and Olympic Games by the FIG.

The participants proceeded to the laboratory, where after a short interview, they were seated and their head was restrained using a chin rest. Movements of the right eye of each participant were recorded using the ISCAN ETL-200 camera (sampling rate, 240 Hz). Visual stimuli were rear-projected using a projector (TDP-T40; Toshiba). The resolution of the projection area was 1024 × 768 pixels, with respective dimensions of 39.4 cm (width) and 30 cm (height). The display dimensions were specified to correspond to 20° of maximum horizontal deviation and 15° of maximum vertical deviation from the center of the projection. The precision of eye movement measurements was at 0.07° and 0.12° at the horizontal and vertical axes, respectively. Subsequently, each participant performed a calibration procedure using a nine-point grid and a moving white circular dot (the initial scene of each video clip was used as background). The calibration procedure was repeated before and after each video to reevaluate the measurements' precision. Then, the two videotaped RG routines were projected, with a resting period of a few minutes in between. There was no sound attached to the videos. Before the initiation of each video projection, an official scoring sheet was placed on a lectern in front of the participant and the participant was instructed to evaluate the gymnasts' routine to simulate the evaluation procedure that takes place during official competition events. This procedure was performed according to the FIG Code of Points (version used, 2008–2012). The evaluation system of the FIG consists of three elements, as follows: difficulty, which is further divided into body difficulty (D1) and apparatus difficulty (D2), execution, and artistic value. Each participant was instructed to evaluate the same two elements for both routines (either difficulty and execution or difficulty and artistic value) one at a time. The difficulty element was either D1 or D2.

Behavioral data analysis. In this analysis, we used only the data for difficulty from each participant, namely, the scores on D1 or the scores on D2. The total value of the D1 scoring of the H routine was 10 points and was the sum of scores of 14 separate exercises. Similarly, the total value of the D2 of the same routine was 10 points and was the sum of 39 separate exercises. Each subject assigned a penalty score to each exercise, corresponding to a particular error in performance (zero indicating no penalty). A reevaluation of the D1 and D2 scoring of the videos was performed by each of the five most experienced international judges 6–8 months after the first testing. Each judge sat with two of the researches, one of the latter being an international RG judge (D. T.) and watched each video exhaustively (frame by frame, slow motion, and repeated replay). Each judge was interviewed by the two researches, and her comments were recorded for each judgment she made. The data from the interviews were all pooled together and evaluated by the two researches. In cases of complete agreement among the five judges, which constitute the majority of the cases, and the independent expert judge, a penalty score was assigned as the "true penalty score." In the few cases where no consensus was reached among the judges, the true penalty score was assigned by the independent researcher–judge (D. T.). The absolute difference between the true penalty score (for each exercise, category of judgment, and routine) and the corresponding penalty score from each participant was calculated. This measure was then subjected to a factorial ANOVA, with judge level (three levels), category of judgment (D1 or D2), and program (H or RR) as categorical factors. This analysis was repeated using the value of each exercise as covariate (ANCOVA). Finally, the mean difference score for each level of judges was compared with zero (perfect performance) by means of a one-sample *t*-test.

A second analysis was performed by pooling the judgment scores for all exercises for both routines for all participant judges of each level. A categorical classification (0 or 1) of scores was then made to the following:

1) hits, a true error was also scored as an error by the participant;
2) false alarms, an error was scored where there was no true error;
3) correct rejections, no error was scored and there was no true error; and
4) misses, a true error occurred but was not scored as an error by the participant.

Using these parameters, we defined the percentage of hits, as equation 1:

$$\mathrm{P\ (h)} = \frac{\text{number of hits}}{\text{number of hits} + \text{number of misses}} \qquad [1]$$

We also defined the percentage of false alarms, as shown in equation 2:

$$\mathrm{P\ (fa)} = \frac{\text{number of false alarms}}{\text{number of false alarms} + \text{number of correct rejections}} \qquad [2]$$

Then, we defined a signal/noise ratio for each level of participant judges, as shown in equation 3:

$$\text{d-prime} = \mathrm{ZP\ (h)} - \mathrm{ZP\ (fa)} \qquad [3]$$

The d-prime score is a measure of detection sensitivity that is free of response bias (17). This analysis used a pooled d-prime score for all participants of each judge category instead of individual d-prime scores for each participant as well as the 95% confidence interval (CI) of the pooled d-prime (see document, Supplemental Digital Content 1, Supplementary methods, http://links.lww.com/MSS/A417). The reason was that we had a small number of scores for each participant judge (14 + 39 = 53 or 14 + 38 = 52) and most of them were correct rejections, meaning errorless judgments. The calculation of d-prime with only a few trials per subject is prone to errors because of large random variation (17). Pooling data from many subjects gives a more reliable estimate of the true d-prime of the population. The d-prime of the pooled data is a better estimate than the mean of the d-primes for each subject (17).

Eye position data recording and preprocessing. The *x*, *y* position data for the right eye were recorded during each evaluation. These data were first calibrated offline and were transformed to degrees of visual angle (vertical and horizontal components). Then, the eye movement recording data were processed with a software tool that was developed by our group to define blinks and saccadic eye movements larger than 0.5° (see document, Supplemental Digital Content 1, Supplementary methods, http://links.lww.com/MSS/A417).

After excluding blinks and other artifacts as well as saccadic eye movements greater than 0.5°, the resulting eye position data that contained fixations and smooth eye pursuit were assigned to each frame of the video that was projected to each participant. Thus, for every 4-ms frame duration (60 frames per second or 1 frame per 4 ms), the eye position data were superimposed on the frame and the original eye position records were reduced to one point (horizontal and vertical coordinates in degrees of visual angle) per frame.

Eye position data analysis. To define the errors in the two video routines spatially and temporally, we developed an interactive software tool in MATLAB (MathWorks version 2010b). A grid of 41 (horizontal) × 31 (vertical) rectangular wire boxes was superimposed to each video frame. The wire boxes were designed so that their sides would correspond to a 4° eye movement that represents the visual angle covered by the fovea. Two of the researchers viewed each video (frame by frame) and located the true penalty errors that have been previously reported by the expert judges by marking the rectangular boxes that covered the corresponding areas where errors occurred. This resulted in the description of errors as areas defined by one or more rectangular wire boxes (see Figure, Supplemental Digital Content 2, A typical frame of our research where an error occurs, http://links.lww.com/MSS/A418). This marking was performed for every frame of the two videos for D1 and D2 errors separately. According to this procedure, a subgroup of errors that were well defined in space and time was isolated for further processing, excluding errors that were widely distributed in space. This process resulted in the identification of two D1 errors and three D2 errors for the exercises of the H routine plus five D1 errors and four D2 errors for the exercises of the RR routine. The scores for the selected errors were separately analyzed (14 exercises × 15 participants = 210 scores) behaviorally to test for differences in performance among the three levels of participants using hit rate and absolute difference in score as defined previously. We did not test for the effects of category (D1 and D2) or routine (H and RR) for these data, and all scores were pooled together for each level of participants.

Subsequently, an automated procedure was used to examine whether each error was spotted by each of the subjects, namely, whether the subject's eye position was within the limits of the error area for a certain amount of time. To do that, we modeled the eye position of each subject as a 2° radius circle centered at the coordinates recorded by the ISCAN and then checked whether there was an overlap with at least one of the wire boxes that defined the area marked as error. The program calculated the maximum number of continuous frames where such overlap occurred. For every error, a categorical variable determined whether the eyes overlaid the area of the error for at least 100 ms. A second continuous variable was defined as the maximum continuous time interval (ms) that the eyes were within the error area (this variable was set to zero if the eyes did not pass over the error area at all). These data were then subjected to a *t*-test with separate variance estimates, comparing the mean time interval for errors that were reported by the subject (score, >0) and errors that were not reported by the subject (score, 0). This test was performed separately for each level of participants.

Using the categorical eye position variable described previously, we defined the following parameters separately for each level of participants:

1) hits, cases where the eyes captured the error (categorical eye variable, 1) and the subjects reported the error (subjective score, 1);
2) false alarms, cases where the eyes did not capture the error (categorical eye variable, 0) and the subjects reported the error (subjective score, 1);
3) correct rejections, cases where the eyes did not capture the error (categorical eye variable, 0) and the subjects did not report the error (subjective score, 0); and
4) misses, cases where the eyes captured the error (categorical eye variable, 1) and the subjects did not reported the error (subjective score, 0).

Using these parameters, we defined a d-prime (eye) signal/noise ratio for each level of participants using equation 3. This d-prime measure indicates how much the subject relies on eye position to make a decision on whether an error in performance occurred. If the subject relies heavily on eye position, then hits (that are the cases where the eyes foveated the error and the subject reported the error) should be much larger compared with false alarms (that are the cases where the subject reported the error but the eyes did not foveate it). The 95% CI for each d-prime (eye) score for each level of participants was calculated (see document, Supplemental Digital Content 1, Supplementary methods, http://links.lww.com/MSS/A417).

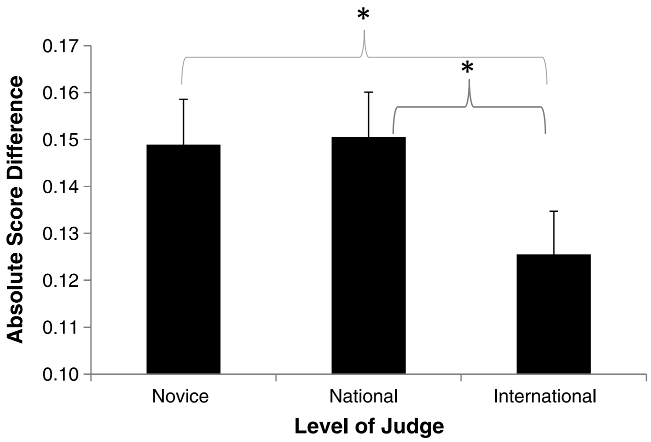

FIGURE 1–Means of absolute score difference between the judges' score and the true score for all three levels of judges (*bars* denote SEM, and significant differences between groups are marked with an *asterisk* (*)).

RESULTS

Behavior. The ANOVA on the absolute difference between judges' score and the true score confirmed a significant effect of the level of judge ($F_{2,1562} = 4.6$, $P < 0.01$). Figure 1 shows that the absolute score difference was smaller for the international-level participants than that for the national and novice participants. In addition, there was no difference between the latter two groups. The absolute score difference was also significantly different from the optimum zero value for all levels of participant judges (one sample $t_{524} = 13.7$ and $P < 10^{-6}$ for international judges; $t_{524} = 15.6$ and $P < 10^{-6}$ for national judges; $t_{524} = 14.4$ and $P < 10^{-6}$ for novice judges). The ANOVA also showed a significant effect of the category of judgment ($F_{1,1562} = 394.7$, $P < 10^{-6}$), significant effect of routine ($F_{1,1562} = 33.2$, $P < 10^{-6}$), and significant interaction of routine versus category of judgment ($F_{1,1562} = 11.4$, $P < 10^{-3}$). Figure 2 shows that the performance of the judges in D1 evaluation was worse (larger absolute differences) than the performance in D2 evaluation. Moreover, the performance of the judges in the RR routine evaluation was worse than the performance in H routine evaluation. The D1-versus-D2 difference was more prominent for the RR routine than that for the H routine. Most importantly, there was no interaction of the level of the judge and the judgment category ($F_{2,1562} = 2.13$, $P < 0.12$) as well as of the level of the judge and the routine ($F_{2,1562} = 1.04$, $P < 0.35$). Finally, there was no significant three-way interaction, namely, level of the judge × the judgment category × routine ($F_{2,1562} = 1.03$, $P < 0.35$). The analysis was repeated (ANCOVA) using the value of each exercise as covariate, and the results for the level of

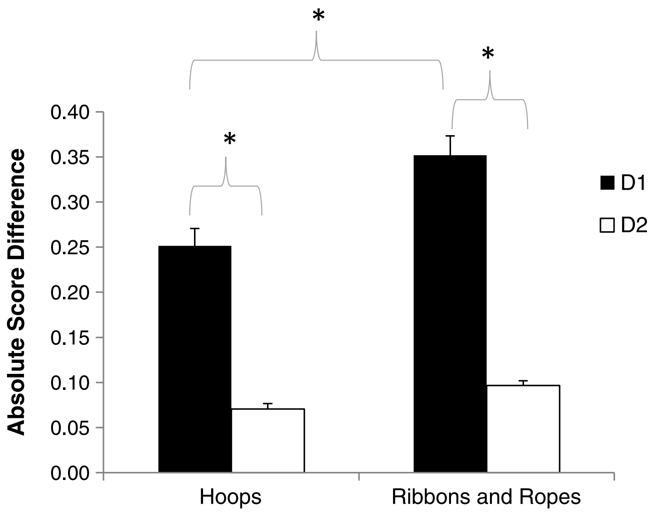

FIGURE 2–Means of absolute score difference between judge score and the true score for both categories of judgment and both routines (*bars* denote SEM, and significant differences between groups are marked with an *asterisk* (*)).

TABLE 1. Categorical classification of judge performance.

	Novice	National	International
Hits	59	62	105
False alarms	29	43	37
Correct rejections	255	254	255
Misses	211	208	165
Hit rate (%)	21.85	22.96	38.89
False alarm rate (%)	10.21	14.48	12.67
d-Prime	0.49*	0.32*	0.86*
Low to high 95% CI	0.43–0.55	0.25–0.38	0.77–0.95

The 95% CI for all d-primes did not overlap, indicating that all pairwise d-prime differences were significant.

*All d-prime scores are significantly different from zero (no 95% CI includes zero).

judge, category of judgment, and type of routine were the same as those for the ANOVA. The value of exercise had a significant effect as a cofactor on the absolute score ($F_{1,1562}$ = 102.6, $P < 10^{-6}$). The mentioned analysis confirmed that differences in judgment performance due to the category of judgment, program type, and the particular value of each exercise were the same for all levels of judges. In the analysis that follows, we pooled together the individual judgments for all routines and categories of judgment for each judge.

Table 1 presents the categorical classification for the scores of the participants and the d-prime score for each one of the three levels of judges. The first important observation to be made was that hit rate was much lower than the optimal 100%. The best group of international judges performed at a mediocre level of approximately 40%; thus, they reported less than half of the true errors. The d-prime scores were all significantly higher than zero, as indicated by their 95% CI values in Table 1, confirming that all three judge groups performed better than chance. The second important observation to be made from Table 1 was that international judges were significantly better than national and novice judges in their overall performance. Interestingly, the national judges were significantly worse in their overall performance than novice judges, and this was due to the larger proportion of false alarms for this group compared with that for the group of novice judges.

Eye position and performance. The ANOVA for the percentage of time that the eyes were not viewing the screen showed that there was no effect of routine ($F_{1,48}$ = 0.01, P = 0.9) and no effect of category of judgment ($F_{1,48}$ = 1.59, P = 0.2) as well as no interaction of category versus routine ($F_{1,48}$ = 0.01, P = 0.9). There was a highly significant effect of level of judge ($F_{2,48}$ = 5.9, P = 0.005). Figure 3 shows that novice judges spent less of their time looking away from the screen and at the scoring sheet compared with the national and international judges. There was also no difference in the percentage of time spent off the screen for national compared with that for international judges. Finally, there was no significant interaction of level of judge with category of judgment ($F_{2,48}$ = 0.95, P = 0.4) or routine ($F_{2,48}$ = 0.12, P = 0.9) and no three-way interaction ($F_{2,48}$ = 0.27, P = 0.8).

The behavioral analysis was repeated for the selected subgroup of errors (see Methods) to combine it with the eye position analysis. The categorical error analysis showed that the hit rate for these errors was 26% for novice judges, 38% for national judges, and 37% for international judges. Thus, for this selection of very prominent and easily identified errors, national judges performed equally well as international judges whereas novices were still worse, although the overall differences among the different groups of judges did not reach significance (X^2_2 = 2.9, P < 0.23). The ANOVA for the absolute score difference showed a significant difference among the three levels of judges ($F_{2,207}$ = 3.3, P < 0.04). The mean absolute difference for novices was 0.4 (SD, 0.25), whereas for national judges, it was 0.32 (SD, 0.27), and for international judges, it was 0.34 (SD, 0.26). Again, the significant difference was due to the worse performance of novices compared with the other two levels of judges.

Figure 4 presents the mean continuous time interval, that the eye position overlaid the error position, separately for errors that were reported and errors that were not reported by the three levels of judges. The mean time interval for reported errors was significantly larger than that for unreported errors for the novices ($t_{18.2}$ = 2.4, P < 0.03), whereas the mean time interval for reported errors versus unreported errors was marginally significant for the national judges ($t_{18.2}$ = 1.84, P < 0.07) and not significant for the international judges ($t_{18.2}$ = 0.7, P < 0.4). It can be seen in this figure that the difference in the time interval between reported and unreported errors decreased in a dose-response–related fashion with the increasing level of judge expertise from novice to national and international judges.

Table 2 shows the d-prime (eye) measure indicating how strong the effect that eye position had on a decision about whether an error in performance occurred. As can be seen in Table 2, only national judges showed a d-prime score that was significantly greater than zero (95% CI for d-prime do not include zero), indicating that this was the only group of judges that eye position had an effect on the detection of errors. On the contrary, eye position had no significant effect

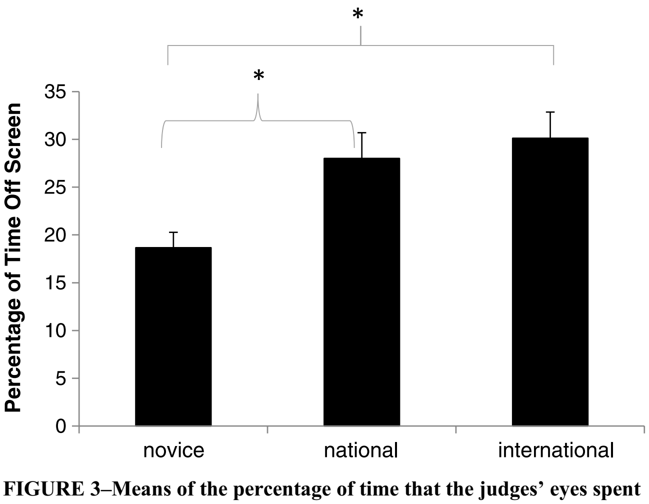

FIGURE 3–Means of the percentage of time that the judges' eyes spent off screen for all three levels of judges (*bars* denote SEM, and significant differences between groups are marked with an *asterisk* (*)).

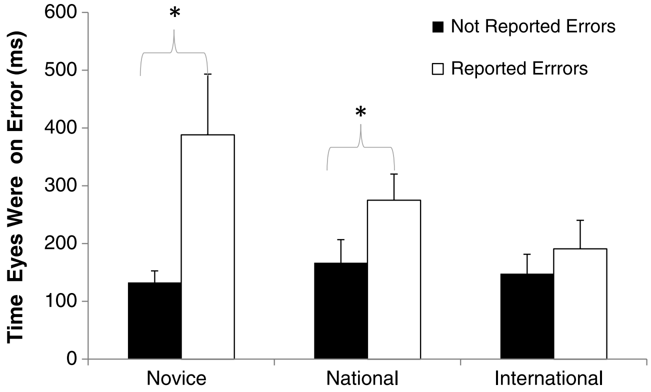

FIGURE 4–Mean time (ms) that the judges' eyes overlaid the error (averaged across members of each group and across errors) for all three levels of judges and for both reported and not reported errors (*bars* denote SEM, and significant differences between groups are marked with an *asterisk* (*)).

on either the novice group or the group of international judges (d-prime scores did not differ significantly from zero).

DISCUSSION

The first question that was addressed in this study was the success of judges of RG in detecting errors in the performance of five expert gymnasts executing an RG routine. In the laboratory setting of a judgment of an RG routine, an ideal situation was simulated in which they watched a video of the gymnasts performing the routine in a silent room and was not distracted by crowd noise or flashes that are usually distracting judgment in real sport events. In addition, the bias (e.g., patriotism, reputation) for judgment should be minimal in our case. This is the first study, to our knowledge, to evaluate the performance of RG judges compared with ideal performance, namely, the detection of all true errors and the assignment of true penalty scores. In this report, only errors in difficulty (D1 and D2) were considered. It should be noted here that the definition of true errors and calculation of true penalty scores were based on a reevaluation of the videos by the group of international judges and one more international judge that did not participate in the original scoring (author D. T.). The fact that the same five judges participated in the original experimental scoring could introduce a problem of external validity for the true penalty scoring procedure. We believe this problem was minimal for two reasons. The first is that the true penalty scoring was performed 6 months or longer after the original experimental scoring, so there were no carry-over effects, and the second and most important is that the true scoring was based on a consensus procedure among all judges (the five internal ones and the new external judge).

The main finding was that judge performance was dramatically lower than ideal although all groups performed scoring better than chance (d-prime scores were significantly above zero). Even international-level judges that had participated in several world competition events, such as the Olympic Games and the World Championship in RG, performed at a mediocre 40% error detection level. Greek national-level judges with much less experience in judging international RG events performed at a poor 20% detection level that did not differ from that of novice judges with no experience in judging official competition events. In fact, their performance overall was significantly worse compared with the performance of novices (significantly lower d-prime scores) and this was due to the larger number of false alarms, namely, cases where they gave a penalty to errorless performance. Our results clearly demonstrated that the level of experience has a highly significant effect on judging performance because international-level judges were twice as efficient as novice and national judges. Previous studies in the sports domain have also reported that experts were superior to novices in decision making (9,14,18).

The poor performance of RG judges is not surprising, considering the complexity of the task at hand. In these particular RG events, judges have to keep track of five gymnasts performing complex body movements at a high speed while they also have to keep track of the gymnast's interaction with an apparatus that also moves in complex trajectories at a high speed. The gymnasts move within a large area, so that sometimes, gymnasts obstruct view of others from the judges and, at other times, they spread so far apart that it becomes impossible for judges to track them simultaneously. While evaluating the routine, the judge repeatedly has to move his/her focus of attention from the field to the scoring sheet (looking down) to be informed about the upcoming exercises and to keep notes of the particular scores and then refocus their attention on the performing gymnasts. Indeed, our analysis of the time spent off screen showed that all judges spent a considerable percentage of their time not viewing the video. Interestingly, the novice judges spent significantly less time off screen (about 20%) compared with national and international judges, who spent about 30% of their time off screen. This applied to both categories of judgment (D1 and D2) and both routines (H and RR). This result was in contrary with previous reports by Ste-Marie (29), which showed that novice judges of gymnastics spend more time looking at the scoring sheet. This might be explained by a different strategy being used by novice judges during the evaluation procedure to cope with the increased

TABLE 2. Categorical classification of eye position in relation to error detection.

	Novice	National	International
Hits (eyes on detected error)	10	20	15
False alarms (eyes away from detected error)	8	7	11
Correct rejections (eyes away from undetected error)	23	25	27
Misses (eyes on undetected error)	26	25	27
d-Prime	0.06	0.64*	0.19
Low to high 95% CI	−0.19 to 0.32)	0.35 to 0.92	−0.08 to 0.46

All other d-prime scores were not significantly different from zero (95% CI includes zero).
*Only the d-prime score for national judges was significantly higher than zero (95% CI does not include zero).

complexity of this particular task, namely, the RG ensembles, where five gymnasts participate, in contrast to a task in gymnastics where only one gymnast participates. In addition, it was observed that many novice judges elected to keep score after they had finished watching the projected video while they kept their eyes on the screen during the projection.

The second part of this study considered the eye fixations of judges on particular errors. To avoid ambiguity at matching the eye fixation location with the location of the error, we selected a subset of D1 and D2 errors that satisfied the following conditions: the errors were very conspicuous (for example, a drop of the apparatus from the gymnast's hand), their location was composed of a single spatial area (for example, right–left corner of the screen of the video projection), and no other errors occurred simultaneously. The behavioral analysis of these errors showed that international-level judges were more efficient at detecting these errors (about 40%) compared with the other groups (about 23%). The analysis of eye movement fixations provided two important insights on the behavior of judges. The first insight was provided by measuring the total time that judges fixated at errors that were detected compared with that of errors that were not detected. The total time difference was significant only for novice judges. More specifically, novice judges spent significantly more time fixating at potential errors that they finally reported as errors compared with the time spent fixating at those they finally did not report as errors. The captivation of visual fixation at a particular location to process an error might be explained by the lack of experience of these novice judges at performing the task of judgment, resulting in spending an excessive amount of time processing particular visually fixated errors. This would inevitably result in loss of the capacity to simultaneously process other events, like the case of multiple errors, as indicated by the poor performance of these subjects.

The second insight came from the evaluation of the signal detection analysis results for all three categories of judges. In this analysis, a measure of efficiency of the use of eye fixation to detect errors was calculated (d-prime). This measure indicates whether fixated errors are actually more probable to finally be reported as errors. This analysis showed that novice judges, although spending a lot of time fixating the detected errors, were not efficient at using fixation to detect errors (the efficiency measured as d-prime score was not significantly different from zero). On the other hand, national-level judges were much more efficient at using visual fixation to detect errors, in a sense that the proportion of reported errors that were fixated was larger than the proportion of reported errors that were not fixated (d-prime of 0.6 that was significantly larger than zero). This efficiency difference also correlated with the difference in overall performance of the national-level judges compared with that of the novice judges on this particular subset of errors. On the basis of these results, one would expect international-level judges to make an efficient use of visual fixation to detect errors. Interestingly enough, our results clearly indicated this not to be the case. International-level judges did not rely on using visual fixation to detect errors, whereas national-level judges did, and their overall d-prime was not significantly greater than zero, as was the case for the novice judges, whereas their error detection performance was similar to that of the national judges. This result suggests a different strategy used by international judges for error detection, which does not depend on visual fixation of the errors. International judges probably rely on more complex cognitive strategies based on extensive experience and a larger knowledge base compared with those of novice judges (3,5,13,26,28). Such strategies that might not be based at all on specific visual perception mechanisms could help them detect a larger number of errors overall. Perceptual anticipation is one of these strategies that have been reported in sports literature (28,31). Anticipation of an upcoming gymnastic element is likely to be based on advanced visual cues that expert judges may be able to identify earlier than novice judges in a given gymnastic sequence. This information could result in reduction in the demands of processing when that gymnastic element is performed (28,31).

In conclusion, this study showed that the performance of judges of RG is far below optimum at the five-gymnast ensemble routines. Experience was found to have a significant effect in performance. International-level judges outperformed national-level judges and novices, whereas national-level judges were as bad, or even worse, than novices. The visual fixation of novice judges was captivated by particular errors, whereas national judges were significantly more efficient at using eye fixation to detect errors. Finally, international-level judges did not rely on eye fixation to detect errors but probably made use of other cognitive strategies, increasing their overall error detection efficiency.

This research was kindly supported by the Greek Organization of Football Prognostics S. A. (OPAP).

Dimitris Mylonas has received scholarship from the European Union (European Social Fund) and Greek national funds through the operational program 'Education and Lifelong Learning' of the National Strategic Reference Framework–Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund. For the remaining authors, none were declared.

The results of the present study do not constitute endorsement by the American College of Sports Medicine.

REFERENCES

1. Abernethy B. Anticipation in squash: differences in advance cue utilization between expert and novice players. *J Sports Sci*. 1990; 8:17–34.
2. Abernethy B, Russell DG. Expert-novice differences in an applied selective attention task. *J Sport Psychol*. 1987;9:326–45.
3. Abernethy B, Neal RJ, Koning P. Visual perceptual and cognitive differences between expert, intermediate, and novice snooker players. *Appl Cogn Physiol*. 1994;8:185–211.
4. Allard F, Starkes JL. Perception in sport: volleyball. *J Sport Psychol*. 1980;2:22–33.

5. Allard F, Graham S, Paarsalu ME. Perception in sport: basketball. *J Sport Psychol*. 1980;2:14–21.
6. Ansorge CJ, Scheer JK. International bias detected in judging gymnastic competition at the 1984 Olympic Games. *Res Q Exerc Sport*. 1988;59:103–7.
7. Bard C, Fleury M, Carriere L, Halle M. Analysis of gymnastic judges' visual search. *Res Q Exerc Sport*. 1980;51:267–73.
8. Boen F, Van Hoye K, Auweele YV, Feys J, Smits T. Open feedback in gymnastic judging causes conformity bias based on informational influencing. *J Sports Sci*. 2008;26(6):621–8.
9. Catteeuw P, Helsen W, Gilis B, Van Roie E, Wagemans J. Visual scan patterns and decision-making skills of expert assistant referees in offside situations. *J Sport Exerc Psychol*. 2009;31:786–97.
10. Cavanagh P, Alvarez G. Tracking multiple targets with multifocal attention. *Trends Cogn Sci*. 2005;9:349–54.
11. Dosseville F, Laborde S, Raab M. Contextual and personal motor experience effects in judo referees' decisions. *Sport Psychol*. 2011;25:67–81.Available from: http://journals.humankinetics.com/tsp.
12. Duchowski AT. A breadth-first survey of eye-tracking applications. *Behav Res Methods Instrum Comput*. 2002;34(4):455–70.
13. Gobet F, Simon HA. The roles of recognition processes and look-ahead search in time-constrained expert problem solving: evidence from a grand-master-level chess. *Psychol Sci*. 1996;7:52–5.
14. Hancock D, Ste-Marie D. Gaze behaviors and decision making accuracy of higher- and lower-level ice hockey referees. *Psychol Sport Exerc*. 2013;14:66–71.
15. Intriligator J, Cavanagh P. The spatial resolution of visual attention. *Cogn Psychol*. 2001;43:171–216.
16. Jacob RJK, Karn KS. Eye tracking in human-computer interaction and usability research: ready to deliver the promises. In: Hyönä J, Radach R, Deubel H, editors. *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*. Amsterdam: Elsevier; 2003. pp. 573–605.
17. Macmillan NA, Creelman CD. *Detection Theory. A User's Guide*. 2nd ed. Lawrence Erlbaum Associates; 2005. pp. 16–21, 323–37.
18. Mann DT, Williams AM, Ward P, Janelle CM. Perceptual-cognitive expertise in sport: a meta-analysis. *J Sport Exerc Psychol*. 2007;29: 457–78.
19. Moreno FJ, Reina R, Luis V, Sabido R. Visual search strategies in experienced and inexperienced gymnastic coaches. *Percept Mot Skills*. 2002;95:901–2.
20. Nevill AM, Balmer NJ, Williams AM. The influence of crowd noise and experience upon refereeing decisions in football. *Psychol Sport Exerc*. 2002;3:261–72.
21. Oksama L, Hyönä J. Is multiple object tracking carried out automatically by an early vision mechanism independent of higher-order cognition? An individual difference approach. *Vis Cogn*. 2004;11:631–71.
22. Oksama L, Hyönä J. Dynamic binding of identity and location information: A serial model of multiple identity tracking. *Cogn Psychol*. 2008;56:237–83.
23. Pylyshyn ZW, Storm RW. Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spat Vis*. 1988;3: 179–97.
24. Ripoll H, Kerlizin Y, Stein J-F, Reine B. Analysis of information processing, decision making, and visual strategies in complex problem solving sport situations. *Hum Mov Sci*. 1995;14(3): 325–49.
25. Singer RN, Cauraugh JH, Chen D, Steinberg GM, Frehlich SG. Visual search, anticipation and reactive comparisons between highly-skilled and beginning tennis players. *J Appl Sport Psychol*. 1996;8:9–26.
26. Starkes J, Allard F, Lindley S, O'Reilly K. Abilities and skill in basketball. *Int J Sport Psychol*. 1994;25:249–65.
27. Stefani R. Predicting outcomes. In: Bennett J, editor. *Statistics in Sport*. London (United Kingdom): Arnold; 1998. pp. 249–75.
28. Ste-Marie D. Expert-novice differences in gymnastics judging: an information processing perspective. *Appl Cogn Psychol*. 1999; 13:269–81.
29. Ste-Marie DM. Expertise in women's gymnastic judging: an observational approach. *Percept Mot Skills*. 2000;90:543–6.
30. Ste-Marie DM. Expertise in sport judges and referees: circumventing information-processing limitations. In: Starkes JL, Ericsson KA, editors. *Expert Performance in Sport*. Leeds (United Kingdom): Human Kinetics; 2003. pp. 169–90.
31. Ste-Marie DM, Lee TD. Prior processing effects on gymnastic judging. *J Exp Psychol Learn Mem Cogn*. 1991;17:126–36.
32. Vickers JN. Knowledge structures of expert-novice gymnasts. *Hum Mov Sci*. 1988;7:47–72.
33. Williams AM, Davids K. Visual search strategy, selective attention, and expertise in soccer. *Res Q Exerc Sport*. 1998;69:111–28.