Journal of Sports Sciences
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rjsp20

Compensating tendencies in penalty kick decisions of referees in professional football: Evidence from the German Bundesliga 1963-2006

Wolf Schwarz [a]
[a] Fakultät für Humanwissenschaften, University of Potsdam, Potsdam, Germany
Published online: 10 Jan 2011.

Journal of Sports Sciences, March 1st 2011; 29(5): 441–447

Compensating tendencies in penalty kick decisions of referees in professional football: Evidence from the German Bundesliga 1963–2006

WOLF SCHWARZ

Fakultät für Humanwissenschaften, University of Potsdam, Potsdam, Germany

(Accepted 5 November 2010)

Abstract

Using a large representative database (12,902 matches from the top professional football league in Germany), I show that the number (441) of two-penalty matches is larger than expected by chance, and that among these 441 matches there are considerably more matches in which each team is awarded one penalty than would be expected on the basis of independent penalty kick decisions (odds ratio = 11.2, relative risk = 6.34). Additional analyses based on the score in the match before a penalty is awarded and on the timing of penalties, suggest that awarding a first penalty to one team raises the referee's penalty evidence criterion for the same team, and lowers the corresponding criterion for the other team.

Keywords: *Decision-making, biased decisions, Bundesliga, sports statistics*

Introduction

In many team sports, referees need to make difficult and close decisions under extreme time pressure, decisions that in sports such as professional football can have considerable public and economic impact. Accordingly, there has been growing research interest in improving understanding of perceptual (e.g. Oudejans et al., 2000) and cognitive limitations and decisional biases associated with such decisions (for recent reviews, see Catteeuw, Helsen, Gilis, & Wagemans, 2009; Mascarenhas, O'Hare, & Plessner, 2006).

One line of evidence from this research suggests that part of the intricacy of the referee's decisions may actually result from the referee's own previous decisions during the same match. In an experimental study of penalty kick decisions in football involving 115 participants (57 licensed referees, 58 players), Plessner and Betsch (2001) found that the participants' decisions about videotaped scenes earlier in the match had a clear impact on their decisions about events later in the same match. Specifically, if a penalty had already been awarded to one team, then the same team was less likely, and the other team more likely, to be awarded another penalty in an ambiguous situation arising later in the same match. This pattern may be interpreted as a "compensating bias" of the participants, or an equalizing tendency to award (or refuse) penalties in a manner that leads to a balanced penalty distribution – one that is more even than objective decisions would require (for a corresponding analysis of NCAA basketball matches, see Anderson & Pierce, 2009).

One concern about controlled laboratory studies is the extent to which they can be generalized to a more natural setting of expert refereeing (Mascarenhas, Collins, & Mortimer, 2002). For example, laboratory studies are typically conducted within a relatively pressure-free context in which decisions have no real-world impact; they also present video scenes from a spectator's – or sideline – viewing position that is quite different from the typical perspective taken by a referee. In contrast, as noted earlier, real-world referees have to make decisions under extreme and public time pressure, and their decisions often have considerable consequences. In the view of Mascarenhas et al. (2002), these factors are better understood within a natural decision-making context, for example, by directly asking real-world referees about the reasons underpinning their decisions. Plessner and Betsch (2002, p. 336) argue that this natural decision-making approach has its own weaknesses, and suggest a combination and integration of different approaches (see, for example, Brand, Schmidt, & Schneeloch, 2006).

Correspondence: W. Schwarz, University of Potsdam, Fakultät für Humanwissenschaften, Haus 24, Karl-Liebknecht Strasse 24–25, D-14476 Potsdam, Germany. E-mail: wschwarz@uni-potsdam.de

ISSN 0264-0414 print/ISSN 1466-447X online © 2011 Taylor & Francis
DOI: 10.1080/02640414.2010.538711

Mascarenhas et al. (2002) suggest that it would be most appropriate to study the referee's decision-making in a naturalistic context of competitive matches. Arguably, then, the most relevant and valid naturalistic databases available are those that describe in considerable detail top football league competitions. Thus, it was the aim of the present study to provide empirical evidence and detailed statistical analyses of potential "compensating tendencies" in the context of penalty kick decisions made in "real-world" professional football competitions. To this end, a large and representative database describing the first 12,902 matches (during the years 1963–2006) from the top German football league (the "Bundesliga") was analysed. The main strategy of data analysis was to determine whether successive penalty decisions within a match are made independently. I consider in detail if or when lack of independence can be interpreted as evidence in favour of compensating tendencies.

Data

The database was originally created for commercial purposes by the IMP AG, Munich, Germany (I thank Holger Rahlfs and Jörn Wendland for making these data available). The database consists of a complete record of all penalties in all matches of the first 42 Bundesliga seasons 1963/64 to 2004/05, plus the first 12 matchdays of the season 2005/06, for a total of 12,902 matches. The correctness of the records was verified for two randomly selected matchdays per year. For each penalty, information is provided about the minute of the match (1–90) in which it was awarded, the score of the match at that instant, the team (home vs. away) to which it was awarded, and whether the penalty was converted or missed.

Results

The distribution of penalties across all matches

In the 12,902 matches, 3713 penalties were awarded, an average of 0.288 per match, about two penalties every seven matches. Of all penalties, 2621 were awarded to the home team, yielding an overall home-penalty proportion of 0.706.

Given these base rates, how should the 3713 penalties be distributed across the 12,902 matches under a model of "complete randomness"? The standard notion of a completely random point process is the Poisson process (cf. Agresti, 2002; Fleiss, Levin, & Paik, 2003; for applications of the Poisson model to football, see Baxter & Stevenson, 1988; Ridder, Cramer, & Hopstaken, 1994). In the Poisson counting model, at each moment t of the match there is an instantaneous event (i.e. penalty) rate, $\mu(t)$. This rate may not be constant during any one match, but from the current data the best estimate of the overall penalty count per match (which is the integral of $\mu(t)$ across one match) is $\lambda = 3713/12902$. Table I shows the observed frequencies of matches with 0, 1, 2, 3, 4, and 5 penalties, together with the frequencies theoretically expected under the Poisson counting model.

Table I. Observed and predicted number of matches with 0, 1, 2, 3, 4, and 5 penalties.

Penalties	Matches observed	Matches predicted	Total penalties	Contribution to χ^2
0	9720	9675.51	0	0.20
1	2699	2784.46	2699	2.62
2	441	400.66	882	4.06
3	37	38.44	111	0.05
4	4	2.77	16	} 1.46
5	1	0.16	5	
Total	12,902	12,902	3713	8.39

Note: Predicted frequencies are derived from a Poisson model with $\lambda = 3713/12{,}902$. Due to small expected counts, the last two categories were collapsed to compute their contribution to χ^2.

Overall, the Poisson fit appears not uniformly inadequate, but two discrepancies from the expected Poisson frequencies stand out, leading to a significant $\chi^2(\text{d.f.} = 3) = 8.39$ $(P = 0.039)$. These discrepancies are: (1) there are fewer matches than expected with exactly one penalty, and (2) there are considerably more matches than expected with exactly two penalties.

At this gross level of analysis, the observed pattern appears to be consistent with the view that referees show a slight tendency to avoid the awarding of exactly one penalty in a match, and that there is a marked excess (10% more than expected) of matches with exactly two penalties. Evidently, these data are just correlational in nature, and therefore alternative interpretations need to be considered as well (cf. Kuss, Kluttig, & Stoll, 2007). To this end, the evidence will be examined in more detail by turning to a more informative analysis of those 441 matches in which exactly two penalties were awarded.

The dependency between successive penalty decisions in two-penalty matches

The basic issue to be addressed is that under a model of independent (of course, not necessarily equal) awarding of penalties to the home and away team, the identity of the team to which the first penalty is awarded should not influence the probabilities for a second penalty to be awarded to one or other team. That is, the second penalty should be awarded independent of the team to which the first penalty was awarded.

Table II provides a breakdown of the 441 matches involving exactly two penalties, showing the team (home vs. away) to which the first and the second penalty were awarded. Let the notation 1 = H stand for the event that the first penalty was awarded to the home (H) team and 1 = A that it was awarded to the away (A) team; 2 = H and 2 = A are used similarly for the second penalty. Of these 882 penalties, 584 were awarded to the home team, an overall proportion (0.662) that conforms closely to the grand average home-penalty proportion of 0.706. However, inspecting the entries in Table II, a gross violation of independence is immediately apparent. For example, if the first penalty was awarded to the away team (a total of 146 matches), then the second penalty was again awarded to the away team in only 11 of those matches, yielding a proportion of $P(2 = A|1 = A) = 0.075$. In contrast, if the first penalty was awarded to the home team (a total of 295 matches), then the second penalty was awarded to the away team in 141 of those matches, a proportion of $P(2 = A|1 = H) = 0.477$. That is, for the away team the probability to be awarded the second penalty is increased markedly from 7.5% to about 48% if it had versus had not been awarded the first penalty – a more than eleven-fold (11.2) increase of the corresponding odds, or a relative risk of 6.34. Correspondingly, relative to the expectations under independence, Table II shows far too many counts in the main diagonal (both teams awarded one penalty) and too few counts in the off diagonal (both penalties awarded to the same team).

Table II. Breakdown of penalties in 441 matches with exactly two penalties.

First penalty	Second penalty Home	Away	Total
Home	154 (193.3)	141 (101.7)	295
Away	135 (95.7)	11 (50.7)	146
Total	289	152	441

Note: The expected proportion of second penalties awarded to the home team independent of the team to whom the first penalty was awarded is 0.665. The observed proportions are 0.522 if the first penalty was awarded to the home team, and 0.925 if it was awarded to the away team. Counts expected under independence are given in brackets. $\chi^2(\text{d.f.} = 1) = 70.09$, $P < 0.001$. The odds ratio $\omega = 11.24$ (relative risk 6.34), and its 95% confidence interval is [5.8, 21.6].

These observations are confirmed by more formal analyses, $\chi^2(\text{d.f.} = 1) = 70.09$ ($P < 0.001$), for the usual χ^2-test of independence. The corresponding odds ratio equals 11.24 (corresponding to a relative risk of 6.34) and its 95% confidence interval is [5.8, 21.6].

If there is a general compensation bias, it should not be limited just to matches involving exactly two penalties. As shown in Table III, there were 37 matches in which exactly three penalties were awarded; of these 111 penalties, 72 were awarded to the home team. Assuming independent penalty decisions, binomial variation with $P = 72/111 = 0.649$ would lead to the expected frequencies shown in Table III. Evidently, the observed data deviate significantly from the expectation under independence ($\chi^2(\text{d.f.} = 6) = 26.85$, $P < 0.001$). To understand the nature of these discrepancies, note there are much fewer matches (3) than expected (11.7) in which all three penalties were awarded to the same team. That is, the distribution of the three penalties among the two teams is more evenly balanced than predicted by a model of independent decisions, in line with what is expected if compensation bias is present. Specifically, consider that $P(3 = A|1 = H,\ 2 = H) = 0.667$. That is, after the home team has been awarded the first two penalties, the probability for the away team to be awarded the third penalty is nearly doubled compared with the baseline of 0.351.

Similarly, the notion of a compensation bias suggests that following the first penalty award, this bias should be present if that penalty was converted, but should be weaker or even absent if that penalty was missed. Table IV provides a breakdown of all 295 two-penalty matches in which the first of these penalties was awarded to the home team (1 = H). These matches were classified according to whether the home team converted or missed this first penalty, and whether the second penalty was then awarded to the home team or to the away team. If the home team converted the first penalty (which happened in 219 of these matches), in 117 matches the second penalty

Table III. Distribution of penalties for 37 matches with exactly three penalties.

Order:	HHH	HHA	HAH	AHH	HAA	AHA	AAH	AAA	Total
Observed	3	6	14	9	0	1	4	0	37
Expected	10.1	5.5	5.5	5.5	3.0	3.0	3.0	1.6	37

Note: Number of matches with a given order of penalties for all 37 matches involving exactly three penalties. For example, the entry HAH stands for a match in which the first and third penalties were awarded to the home team, the second penalty to the away team. Given that overall 72 of these 111 penalties were awarded to the home team, expected frequencies under independence are calculated on the basis of binomial variation with $P = 72/111 = 0.649$. The observed data deviate significantly from the frequencies expected under independence ($\chi^2(\text{d.f.} = 6) = 26.85$, $P < 0.001$).

Table IV. Breakdown of penalties according to converted vs. missed first penalty.

First penalty	Second penalty Home	Away	Total
Converted	102	117	219
Missed	52	24	76
Total	154	141	295

Note: The data relate to two-penalty matches in which the first penalty was awarded to the home team. $\chi^2(\text{d.f.}=1)=10.79$, $P<0.001$. The odds ratio $\omega=2.485$ (relative risk 1.69), and its 95% confidence interval is [1.39, 4.52].

Table V. Breakdown of penalties in 441 matches with exactly two penalties.

First penalty	Second penalty Home	Away	Total
I. Home team leads before second penalty is awarded			
Home	102 (116.4)	111 (95.6)	213
Away	39 (24.6)	6 (20.4)	45
Total	141	117	258
II. Away team leads before second penalty is awarded			
Home	19 (23.1)	7 (2.9)	26
Away	61 (56.9)	3 (7.1)	64
Total	80	10	90

Note: For the upper part (I) of the table, the expected proportion of second penalties awarded to the home team under independence of the team to whom the first penalty was awarded is 0.547. The observed proportions are 0.479 if the first penalty was awarded to the home team, and 0.867 if it was awarded to the away team. For the lower part (II) of the table, the expected proportion is 0.889. The observed proportions are 0.731 if the first penalty was awarded to the home team, and 0.953 if it was awarded to the away team. Counts expected under independence are given in brackets. For the upper table (I), $\chi^2(\text{d.f.}=1)=22.54$, $P<0.001$. The odds ratio $\omega=7.07$ (relative risk 3.91), and its 95% confidence interval is [2.9, 17.4]. For the lower table (II), $\chi^2(\text{d.f.}=1)=9.26$, $P<0.003$. The odds ratio $\omega=7.49$ (relative risk 5.74), and its 95% confidence interval is [1.8, 31.8].

was awarded to the away team. That is, $P(2=A|1=H \text{ and converted})=0.534$, which is significantly larger than the away team's base probability of 0.294, in line with the prediction of a compensating bias. In contrast, if the home team missed the first penalty (76 matches), then in only 24 matches was the second penalty awarded to the away team, i.e. $P(2=A|1=H \text{ and missed})=0.316$, which is not significantly larger than the away team's base probability of 0.294. This difference between converted and missed first penalties is highly significant, $\chi^2(\text{d.f.}=1)=10.79$ $(P<0.001)$; it suggests that referees show a compensating bias if the first penalty is converted, whereas second penalties are awarded "as usual" if the first penalty is missed. A corresponding breakdown for matches in which the first penalty was awarded to the away team produces a similar qualitative pattern, but too few matches for further statistical analysis.

The data in Tables II–IV are clearly consistent with the hypothesis that referees show a compensating bias that acts against the team to which the first (or the first and second) penalty was awarded. More specifically, awarding a first penalty to one team may raise the decision criterion for the evidence required to award a second penalty to the same team, and lower the corresponding criterion for the other team. However, as noted above, the data are correlational in nature, so that alternative accounts need to be considered as well. Perhaps the most general and relevant alternative interpretation (cf. Mascarenhas et al., 2002, p. 330) is that as a consequence of the first penalty (and, in particular, of its conversion) awarded to one team, the other team will often be behind, and thus adopt a more offensive playing strategy than before. This in turn might lead to more action in the penalty area of the team that had been awarded the first penalty and therefore tend to balance out penalties. Specifically, based on this view, the large entries in the main diagonal of Table II would not indicate any compensating bias in the referees' decision-making but simply reflect an effect of a strategy change due to the changed score in the match.

Alternative accounts of the dependency between successive penalty decisions

To address the alternative interpretation described in the last paragraph further, the data in Table II were broken down to distinguish between cases in which during the period before the second penalty was awarded which of the home versus the away team was in the lead ("H leads", "A leads"). Under the alternative "changed score, changed strategy" account, the large effect seen in Table II simply reflects that the first penalty will often change the score, and this in turn will lead to a corresponding change of playing strategy. If true, then one would expect that if one conditions on the team that leads during the period before the second penalty is awarded, then it should be irrelevant who had received the first penalty. For example, suppose that the away team is in the lead at the moment before the second penalty is awarded. According to the alternative account, this should imply an offensive strategy of the home team at this point of time, no matter who had been awarded the first penalty. Put more formally, under the alternative account it is expected that $P(2=H|1=A, \text{A leads})=P(2=H|1=H, \text{A leads})$, and similarly it is expected that $P(2=H|1=A, \text{H leads})=P(2=H|1=H, \text{H leads})$.

The data shown in Table V clearly contradict this prediction: even after conditioning on the team that was in the lead before the second penalty, there is still a highly significant influence of the first penalty on the probabilities for each team to be awarded the second penalty. In fact, conditioning on the leading team produces only a modest lowering of the odds ratio of 11.2 in Table II to 7.1 and 7.5 for the top and bottom of Table V respectively, values that still differ significantly from 1 (see Table V).

A slightly different way of addressing this issue is to compare the division of the second penalties to the home versus away team shown in Table V to the situation in which no penalty has yet been awarded. To this end, all 2699 matches involving exactly one penalty were analysed, the conditional probabilities P(1 = H|H leads) and P(1 = H|A leads) computed, and then separately for each case of leading team I evaluated if a previously awarded (first) penalty had an effect, relative to these baselines. In one-penalty matches, P(1 = H|H leads) = 0.667 and P(1 = H|A leads) = 0.743, confirming that home teams are generally awarded more penalties, but especially so if the away team is in the lead. Using the value of P(1 = H|H leads) = 0.667 for the second penalty data in Table V (I), and P(1 = H|A leads) = 0.743 for the second penalty data in Table V (II), significant differences were again found in the expected direction between the ways in which first and second penalties are awarded, even after controlling for the team that was in the lead when the penalty was awarded (χ^2(d.f. = 1) = 41.90, $P < 0.001$ when H leads, and χ^2(d.f. = 1) = 18.62, $P < 0.001$ when A leads).

It is thus concluded that even after conditioning on the score of the match before the second penalty is awarded, the "history" of who had been awarded the first penalty remains a crucial factor, as predicted by the "compensating bias" account of the data in Table II, but not by the alternative account. This conclusion should, however, not be taken to mean that the "changed score, changed strategy" effect does not exist. In fact, apart from the modest lowering of the odds reported in Table V, other evidence from the database suggests that the score immediately before a penalty is awarded does have a systematic (albeit relatively small) effect on which team will be awarded that penalty. For example, it is instructive to look retrospectively at the score before one of the 3182 first penalties of all 12,902 matches was awarded. When the home team was awarded the first penalty, it was already leading before that penalty in 36.3% of all cases. In comparison, when the away team was awarded the first penalty, the home team was leading before that penalty in 46.4% of all cases. This significant (χ^2(d.f. = 1) = 27.85, $P < 0.001$) increase presumably reflects a more offensive playing strategy adopted by the away team when it is behind. However, the present analysis (Table V) clearly suggests that the large effect that is evident in Table II cannot be reduced to this adaptive mechanism. It should also be noted that it cannot logically be ruled out that officials may be more inclined to award a penalty kick to a trailing team, or less inclined to award a penalty to a leading team. However, the current score is usually (for example, if not due to a previous penalty) relatively independent of the referee's own active previous decisions and thus cannot normally be attributed by him to himself (but rather to the performance of the teams). Therefore, it is less obvious that or why a referee would change his inclination simply on the basis of changing scores. On the other hand, almost inevitably a referee must attribute the consequences of his own previous penalty decisions much more directly to himself, which in turn would appear to provide a much stronger context for "compensating decisions".

Temporal aspects of successive penalty decisions

According to the hypothesis advanced above, referees tend to compensate their first penalty decision by a tendency to award a second penalty to the other team, and are more hesitant than before the first penalty to award another penalty to the same team. If correct, this interpretation leads one to expect that the time (in minutes of the match) from the first to the second penalty should on average be shorter for those matches in which one penalty is awarded to each team (i.e. the case of a "compensation" that referees tend to prefer) than for those matches in which one team is awarded both penalties (the unbalanced case that referees tend to avoid). An important methodological prerequisite for such a comparison to be valid is that there exist no baseline differences in the time of the first penalty between these groups of matches. If such baseline differences existed, any further difference in the latency from the first to the second penalty could simply be a consequence of the different timing of the first penalty (for example, the time left) rather than the effect of a compensation bias.

The left-hand column in Table VI shows the mean minute of the match in which the first penalty was awarded, separately for each of the four combinations shown in Table II. Clearly, these means do not depend on the team to which the first, or on the team to which the second, penalty is awarded, or on whether or not both penalties are awarded to the same team. Table VI (right-hand column) also shows the mean latencies from the first to the second penalty, which differ significantly ($F_{3,437} = 4.51$, $P = 0.004$). Individual *post-hoc* comparisons reveal two significant differences between these four means. The time from the first to the second penalty is significantly shorter (22.7 min) if the first penalty is awarded to the away team and then compensated by a second penalty awarded to the home team,

Table VI. Times of first penalty and from first to second penalty for 441 matches with exactly two penalties.

Team 1st penalty	Team 2nd penalty	Time of 1st penalty (min)	Time from 1st to 2nd penalty (min) (95% CI)	n
Home	Home	38.6	30.7 (27.5, 33.8)	154
Home	Away	39.0	26.7 (23.4, 29.9)	141
Away	Home	39.9	22.7 (19.4, 26.1)	135
Away	Away	38.0	35.3 (23.6, 46.9)	11
Mean		38.9	27.1	441

Note: The four means for the time of the first penalty do not differ ($F_{3,437} = 0.118$, $P = 0.95$). The four means (in brackets: the 95% confidence intervals) for the time from the first to the second penalty differ significantly ($F_{3,437} = 4.51$, $P = 0.004$). An LSD post hoc test indicates that the mean (22.7) for the combination away/home is significantly shorter than for the combinations home/home (mean 30.7, $P = 0.001$) and away/away (mean 35.3, $P = 0.044$).

Table VII. Expected joint home vs. away team probabilities for first and second penalties under a model assuming heterogeneity of the proneness for home team penalties across referees.

	Second penalty		
First penalty	Home	Away	Total
Home	$\mu_\pi^2 + \sigma_\pi^2$	$\mu_\pi(1-\mu_\pi) - \sigma_\pi^2$	μ_π
Away	$\mu_\pi(1-\mu_\pi) - \sigma_\pi^2$	$(1-\mu_\pi)^2 + \sigma_\pi^2$	$1-\mu_\pi$
Total	μ_π	$1-\mu_\pi$	1

Note: The entries in the table give the expected probability of each indicated joint event if across the population of referees the proneness π for home team penalties has a mean of μ_π and a variance of σ_π^2. Note that $\sigma_\pi^2 = 0$ produces independence, and that $\sigma_\pi^2 > 0$ generally implies positive dependence between successive penalty decisions.

compared with the two "same-team" cases – that is, when both penalties are awarded to the home team (mean inter-penalty time of 30.7 min, $P = 0.001$) or both penalties are awarded to the away team (mean inter-penalty time of 35.3 min, $P = 0.044$).

The pattern described in Table VI is consistent with the view that awarding a penalty to one team subsequently lowers the referee's criterion to award a second "compensating penalty" to the other team, and raises the criterion to award another penalty to the same team. Accordingly, this tendency manifests itself as systematically shorter mean inter-penalty times for compensating penalties compared with the mean inter-penalty times when both penalties are awarded to the same team.

Discussion

In an analysis of the first 12,902 matches from the top football league in Germany, fewer one-penalty matches and considerably more two-penalty matches were found than would be expected on a purely random (i.e. Poisson) basis. Specifically, the 441 matches involving exactly two penalties reveal significantly more matches with one penalty for each team than expected under independent decisions, producing an odds ratio of 11.2, corresponding to a relative risk of 6.34. This large effect is only slightly reduced by conditioning on the score before the second penalty, which argues against an interpretation in which this effect simply reflects a change of playing strategies adopted by the teams as a consequence of the (conversion of the) first penalty. This finding supports the interpretation that the first penalty awarded to a team raises the evidence criterion used by referees to award a further penalty to the same team, and lowers his criterion to award a penalty to the other team. This interpretation is supported by an analysis showing that if the first penalty is missed, then no compensating tendency is seen in the awarding of second penalties, and by the finding of a corresponding compensating effect in matches involving exactly three penalties. Further support for an interpretation in terms of a compensating mechanism comes from an analysis of the timing of penalty decisions: although there exist no differences in the mean time of the first penalty, the latency from the first to the second penalty is much shorter for "compensating" penalties (especially if the first penalty was awarded to the away team) compared with matches in which both penalties were awarded to the same team. Overall, these findings support the conclusions reached by Plessner and Betsch (2001), that referees show an equality-oriented bias induced by their own decisions earlier in the same match. Mascarenhas et al. (2002) critically discussed in some detail the external validity of Plessner and Betsch's study, which used a controlled design in which all 115 participants judged the same ambiguous video scenes from the perspective of a spectator. However, the present statistical analysis of 12,902 matches from a representative professional football league suggests conclusions similar to those of Plessner and Betsch (2001) when the penalty decisions of referees in "real" matches are studied. More general analyses modelling, for example, the number of penalties for each match as the dependent variable could be based on generalized linear mixed Poisson models with crossed random effects (see Agresti, 2002, Ch. 12) if a database is available that contains more, and more relevant, predictors than the present one.

One alternative interpretation not considered so far is that the large effect seen in Table II reflects heterogeneous dispositions of individual referees to award penalties to home or away teams. Specifically, even if for any individual referee successive penalty decisions were truly independent, differences in the probability to award penalties to the home versus away team between the referees could still induce dependencies across the set of all matches.

It is straightforward to consider the nature and consequences of heterogeneity across referees (for a detailed treatment of heterogeneity effects in terms of generalized linear mixed models, see Agresti, 2002, Ch. 12). Let π_i be the marginal probability of referee i to award a penalty to the home team rather than to the away team (i.e. the home-team penalty proneness of referee i). Assume that across all referees, π_i has a distribution with mean μ_π and variance $\sigma_\pi{}^2$. Table VII shows the joint probabilities for the first and second penalty to be awarded to either team, across all referees. For example, for an individual referee i the probability to award both penalties to the home team equals π_i^2; the expected overall proportion of such matches is therefore equal to $E[\pi^2] = \mu_\pi{}^2 + \sigma_\pi{}^2$; other entries arise similarly. The table shows that if and only if $\sigma_\pi{}^2 = 0$, the entries are the product of the marginal probabilities so that successive penalty decisions are independent. More important still is that if $\sigma_\pi{}^2 > 0$, then successive decisions are necessarily positively correlated (see also Agresti, 2002, Ch. 12.2): for example, the probability for the home team to be awarded the second penalty is generally larger if the first penalty had already been awarded to the home team compared with it being awarded to the away team, $P(2 = H|1 = H) > P(2 = H|1 = A)$. This is a direct consequence of the dispositional differences assumed across referees: the fact that the first penalty was awarded to the home team increases the posterior probability that the referee is characterized by a large value of π, which in turn makes it more likely that he will award the second penalty to the home team as well. This type of positive dependency is, however, exactly opposite to the pattern reported in Table II, and is thus principally unable to account for the pattern of "compensatory" decisions that are characterized by a negative dependency. In fact, these heterogeneity considerations seem to provide further indirect evidence for the compensatory interpretation advanced above: given that at least some dispositional differences in π across referees seem unavoidable, and given furthermore that such differences generally produce a positive dependency of successive penalty decisions, the data reported in Table II would seem even stronger evidence in favour of compensating mechanisms.

In summary, the present analyses of a large and representative database suggest that awarding a first penalty to one team raises the referee's penalty evidence criterion for the same team, and lowers the corresponding criterion for the other team. This may be interpreted as an adaptive, equality-oriented decision rule that creates contingencies to "balance things out". One can speculate that (implicitly or explicitly) professional players might seek to take strategic advantage of such biased contingencies. Clear documentation, conscious awareness of compensating biases, and explicit consideration of such decisional mechanisms in the training of referees are likely to be important steps to arrive at more objective penalty kick decisions in professional football.

Acknowledgement

I would like to thank Prof. Jeff O. Miller, University of Otago, New Zealand, for his helpful commentaries.

References

Agresti, A. (2002). *Categorical data analysis* (2nd edn.). New York: Wiley.

Anderson, K. J., & Pierce, D. A. (2009). Officiating bias: The effect of foul differential on foul calls in NCAA basketball. *Journal of Sports Sciences*, *27*, 687–694.

Baxter, M., & Stevenson, R. (1988). Discriminating between the Poisson and negative binomial distributions: An application to goal scoring in association football. *Journal of Applied Statistics*, *15*, 347–354.

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

Catteeuw, P., Helsen, W., Gilis, B., & Wagemans, J. (2009). Decision-making skills, role specificity, and deliberate practice in association football refereeing. *Journal of Sports Sciences*, *27*, 1125–1136.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd edn.). New York: Wiley.

Kuss, O., Kluttig, A., & Stoll, O. (2007). The fouled player should not take the penalty himself: An empirical investigation of an old German football myth. *Journal of Sports Sciences*, *25*, 963–967.

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

Mascarenhas, D., O'Hare, D., & Plessner, H. (2006). The psychological and performance demands of association football refereeing. *International Journal of Sport Psychology*, *37*, 99–120.

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

Plessner, H., & Betsch, T. (2001). Sequential effects in important referee decisions: The case of penalties in soccer. *Journal of Sport and Exercise Psychology*, *23*, 200–205.

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

Ridder, G., Cramer, J. S., & Hopstaken, P. (1994). Down to ten: Estimating the effect of a red card in soccer. *Journal of the American Statistical Association*, *89*, 1124–1227.