Zeitschrift für Sportpsychologie, 15 (1), 1–11 © Hogrefe Verlag, Göttingen 2008

Gelb oder kein Gelb?

Persönliche Verwarnungen im Fußball als Kalibrierungsproblem

Daniel Memmert, Christian Unkelbach, Julia Ertmer und Michael Rechner

Ruprecht-Karls-Universität Heidelberg

Zusammenfassung. In einer Datenbankanalyse wird gezeigt, dass in der ersten Viertelstunde eines Fußballspiels signifikant weniger persönliche Verwarnungen („gelbe Karten") ausgesprochen werden als im Rest des Spiels. Neben pragmatischen Gründen (z. B. keine Wiederholungsfouls zu Beginn) erklären wir diesen Effekt als Resultat einer notwendigen Kalibrierung der Schiedsrichter. Nach dem Konsistenzmodell (Haubensak, 1992a) für solche Urteile entwickeln Menschen (Schiedsrichterinnen und Schiedsrichter) zu Beginn einer Stimulusserie (des Spiels) eine Urteilsskala und verwenden diese konsistent (über die gesamte Spielzeit). Diese Urteilsskala muss zunächst kalibriert werden: würden leichte Vergehen zu Beginn als „gelbe Karte" klassifiziert, müssten alle nachfolgenden Vergehen gleicher oder größerer Schwere ebenfalls zu gelben Karten führen. Da aber eine Norm herrscht, persönliche Verwarnungen sparsam einzusetzen, können zu Beginn des Spiels nur wenige gelbe Karten vergeben werden; d. h. objektiv sollte für ein Vergehen zu Beginn mit höherer Wahrscheinlichkeit eine gelbe Karte gegeben werden als dies realiter der Fall ist. Diese Vorhersage wurde in einem Experiment getestet und die Implikationen für die Urteilssituation von Schiedsrichterinnen und Schiedsrichtern werden diskutiert.
Schlüsselwörter: Fußball, Schiedrichter-Expertise, Entscheidungsfindung, Urteilsfehler

Yellow card or no yellow card? Soccer cautioning as a calibration problem

Abstract. The analysis of a database revealed that significantly fewer yellow cards are shown in the first 15 min compared to the rest of a soccer game. Alongside obvious pragmatic reasons (e.g., preventing repeated fouls at the beginning of a game), we explain this effect as the outcome of a necessary calibration process. According to the consistency model (Haubensak, 1992a), people (in this case, referees) need to develop an internal judgment scale for such categorical decisions at the beginning (of a game), and then apply it consistently across the whole stimulus set (i.e., the whole game). This scale first has to be calibrated: If a light/moderate misconduct is already classified to the "yellow card" category at the beginning of a game, each successive misconduct of the same or greater severity will also require a yellow card. However, because the norm is to caution sparingly, few yellow cards can be shown at the beginning. Hence, objectively speaking, the probability of cautioning a misconduct with a yellow card should be greater than it actually is in real games. This prediction was tested and confirmed in an experiment, and the implications for referees' judgment situations are discussed.
Key words: soccer, referees' expertise, decision making, judgment bias

Im Achtelfinale der Fußball-Weltmeisterschaft 2006 zwischen Portugal und den Niederlanden gab Schiedsrichter Valentin Ivanov 12 gelbe Karten, von denen 4 weiter zu gelb-roten Karten und damit zu Platzverweisen führten. Diese Häufung von Verwarnungen wurde in der Presse weithin diskutiert und ist ein eindrucksvolles Beispiel, welche entscheidende Rolle Schiedsrichterinnen und Schiedsrichter durch das Erteilen von persönlichen Verwarnungen in fast allen Sportspielen für den Ausgang eines Spiels haben. Seit den Arbeiten aus der Arbeitsgruppe um Ansorge (u. a. Ansorge, Scheer, Laub & Howard, 1978) wird den psychologischen Prozessen, die Schiedsrichterurteilen zugrunde liegen, größere Beachtung geschenkt (im Überblick: Bar-Eli & Raab, 2006; Mascarenhas, O'Hare & Plessner, 2006; Plessner & Haar, 2006). In dem erwähnten Weltmeisterschaftsspiel wurden in nicht-wissenschaftlichen Quellen zahlreiche Erklärungen geliefert, wie es zu insgesamt 16 Verwarnungen in einem Spiel kommen kann und ob dies eine gute oder eine schlechte Leistung des Un-

Für die Datenerhebung im Rahmen beider Studien möchten wir Sebastian Kreitz, Simon Landa, Michael Pabst, Christoph Rutsch und Benjamin Waldecker herzlich danken. Zudem möchten wir zwei anonymen Gutachtern für ihre konstruktiven Hinweise herzlich danken.

DOI: 10.1026/1612-5010.15.1.1

parteiischen gewesen sei. Basierend auf Erkenntnissen aus der allgemeinen Kognitionsforschung stellen wir im vorliegenden Beitrag ein Modell vor, welches die Häufigkeit von persönlichen Verwarnungen („gelben Karten") im Fußball erklären könnte, und damit auch die zahlreichen Verwarnungen im Spiel Portugal gegen die Niederlande.

Idealerweise sollten Schiedsrichterinnen und Schiedsrichter jede Spielsituation unabhängig vom Kontext wahrnehmen und nach dem vorgegebenen Regelwerk beurteilen (vgl. Regelhandbuch vom DFB, 2006; Plessner & Haar, 2006). Beispielsweise sollte die Entscheidung im Fußball, ob ein Foul im Strafraum gepfiffen und mit einem Elfmeter geahndet wird, nicht davon abhängen, ob bereits ein Elfmeter für diese Mannschaft gepfiffen wurde. Dass Schiedsrichterinnen und Schiedsrichter dies nicht tun, sondern ihre Entscheidungen sehr wohl im Kontext des Spiels treffen, ist bereits in vielen anderen Studien evident geworden. Beispielsweise zeigten Plessner und Betsch (2001), dass gleiche Foulszenen im Strafraum völlig unterschiedlich beurteilt werden (Strafstoß vs. kein Foul und kein Strafstoß), je nachdem, ob derselben Mannschaft bereits ein Elfmeter zugesprochen wurde oder nicht. Im ersteren Fall wurde praktisch nie ein zweiter Elfmeter gepfiffen; wurde dagegen der gegnerischen Mannschaft bereits ein Elfmeter zugesprochen, stieg die Häufigkeit der zuerkannten Elfmeter massiv an. Solche Kontexteffekte, dass Szenen im Kontext anders beurteilt werden als isoliert, wurden ebenfalls bei Bewertungen im Kunstturnen berichtet (Ste-Marie & Lee, 1991; Damisch, Mussweiler & Plessner, 2006), bei Schiedsrichterurteilen im Australischen Football (Mohr & Larsen, 1998) und bei Foulentscheidungen im Basketball (Brand, Schmidt & Schneeloch, 2006). Ob solche Entscheidungen, die im Kontext des Spiels bzw. des Wettkampfes im Sinne eines „Game-Management" getroffen werden, sinnvoll und dem Sport zuträglich sind, ist noch nicht geklärt (vgl. Mascarenhas, Collins & Mortimer, 2002; Plessner & Betsch, 2002).

Wir möchten im Folgenden dagegen ein Modell aus der allgemeinen Kognitionsforschung vorstellen, das die unterschiedliche Beurteilung von Spielszenen (Beurteilung im Kontext eines Spiels/Wettkampfes vs. isolierte Beurteilung) als Resultat eines Kalibrierungsprozesses in der Urteilssituation erklärt. Dies geschieht am Beispiel von persönlichen Verwarnungen (gelben Karten) im Fußball. Die Vorhersagen dieses Modells für die Verteilung gelber Karten innerhalb eines Spiels werden in einer ersten Studie anhand einer Datenbankanalyse überprüft. Eine zweite Studie testet sodann die Vorhersagen des Modells bezüglich der Beurteilung von Spielszenen (Gelb oder kein Gelb).

Theoretischer Hintergrund

Die Entscheidung, ob ein Regelverstoß mit einer persönlichen Verwarnung geahndet wird, kann als Kategorisierungsaufgabe verstanden werden. Nach einem wahrgenommenen Regelverstoß müssen die Unparteiischen entscheiden, ob der Verstoß in die Kategorie „gelbe Karte" fällt oder nicht. Ein Modell der kognitiven Psychologie für diese Art von Kategorisierungsaufgaben ist das Konsistenz-Modell (Haubensak, 1992a), welches eine Alternative zum Range-Frequency-Modell von Parducci und Kollegen bildet (Parducci, 1965; Parducci & Wedell, 1986). Nach dem Konsistenz-Modell wird zur Kategorisierung zunächst eine interne Beurteilungsskala benötigt, da die Kategorisierung des Stimulus (des Regelverstoßes) absolut erfolgt und nicht in Bezug zu anderen Stimuli. „Absolut" bedeutet hier, dass ein Kategorienurteil getroffen werden muss, und nicht, ob ein Verstoß leichter oder schwerer als ein anderer Verstoß war. Die anderen Stimuli beeinflussen das Kategorienurteil natürlich, indem sie den Range der Skala abstecken. Die Skala entwickelt und kalibriert sich nach Haubensak (1992a) zu Beginn der Stimuluspräsentation (hier: Regelverstöße) und wird über die gesamte Stimulusserie (die Spielzeit) verwendet. Die zentrale Aussage bezieht sich dann auf die Konsistenz der Entscheidungen bzw. Urteile: „Because absolute judgments are concerned with subjective impressions only, there can be no right or wrong answers. The only criterion judges can use is the internal consistency of their own responses" (Haubensak, 1992a, p. 304).

Die Diskussion, ob das Konsistenz-Modell oder das Range-Frequency-Modell die besseren Vorhersagen für Absolut-Urteile liefert, ist noch nicht beendet. Einerseits existieren Datensätze, die nicht mit dem Konsistenzmodell vereinbar sind (Parducci, 1992); andererseits argumentiert Haubensak (1992b), dass diese Unvereinbarkeiten vor allem für Stimuli auftreten, die außerhalb des erwarteten oder erinnerten Ranges der Stimuli liegen, was beispielsweise durch einen Wechsel des Kontextes geschehen kann (z. B. Härte von Fouls in einem Männer-Fußballspiel im Vergleich zu einem Frauen-Fußballspiel). Solche Range-Verletzungen müssen zu einer Anpassung der Skala führen, welche im Range-Frequency-Modell möglich ist, nicht jedoch im Konsistenzmodell. Da Kontextwechsel und außerhalb des Ranges liegende Stimuli im Sport eher die Ausnahme sind, wird im Folgenden das Konsistenzmodell verwendet.

Auf die Schiedsrichtersituation angewendet bedeutet dies Folgendes: Zu Beginn eines Spiels existiert im besten Falle eine gedächtnisbasierte Skala; eine Skala für das Spiel muss sich erst noch bilden und kalibrieren. Wird nun ein Regelverstoß zu Beginn

des Spiels mit Gelb geahndet, müssten nach dem Modell alle nachfolgenden Verstöße gleicher oder größerer Schwere ebenfalls mit Gelb bestraft werden. Prinzipiell wären nach dieser Annahme zwei Vorhersagen möglich: zum einen könnten in der frühen Phase des Spiels übermäßig viele gelbe Karten gezeigt werden, bis die Skala für das laufende Spiel kalibriert ist. Zum anderen könnten aber auch unterdurchschnittlich wenig gelbe Karten gezeigt werden. Die Richtung des Konsistenzeinflusses müsste sich durch die zufällige Wahl des Stimulus/Vergehens ergeben, der zuerst in die Kategorie „gelbe Karte" klassifiziert wird. Würde ein leichtes Vergehen zu Beginn mit Gelb bestraft, wäre die Erwartung nach dem Konsistenzmodell, dass in diesem Spiel viele gelbe Karten gezeigt werden müssten. Dies steht aber im Konflikt zu der Norm, dass gelbe Karten sparsam (persönliches Gespräch mit Egon Striegl am 23. Mai 2007, DFB-Schiedsrichter-Lehrwart) und maßvoll eingesetzt werden sollten (vgl. DFB, 2006, S. 147). Wird für ein Vergehen eine persönliche Verwarnung ausgesprochen und später im Spiel für ein Vergehen gleicher Schwere nicht, wird dies als ungerecht erlebt (vgl. Spranca, Minsk & Baron, 1991). Umgekehrt kann jedoch für ein Vergehen gleicher Schwere später im Spiel eine Verwarnung ausgesprochen werden. Die Unparteiischen benötigen zu Beginn des Spiels also eine Beobachtungsphase der Stimulusreihe (der Vergehen), um eine interne Skala zu entwickeln und diese zu kalibrieren, bevor sie ein Vergehen in die extreme Kategorie „gelbe Karte" klassifizieren. Die zu frühe Vergabe einer gelben Karte reduziert die Freiheitsgrade dieser Skala, da es zukünftige Entscheidungen determiniert: Alle nachfolgenden Fouls gleicher oder größerer Härte müssen ebenfalls mit gelb geahndet werden.

Zusammenfassend leitet sich aus dem Konsistenzmodell die Notwendigkeit der Kalibrierung einer internen Urteilsskala ab, und da im Rahmen eines Fußballspiels gelbe Karten sparsam eingesetzt werden sollen, werden in dieser Kalibrierungsphase weniger gelbe Karten vergeben als zu erwarten wäre. Unsere erste Studie prüft nun mittels einer Datenbankanalyse der Fußball-Bundesliga Spielzeiten 1997 bis 2003, ob tatsächlich weniger gelbe Karten für Vergehen zu Beginn eines Spiels gegeben wurden.

Studie 1

Die Hypothese, dass zu Beginn eines Spiels weniger gelbe Karten gegeben werden, kann verhältnismäßig einfach überprüft werden. Es existieren extensive Datenbänke über Fußballspiele, in denen gezielt nach „gelben Karten" gesucht werden kann. Zur Überprüfung der Hypothese wurden die 90 Minuten eines Fußballspiels in sechs 15-Minuten lange Blöcke unterteilt. Das Konsistenzmodell impliziert, dass in den ersten 15 Minuten die Kalibrierung der Skala stattfindet. Aufgrund des Rahmens des Fußballspiels sollten in dieser Phase weniger gelbe Karten vergeben werden als im Rest des Spiels.

Methode

Alle Spiele der Spielzeiten 97/98, 98/99, 99/00, 00/01, 01/02 und 02/03 der 1. Fußball-Bundesliga wurden auf gezeigte gelbe Karten mit entsprechender Spielminute hin analysiert. Die Daten stammten aus der Datenbank „40 Jahre Bundesliga" (Level9 Medienproduktion GmbH, 2004), auf der auch die Analysen der ARD Sportschau beruhen. Zur Vereinfachung der Analyse wurde jedes Spiel in 6 Blöcke à 15 Minuten aufgeteilt und die Summe aus allen in einem Block gegebenen gelben Karten gebildet. Gelbe Karten, die in der Nachspielzeit der ersten Halbzeit vergeben wurden, wurden dem dritten Block (31.–45. Minute) zugeschlagen, während die Nachspielzeit (90. + x Minute) nicht in die Analyse mit eingeht, da die Nachspielzeit für die zu testenden Hypothesen ohne Bedeutung ist.

Ergebnisse

Randverteilungen: Die Ergebnisse dieser ersten Analyse sind in Tabelle 1 dargestellt: insgesamt wurden über die sechs Spielzeiten 7555 gelbe Karten gegeben; bei 1836 analysierten Spielen entspricht dies somit einem Schnitt von 4.12 (SD = 1.80) gelben Karten pro Spiel[1]. Auffällig ist zunächst die hohe Konstanz der absoluten Häufigkeit von gelben Karten über die fünf Spielzeiten. Trotz der hohen Anzahl von Freiheitsgraden liefert weder ein Chi-Quadrat-Test noch eine Varianzanalyse einen signifikanten Unterschied zwischen den Spielzeiten. $\chi^2(5) = 4.874$, *ns*; $F(5, 7550) = 0.98$, *ns*. Auch ein Test jeder einzelnen Spielzeit gegen den Mittelwert der restlichen Spielzeiten liefert kein signifikantes Ergebnis, alle Fs $(7554) < 1.70$, ns.

Entscheidend ist jedoch, wie Tabelle 1 ebenfalls zeigt, dass es einen klaren Effekt der Spielabschnitte auf die Häufigkeit der gelben Karten gibt. Da die Nachspielzeit[2] aus rein pragmatischen Gründen von den übrigen Spielabschnitten abweichen muss, geht

[1] Hinsichtlich der Häufigkeit von gelb-roten (vgl. die sieben Zeit-Kategorien in Tabelle 1: 1/17/31/46/60/88/11; N_{gesamt} = 254) und roten Karten (9/16/26/31/37/51/14; N_{gesamt} = 184) zeigt sich jeweils ein klarer linearer Trend, der als Kumulierungseffekt interpretiert werden kann: Je länger ein Spiel läuft, desto mehr gelb-rote und rote Karten werden vergeben.

[2] Bei einer mittleren Nachspielzeit der 2. Hälfte von zwei Minuten entspricht die Häufigkeit der gelben Karten etwa der letzten Viertelstunde (11%).

Tabelle 1. Absolute Häufigkeit von gelben Karten (ohne gelb-rote Karten) für sechs Spielzeiten der Fußball-Bundesliga

	Spielminute							
Saison	1.–15.	16.–30.	31.–45.	46.–60.	61.–75.	76.–90.	90.+	Gesamt
97–98	99	192	197	212	254	247	25	1226
98–99	122	191	234	198	260	253	24	1282
99–00	91	213	253	187	236	241	24	1245
00–91	96	203	248	213	250	243	31	1284
01–02	112	186	250	227	221	281	25	1302
02–03	86	190	215	213	232	240	40	1216
Gesamt	606	1175	1397	1250	1453	1505	169	7555

Anmerkung: Die Kategorie 31.–45. schließt die Nachspielzeit vor der Halbzeit mit ein.

diese Kategorie nicht mit ein, was zu $N = 7386$ gelben Karten führt. Ohne die Nachspielzeit ergibt sich über alle Spielzeiten summiert (unterste Zeile Tabelle 1) ein Chi-Quadrat-Wert für die Häufigkeitsunterschiede zwischen den Spielabschnitten von $\chi^2(6) = 443.57$, $p < .0001$. $N = 7386$. Eine Varianzanalyse dieser Unterschiede liefert das äquivalente Ergebnis, $F(5, 7381) = 150.11$, $p < .0001$. Die Varianzanalyse erlaubt gegenüber dem allgemeinen Chi-Quadrat-Test auch die Testung spezifischer Unterschiedshypothesen, in diesem Falle dem ersten Block (dem Beginn des Spiels) gegen die restliche Spielzeit (zur Verwendung von ANOVA mit kategorialen Daten, siehe Lunney, 1970). Dieser Kontrast liefert einen hochsignifikanten Effekt. $F(1, 7385) = 702.11$. $p < .0001$: In der ersten Viertelstunde wurden pro Block weniger als halb so viele gelbe Karten als im Rest des Spiels gegeben. Wie Tabelle 1 bereits vermuten lässt, ist dieser Effekt ebenfalls für jede einzelne Spielzeit hoch signifikant ($F_{97-98}(1, 1200) = 112.57$; $F_{98-99}(1, 1257) = 69.71$; $F_{99-00}(1, 1220) = 150.16$; $F_{00-91}(1, 1252) = 143.51$; $F_{01-02}(1, 1276) = 99.43$; $F_{02-03}(1, 1175) = 151.67$; alle $ps < .0001$).

Binnenverteilung: In einem zweiten Analyseschritt wurde die Häufigkeit von gelben Karten zwischen den einzelnen Spielabschnitten in Abhängigkeit der davor getroffenen Entscheidungen geprüft. Dazu wurden drei Kategorien verglichen: Nach 30 Minuten wurde die erste gelbe Karte gezeigt ($M = 3.35$; $SD = 1.69$), nach 15 Minuten wurde die erste gelbe Karte gezeigt ($M = 4.79$; $SD = 1.70$) und innerhalb der ersten 15 Minuten wurde die erste gelbe Karte gezeigt ($M = 4.47$; $SD = 1.68$). Eine Varianzanalyse dieser Unterschiede liefert einen hochsignifikanten Effekt, $F(2, 1833) = 125.65$, $p < .0001$. Der lineare Kontrast zwischen der ersten und letzten Kategorie ist ebenfalls hochsignifikant, $F(2, 1834) = 210.87$, $p < .0001$: Die Anzahl der gegebenen gelben Karten sinkt in einem Spiel, je länger das Spiel andauert, ohne dass eine gelbe Karte verteilt wurde.

Schiedsrichtereffekte: In einem dritten Schritt wurde schließlich der Einfluss der eingesetzten Schiedsrichter auf das Zustandekommen der Häufigkeit der gelben Karten in den ausgewählten Spielzeiten analysiert. Insgesamt wurden 31 Schiedsrichter in den 1836 Partien eingesetzt. Fünf dieser Schiedsrichter, die weniger als 15 Partien leiteten, wurden aus der Analyse ausgeschlossen, was zu einem Ausschluss von 28 Partien führte (verbleibendes $n = 1808$; $M = 4.11$, $SD = 1.80$). Die verbleibenden 26 Schiedsrichter zeigten wenig Varianz in der mittleren Häufigkeit von gelben Karten ($M_{max} = 4.74$ vs. $M_{min} = 3.73$). Für alle 325 paarweisen Vergleiche unterschieden sich nach Tukey nur acht Schiedsrichterpaare signifikant voneinander in der Häufigkeit der Gabe von gelben Karten, nach Bonferroni fünf und nach Scheffé, dem konservativsten Test, kein einziges Paar. Tabelle 2 zeigt nun die mittlere Häufigkeit von gelben Karten für jeden der verbleibenden 26 Schiedsrichter in Abhängigkeit davon, ob sie in der ersten Viertelstunde keine oder eine bzw. mehrere gelbe Karten verteilt haben. Wie Tabelle 2 klar zeigt, werden in Spielen weniger gelbe Karten gezeigt, wenn in der ersten Viertelstunde keine gelbe Karte gezeigt wurde ($M = 3.87$, $SD = 1.78$) als wenn eine oder mehrere gelbe Karten gezeigt wurden ($M = 4.78$, $SD = 1.69$). Dieser Unterschied ist für 50% der Schiedsrichter signifikant; gleichsam zeigt sich kein einziges Mal eine Umkehrung des Effektes.

Diskussion

Natürlich gibt es eine Reihe von rein pragmatischen Gründen, warum zu Beginn eines Spiels weniger gelbe Karten gegeben werden als im Rest des Spiels: die Mannschaften tasten sich gegenseitig erst ab, Spieler sind noch nicht richtig „warm" oder müssen erst „Ins-Spiel-Kommen" und das Aggressionspotenzial mag sich erst während längeren Spielens stei-

Tabelle 2. Häufigkeiten der Gabe von gelben Karten als Funktion für das Aussprechen oder Nicht-Aussprechen von gelben Karten in den ersten 15 Minuten sowie der eingesetzten Schiedsrichter (n = 26) in sechs Spielzeiten der Fußball-Bundesliga

	Anzahl der geleiteten Spiele	Keine gelbe Karte in der 1. Viertelstunde			Eine oder mehr gelbe Karten in der 1. Viertelstunde				
Schiedsrichter	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*F*	*p*
Albrecht	79	3.96	2.09	51 (65%)	4.89	2.03	28 (35%)	3.68	<.06
Aust	89	4.27	1.39	62 (70%)	5.33	1.39	27 (30%)	10.90	<.001
Berg	75	3.35	1.56	48 (64%)	4.41	1.58	27 (36%)	7.80	<.007
Buchhart	29	3.57	1.47	23 (79%)	5.67	0.82	6 (21%)	11.13	<.003
Dardenne	34	4.22	1.93	23 (68%)	5.82	1.72	11 (32%)	5.47	<.03
Fandel	109	4.18	1.95	72 (66%)	4.73	1.56	37 (34%)	2.22	<.14
Fleischer	71	3.76	1.10	46 (65%)	4.60	1.78	25 (35%)	6.04	<.02
Fröhlich	90	3.79	2.00	66 (73%)	4.67	1.47	24 (27%)	3.85	<.053
Gagelman	28	4.54	2.06	26 (93%)	5.00	0.00	2 (7%)	0.10	<.76
Heyneman	69	3.71	1.67	42 (61%)	4.81	1.69	27 (39%)	7.07	<.01
Jansen	77	3.26	1.43	55 (71%)	4.05	1.68	22 (29%)	4.35	<.04
Kemmling	78	4.00	1.73	60 (77%)	4.94	2.24	18 (23%)	3.60	<.06
Keßler	69	4.06	1.89	47 (68%)	5.27	1.91	22 (32%)	6.08	<.02
Koop	55	3.91	1.68	44 (80%)	5.00	1.41	11 (20%)	3.92	<.053
Krug	125	3.52	1.70	90 (72%)	4.66	1.64	35 (28%)	11.40	<.001
Merk	116	3.59	2.08	86 (74%)	4.57	1.55	30 (26%)	5.49	<.02
Meyer	54	3.66	1.67	44 (81%)	5.00	1.94	10 (19%)	4.95	<.03
Sippel	36	3.77	1.72	30 (83%)	3.83	1.33	6 (17%)	0.01	<.93
Stark	81	4.01	1.69	70 (86%)	5.91	1.76	11 (14%)	11.83	<.001
Steinborn	88	3.10	1.70	63 (72%)	4.08	1.71	25 (28%)	5.99	<.02
Strampe	92	4.24	1.80	72 (78%)	4.25	2.00	20 (22%)	0.00	<.98
Wack	85	4.26	1.71	61 (72%)	4.96	1.49	24 (28%)	3.05	<.08
Wagner	74	3.87	1.93	60 (81%)	4.79	1.37	14 (19%)	2.84	<.10
Weber	25	4.58	1.71	19 (76%)	5.17	2.32	6 (24%)	0.46	<.51
Weiner	43	4.26	1.69	34 (79%)	5.11	0.93	9 (21%)	2.06	<.16
Zerr	37	4.58	1.42	26 (70%)	5.10	1.51	11 (30%)	0.98	<.33
Gesamt	1808	3.92	1.72	1320 (73%)	4.87	1.57	488 (27%)	95.59	<.001

Anmerkung: Die Häufigkeit und Prozentzahlen der geleiteten Spiele, in denen jeweils gelbe Karten gegeben oder nicht gegeben wurden, finden sich in der Spalte nach den *M* und *SD* Kennwerten, auf denen diese Kennwerte basieren.

gern. Es steht außer Zweifel, dass solche Ursachen zu dem beobachteten Effektmuster beitragen. Jedoch halten wir es für unwahrscheinlich, dass die beobachtete Verdopplung der gelben Karten nur auf diese Faktoren zurückzuführen ist. Dass die von uns postulierte Kalibrierung für die wenigen gelben Karten zu Beginn eines Spiels verantwortlich ist, lässt sich partiell an dem vorliegenden Datensatz nachweisen. Einige der oben angesprochenen pragmatischen Gründe (Spieler sind noch nicht richtig „warm" nach der Pause oder müssen erst wieder „Ins-Spiel-Kommen") sollten nämlich auch nach Wiederaufnahme des Spiels in der zweiten Halbzeit, d. h. in der Kategorie 46. bis 60. Minute gelten. Und tatsächlich findet sich ein signifikanter Abfall der Häufigkeit in dieser Kategorie im Vergleich zur letzten Viertelstunde der ersten Halbzeit (vgl. Tabelle 1), $F(1, 7385) = 8.17$, $p < .01$ sowie zur mittleren Viertelstunde der zweiten Halbzeit. $F(1, 7385) = 15.28$, $p < .001$. Dennoch werden auch in dieser Viertelstunde mehr als doppelt so viele gelbe Karten gezeigt wie in der Anfangsviertelstunde.

Die konditionale Analyse der Daten hat unsere Kalibrierungshypothese weiter empirisch unterstützt. Wenn Schiedsrichter im 1. Spielabschnitt gelbe Karten vergeben haben, steigt die mittlere Häufigkeit von gelben Karten im gesamten Spiel. Umgekehrt sinkt die Häufigkeit von gelben Karten im Spiel, je länger keine gelben Karten im Spiel verteilt werden. Allerdings ist dieser Befund offen für alternative Interpretationen; beispielsweise könnte eine gelbe Karte in der ersten Viertelstunde ein Hinweis für ein sehr hartes Spiel sein, was wiederum automatisch zu mehr Verwarnungen führen sollte.

Für die Prozedur spricht, dass die a priori Wahrscheinlichkeit für eine gelbe Karte steigt, je länger ein Schiedsrichter kein Gelb gezeigt hat. Trotz dieses unserer Hypothese gegenläufigen Prozesses passt der beobachtete Effekt zur Kalibrierungshypothese: Wird früh eine gelbe Karte gegeben, wird die Kategorie „gelbe Karte" breiter, was automatisch zu einer erhöhten Häufigkeit von gelben Karten führt (mit den diskutierten Einschränkungen). Klar widerlegt wird von diesen Daten, dass Schiedsrichter in der ersten Viertelstunde eine Verwarnung aussprechen, um die Härte aus dem Spiel zu nehmen und insgesamt weniger Verwarnung aussprechen zu müssen. Frühe gelbe Karten führen insgesamt zu mehr gelben Karten.

Die weitere Analyse der Schiedsrichter zeigt, dass der Faktor der Person oder des individuellen Stils keinen systematischen Einfluss auf die Häufigkeit der gelben Karten zu haben scheint. Aktuelle Arbeiten von Dawson (in press) zeigen ebenfalls nur unsystematische Unterschiede in der Vergabe von Verwarnungen verschiedener Schiedsrichter aus der Premier League. Methodisch gesehen muss bei dieser Auswertungsprozedur kritisch angemerkt werden, dass der DFB die Schiedsrichter den Spielen systematisch und nicht zufällig zuteilt. Als gut bewertete Schiedsrichter pfeifen in der Regel die schweren Spiele (z. B. Dortmund gegen Schalke).

Eine weitere triviale Erklärungsmöglichkeit, welches das dargestellte Ergebnismuster (weniger gelbe Karten zu Beginn eines Spiels) bedingen könnte, kann mit den Resultaten der Datenbankanalyse nicht widerlegt werden. Zu Beginn eines Spiels werden nicht immer direkt gelbe Karten gezeigt, sondern zunächst mündliche Verwarnungen an die regelwidrig handelnden Spieler ausgesprochen. Dies kann der Grund für die geringere Anzahl an gelben Karten in der ersten Spielviertelstunde sein, da erst bei wiederholten, also kumulierten Regelverstößen eine gelbe Karte gezeigt wird. Zum anderen zeigen die Daten einen Anstieg in der letzten halben Stunde, der, obwohl nicht vorhergesagt, sehr gut in das Schema einer Kumulierung passt. Mit anderen Worten, gegen Ende des Spiels findet eine Zunahme von gelben Karten aufgrund wiederholter Verstöße statt. Diese Kumulationserklärung spielt mit großer Sicherheit eine Rolle für das beobachtete Muster.

Das Problem der gesamten Datenbankanalyse besteht natürlich darin, dass keine Basisrate der kritischen Ereignisse (d. h. tatsächlich gelbwürdige Fouls) vorliegt und alle Auswertungen und Ergebnisse offen für alternative Erklärungen sind. Soweit konnten wir nur zeigen, dass die Daten der Kalibrierungshypothese nicht widersprechen. Stärkere Evidenz wird im Folgenden in Studie 2 geliefert.

Studie 2

Um die Kalibrierungshypothese als Grund für die geringe Anzahl von gelben Karten in der Anfangsviertelstunde unabhängig von Kumulierung und anderen Einflüssen zu zeigen, wurde ein quasiexperimentelles Untersuchungsdesign gewählt. Der Grundgedanke kann an einem klassischen Kalibrierungsexperiment illustriert werden: Probanden müssen dabei beispielsweise Rechtecke nach ihrer Größe den Kategorien „klein", „mittel" oder „groß" zuordnen. Wenn ihnen nun das erste Recheck gezeigt wird, haben sie zunächst keine Hinweise darauf wie das Kategoriensystem zu verwenden ist, beispielsweise ob die nachfolgenden Rechtecke alle kleiner oder größer sein werden. Da die Probanden bei ihren ersten Kategorisierungen noch nicht kalibriert sind, werden sie sich nach beiden Richtungen „verschätzen", z. B. mittlere Rechtecke als groß oder kleine als mittlere bewerten. Schiedrichter befinden sich nun in derselben Situation wie Probanden zu Beginn des Versuchs. Bei der Beurteilung eines Fouls als „leicht", „mittel" oder „schwer" (d. h. Verwarnung mit gelber Karte), wissen sie nicht, wie die nachfolgenden Fouls aussehen werden und wie das Kategoriensystem einzusetzen ist. Da gelbe Karten zudem die Freiheitsgrade ihrer Entscheidungen einschränken, sollten sie zu Beginn keine gelben Karten vergeben.

Bei unserem Experiment sollten Schiedsrichter Foulszenen nun isoliert beurteilen, d. h. aus dem Kontext des Spiels herausgelöst. Ohne diesen Kontext spielen Kalibrierung und Freiheitsgrade keine Rolle, da die Entscheidungen unabhängig voneinander sind und eine gegebene gelbe Karte keine Konsequenzen für zukünftige Entscheidungen hat. Die Entscheidungen der Schiedsrichter im Experiment können somit als die regelbasierte Norm gesehen werden.

Bei den zu beurteilenden Szenen wurden zwei Variablen orthogonal manipuliert. Zum einen wurden Szenen ausgewählt, bei denen tatsächlich eine gelbe Karte bzw. keine Karte gegeben wurde. Zum anderen wurden gleich verteilt Szenen aus der ersten oder der vorletzten Viertelstunde ausgesucht.

Ausgehend von der Kalibrierungsidee ist die Vorhersage, dass die Schiedsrichter in einer kontextfreien Bedingung für Foulszenen aus der 1. bis 15. Minute im Gegensatz zur 60. bis 75. Minute mit höherer Wahrscheinlichkeit von den tatsächlich im Spiel getroffenen Entscheidungen abweichen. Wie oben angedeutet, kann sich dieser Effekt in beide Richtungen zeigen, d. h. zu einem frühen Zeitpunkt unterscheiden sich die teilnehmenden Schiedsrichter von den tatsächlichen Urteilen im Spiel sowohl bei der Vergabe als auch bei Nicht-Vergabe von gelben Karten. Spezifischer erwarten wir, dass die Schiedsrichter im Expe-

riment besonders in der ersten Viertelstunde mehr gelbe Karten zeigen als die Schiedsrichter im tatsächlichen Spiel. Begründet werden kann dies dadurch, dass ohne spezifisches Wissen über das aktuelle Spiel oder die Spielminute, in der das ausgewählte Foul passiert, die teilnehmenden Schiedsrichter in unserem Experiment ihre Entscheidungen regelbasiert treffen sollten, ohne auf die Einschränkung ihrer Freiheitsgrade für folgende Entscheidungen Rücksicht nehmen zu müssen. Schiedsrichter im aktuellen Spiel befinden sich dagegen noch in einer Kalibrierungsphase und würden durch eine zu frühe Vergabe einer gelben Karte ihre Freiheitsgrade für zukünftige Entscheidungen einschränken.

Damit würde die Kumulationshypothese widerlegt werden, die eine gegenläufige Ergebniserwartung nahe legt, nämlich geringe Abweichungen zwischen den tatsächlichen Entscheidungen und den Urteilen der teilnehmenden Schiedsrichter in der ersten Viertelstunde, größere gegen Ende des Spiels. Da in der isolierten Beurteilung einer Szene die vorangegangen Fouls keine Rolle spielen können, dürfen hier in der kontextfreien Situation keine Informationen wie beispielsweise das Aussprechen von Verwarnungen berücksichtigt werden.

Methode

Teilnehmer

Es nahmen 17 männliche Schiedsrichter (Alter: $M = 30.06$; $SD = 7.75$) freiwillig im Rahmen eines Schiedsrichterlehrgangs teil. Sie hatten im Durchschnitt seit 11.29 Jahren ($SD = 5.22$) die Schiedsrichterlizenz. Ihr höchstgeleitetes Spiel hatten zwei in der Kreisliga, sieben in der Landesliga, sieben in der Verbandsliga und einer in der Oberliga.

Material und Design

Aus 300 Spielen der Fußball-Bundesliga in der Spielzeit 1995/1996 sowie der Regionalliga aus der Spielzeit 2001/2002 wurden 120 Foulspiele ausgesucht, die nach Beurteilung von Fußball-Experten als gelbwürdig eingeschätzt wurden. Diese früheren Spielzeiten wurden ausgewählt, um zum einen die Wahrscheinlichkeit zu minimieren, dass Schiedsrichter die Spiele kennen. Zum anderen musste sichergestellt werden, dass in der Folgezeit keine gravierenden Regeländerungen stattgefunden hatten. Tatsächlich kannte keiner der Teilnehmer die ausgewählten Spielsequenzen.

Wichtig war, dass ausschließlich Szenen gewählt wurden, die aus den ersten 15 Minuten sowie zwischen der 60. und der 75. Spielminute stammten. Bei beiden Zeitintervallen wurden zur Hälfte auch Situationen ausgewählt, bei denen keine gelben Karten gezogen wurden. Aus allen selektierten Spielsequenzen wurde von einem ehemaligen Bundesligaschiedsrichter und derzeitigem Lehrwart eines Landes-Fußballverbandes eine weitere Auswahl getroffen. Diese Person kannte die Fragestellung der Studie nicht. Es wurden diejenigen Szenen ausgewählt, die nach dem Regelwerk (vgl. DFB, 2006) *nicht* eindeutig und unmissverständlich den Kategorien gelbe Karte bzw. keine gelbe Karte zuzuordnen sind (z. B. keine Grätschen von hinten; am Trikot halten; nach Unterbrechung Ball wegschlagen, etc.). Dabei war dem Bundesligaschiedsrichter nicht bekannt, in welcher Spielzeit die Szenen stattfanden. Dieser Selektionsprozess wurde gestoppt, als jeweils 15 Szenen zu den vier Kategorien ausgewählt waren: Die Kategorie *früh – kein gelb* beinhaltete Szenen der Anfangsviertelstunde, bei welchen keine gelben Karten gegeben wurde. Die Kategorie *früh – gelb* umfasste Sequenzen, bei denen der Schiedsrichter in den ersten 15 Minuten eine gelbe Karte gegeben hatte. Die anderen beiden Kategorien folgten dem gleichen Schema, wobei alle Foulspiele zwischen der 60. und 75. Minute stattgefunden hatten (spät – kein gelbe Karte; spät – gelbe Karte).

Ablauf

Die Schiedsrichter bewerteten die Szenen im Rahmen eines Ausbildungslehrgangs. Zur Einführung wurde erklärt, dass sie an einer psychologischen Studie teilnehmen würden, um Faktoren bei der Vergabe von gelben Karten zu ermitteln. Zur Beurteilung der Szenen wurde ein Fragebogen ausgegeben, eingeleitet durch die dazu nötige Erklärung des Antwortformats auf der ersten Seite. Zudem wurde diesen Schiedsrichtern die Anweisung gegeben, bei jeder Szene zu entscheiden, ob sie eine gelbe Karte zeigen würden oder nicht. Die Szenen wurden digitalisiert über einen Videoprojektor in randomisierter Reihenfolge dargeboten, d. h. die Schiedsrichter wussten bei den Szenen nicht, zu welchem Zeitpunkt im Spiel das jeweilige Foul stattgefunden hatte. Absprachen unter den Teilnehmern waren nicht möglich. Nachdem die Teilnehmer die 60 Szenen bewertet hatten, sammelten die Versuchsleiter die Bögen ein und klärten die Teilnehmer über das vermutete Kalibrierungsmodell auf.

Ergebnisse

Vier Szenen wurden aus den nachfolgenden Analysen ausgeschlossen, da sich alle 17 Schiedsrichter bei ihnen einig waren, keine gelbe Karte zu geben. Aber

auch bei den restlichen Szenen zeigte sich eine hohe Übereinstimmung: Über alle Szenen hinweg ergab sich ein Cronbach's α von .953; mit anderen Worten, obwohl die Anweisung an den Obmann gelautet hatte, Szenen auszuwählen, bei denen die Entscheidung „Gelb" oder „kein Gelb" nicht eindeutig sein sollte, waren sich die Schiedsrichter in allen Szenen in ihren Entscheidungen einig.

Von größerem Interesse sind natürlich die Gelb-Entscheidung in jeder der vier Kategorien, die sich aus der orthogonalen Kombination von Spielzeit (früh vs. spät) und tatsächlicher Entscheidungen (Gelb vs. kein Gelb) ergeben. Für jeden Schiedsrichter wurde der Anteil Gelb-Entscheidungen in jeder Kategorie als Abweichung von der im tatsächlichen Spiel getroffenen Entscheidung berechnet. Dieser Abweichungsindex wurde über alle 17 Schiedsrichter gemittelt und ist in Tabelle 3 als Funktion der tatsächlichen Spielzeit und der tatsächlichen Entscheidungen im Spiel dargestellt. Dieser Prozentsatz kann auch als Grad der Wahrscheinlichkeit interpretiert werden, von der Entscheidung des Schiedsrichters im tatsächlichen Spiel abzuweichen.

Tabelle 3. Abweichung der Gelb-Entscheidungen (in Prozent) der 17 Schiedsrichter aus Studie 2 von den tatsächlichen Entscheidungen als Funktion der tatsächlichen Spielminute und der tatsächlichen Entscheidung im Spiel (Standardabweichungen in Klammern)

	Spielminute	
	1.–15.	**60.–75.**
gelbe Karte	26.3 (10.1)	18.9 (10.7)
keine gelbe Karte	39.4 (16.3)	34.8 (11.1)

Tabelle 3 zeigt nun zwei klare Haupteffekte: Wie vorhergesagt, wichen die Schiedsrichter in *frühen* Spielszenen mit einer mittleren Wahrscheinlichkeit von 32.8% (SD = 9.5%) von den tatsächlichen Entscheidungen ab, während dieser Wert für späte Spielszenen auf 26.9% (SD = 8.3%) sank, $F(1, 16) = 4.51$, $p < .05$. Je länger ein Spiel dauerte, umso mehr glichen die im Spiel getroffenen Entscheidungen den Entscheidungen unserer 17 Schiedsrichter, obwohl letztere keinen Hinweis auf die tatsächliche Spielzeit hatten. Dies spricht für die Kalibrierungshypothese und gegen die Idee einer Kumulation. Neben dem Haupteffekt für die Spielzeit gibt es zum anderen im Mittel größere Abweichung von den im Spiel getroffenen Entscheidungen für *nicht* gezeigte gelbe Karten (M = 37.1%, SD = 13.9%) als für tatsächlich gezeigte gelbe Karten (M = 22.6%, SD = 10.9%), $F(1, 16) = 20.27$, $p < .001$. Mit anderen Worten, unsere 17 Schiedsrichter waren strenger als die Schiedsrichter im Spiel. Dieser Effekt erklärt sich vor allem daraus, dass die Schiedsrichter im Experiment gelbe Karten ohne Rücksicht auf zukünftige Entscheidungen vergeben konnten. Die Interaktion von Spielminute und tatsächlicher Entscheidung war nicht signifikant, $F(1, 16) = 0.30$, *ns*.

Unsere spezifischere Hypothese war jedoch, dass besonders in der ersten Viertelstunde mehr gelbe Karten vergeben werden sollten, da Kalibrierung und Freiheitsgrade im Experiment keine Rolle spielten. Um diese Hypothese angemessen zu testen, muss man bedenken, dass die Wahscheinlichkeit gegen Null geht, dass alle teilnehmenden Schiedsrichter bei allen tatsächlichen nicht-gelb würdigen Szenen keine gelbe Karte geben bzw. bei allen gegebenen gelben Karten auch auf eine gelbe Karte entschieden hätten. Es muss immer Abweichungen von der im Spiel getroffen Entscheidung geben. Die oben verwendete ANOVA zeigt zwar den Haupteffekt für Spielzeit, jedoch nur unter der Annahme, dass diese zufälligen Abweichungen für alle vier Zellen gleich sind. Ein angemessener Test ist die binominale Wahrscheinlichkeit, dass die Abweichung von der im Spiel getroffenen Entscheidung zufällig oder systematisch ist. Ein signifikantes Ergebnis dieses Tests bedeutet, dass die Entscheidungen systematisch mit den im Spiel getroffenen Entscheidungen übereinstimmen, während umgekehrt ein nicht-signifikantes Ergebnis eine Nichtübereinstimmung ausdrückt.

Für diese Analyse wurden zunächst für alle vier Kategorien (frühe/späte Spielzeit: gelb/nicht-gelb) für alle 17 Schiedsrichter die Wahrscheinlichkeit berechnet, ob das Antwortmuster zufällig oder systematisch ist (unter der Annahme, dass p = .5). Diese Wahrscheinlichkeit wurde dann über alle Schiedsrichter hinweg gemittelt. Die entsprechenden mittleren Wahrscheinlichkeiten waren für eine frühe Spielzeit „keine gelbe Karte" (binominal $p < .157$) und für „gelbe Karte" (binominal $p < .045$) sowie für eine späte Spielzeit „keine gelbe Karte" (binominal $p < .052$) und für eine „gelbe Karte" (binominal $p < .038$). Dies zeigt, dass die drei Kategorien („spät/gelbe Karte", „spät/keine gelbe Karte", „früh/gelbe Karte") systematisch mit den Entscheidungen im Spiel übereinstimmen und die Abweichungen als Zufall zu sehen sind (mit Einschränkung für die Kategorie „spät keine gelbe Karte"). Einzig die Kategorie „früh/keine gelbe Karte" zeigt keine signifikante Übereinstimmung mit den Entscheidungen im realen Spiel. Damit stützen die Daten die Kalibrierungshypothese und sprechen gegen eine Kumulierungshypothese, da diese vor allem Abweichungen für die späten Szenen vorhersagen würde.

Diskussion

Zusammenfassend weisen die Daten die vorhergesagte größere Abweichung in der ersten Viertelstunde für gezeigte und nicht-gezeigte gelbe Karten auf. Spezifischer zeigt sich, wenn man für die zufälligen Abweichungen innerhalb einer Kategorie mittels eines Binominaltests („gelb vs. kein gelb") korrigiert, dass die Schiedsrichter im Experiment vor allem nicht mit den „früh/kein gelb" – Entscheidungen übereinstimmen. Obwohl die Schiedsrichter kein Wissen darüber besaßen, zu welchem Zeitpunkt (früh/spät) die Szenen aus dem Spiel waren, ergab sich dennoch ein klarer Effekt der Spielzeit auf das Entscheidungsverhalten. Unsere teilnehmenden Schiedsrichter wichen bei der Beurteilung der Foulspiele, die in den ersten 15 Minuten stattgefunden haben, deutlich von den tatsächlichen Bewertungen der Schiedsrichter im Spiel ab. Da die Schiedsrichter in unserer Studie alle Szenen hoch konsistent beurteilten, zeigt dieser Effekt die fehlende Kalibrierung der Schiedsrichter in den ersten 15 Minuten des tatsächlichen Spiels. In den späteren Situationen, unabhängig ob gelbe Karten gezeigt wurden oder nicht, konvergierten die Urteile stärker mit den Entscheidungen im tatsächlichen Spiel. Dadurch scheidet die Kumulations-Hypothese als ein Erklärungsansatz für das Resultatsmuster aus. Wären die Schiedsrichter von Anfang an kalibriert und die Kumulationshypothese alleine für den Effekt aus Studie 1 verantwortlich, sollte sich ein umgekehrtes Muster ergeben: Höhere Übereinstimmung zu Beginn des Spiels und geringere Übereinstimmung zum Ende hin, da im tatsächlichen Spiel nun kumulierte Karten vergeben würden, die unsere teilnehmenden Schiedsrichter in der kontextfreien Beurteilung nicht vergeben würden und könnten, da für sie keine Kumulierung von Fouls über die Szenen hinweg stattfand. Gleichzeitig findet sich der Effekt, dass die Abweichung von den tatsächlichen Entscheidungen für die *nicht* gezeigten Karten stärker ist, was unsere Annahme der Datenbankanalyse stützt, dass eine Norm herrscht, gelbe Karten sparsam einzusetzen. Die Schiedsrichter in unserer Studie unterlagen dieser Norm nicht, sondern konnten jede Szene isoliert beurteilen. Man könnte nun argumentieren, dass die Schiedsrichter in unserer Studie keine absolute, gedächtnis- und regelbasierte Skala verwendeten, sondern sich zunächst auch an die Szenen kalibrieren mussten, und möglicherweise dadurch das Effektmuster zustande kommt. Dagegen sprechen jedoch ein theoretisches und ein empirisches Argument: zunächst kann man nach dem Konsistenzmodell jede neue Spielszene kontextfrei als eine neue Stimulusserie betrachten, womit die Kalibrierung entfallen würde. Betrachtet man die 60 Szenen aber als eine Serie, könnte sich jedoch ein Kalibrierungseffekt ergeben. Dass dem nicht so ist, zeigen unsere Daten. In den ersten 15 Szenen der Zufallsabfolge werden mehr gelbe Karten gegeben (162) als in den verbleibenden 15er Blöcken (119, 134 und 141). Und die Unterschiede zwischen diesen Blöcken sind statistisch nicht signifikant, $\chi^2(3) = 6.89$, $p > .10$. Somit können wir mit einiger Sicherheit feststellen, dass die notwendige Kalibrierung zu Beginn des Spiels zu weniger gelben Karten in den ersten 15 Minuten führt. Die tatsächliche Größe dieses Einflusses im Vergleich zu anderen Faktoren zu ermitteln, muss in zukünftigen Studien untersucht und herausgefunden werden.

Abschließende Diskussion

Mit einer Datenbankanalyse (Fußball-Bundesliga Spielzeiten 1997–2003) konnte gezeigt werden, dass in der ersten Viertelstunde eines Fußballspiels deutlich weniger persönliche Verwarnungen ausgesprochen werden als in der restlichen Spielzeit. Neben pragmatischen Gründen (z. B. keine Kumulation, bzw. keine Wiederholungsfouls zu Beginn) kann dieser Effekt als das Resultat einer notwendigen Kalibrierung in der Urteilssituation der Schiedsrichter erklärt werden. Nach dem Konsistenzmodell (Haubensak, 1992a) für solche Urteile entwickeln Schiedsrichterinnen und Schiedsrichter zu Beginn eines Spiels eine Urteilsskala und verwenden diese konsistent über die gesamte Spielzeit. In einem quasi-experimentellen Design konnten Kumulationseffekte als alleiniger Grund für die geringe Anzahl gelber Karten in der Anfangsviertelstunde weitestgehend ausgeschlossen werden. Vielmehr deuten die Daten auf eine notwendige interne Kalibrierung der Schiedsrichter zu Spielbeginn hin. Dass bedeutet, dass die Unparteiischen zu Beginn eines Spiels eine Urteilsskala aufbauen und diese konsistent verwenden. Werden also leichte Vergehen zu Beginn in die Kategorie „gelbe Karte" klassifiziert, müssten alle nachfolgenden Vergehen ebenfalls zu gelben Karten führen. Da aber gleichzeitig eine Norm herrscht, persönliche Verwarnungen sparsam einzusetzen, können zu Beginn des Spiels nur wenige gelbe Karten vergeben werden.

Die Ergebnisse der Datenbankanalyse sowie der vorgestellten quasiexperimentellen Studie wurden bereits mit Folgeexperimenten abgesichert. Dazu wurde einer Gruppe von Unparteiischen Foulsituationen in chronologischer Reihenfolge des Spiels gezeigt. Den anderen Probanden wurden dieselben Szenen in zufälliger Anordnung präsentiert. Auch hier deuten die Resultatsmuster auf einen Kalibrierungseffekt hin, weil die erstgenannten Schiedsrichter in einer frühen Spielzeit im Gegensatz zu der zweiten Schiedsrichtergruppe weniger oft auf gelbe Karten entschieden haben (Unkelbach & Memmert, 2007).

Unklar bleibt nach wie vor, ob Schiedsrichterinnen und Schiedsrichter bewusst oder unbewusst in den ersten Minuten eines Spiels bemüht sind, eine Urteilsskala zu entwickeln, d. h. ein Gefühl für das Spiel zu bekommen. Andererseits könnte unabhängig von der Kalibrierungsproblematik eine breitere Regelauslegung – nicht sofort eine gelbe Karte zu geben, obwohl sie nach den Regeln angebracht wäre – für die gesamte Leitung des Spiels sinnvoll sein. Für diese Fähigkeit, ein Spiel in seiner Gesamtheit zu leiten, haben Mascarenhas, Collins und Mortimer (2002) den Begriff des „Game-Management" eingeführt. Damit meinen sie, dass Schiedsrichter in der Lage sein müssten, ein Gefühl für das Spiel zu haben um dieses flüssig und gerecht zu leiten. Während die Kalibrierung ein eher unbewusster Prozess ist, würde eine bewusste Zurückhaltung mit persönlichen Verwarnungen zu Beginn auch im Sinne eines Game-Managements erklärbar sein, da die Unparteiischen zu Beginn den Spielfluss noch nicht mit harten Strafen unterbrechen wollen. Mittlerweile haben Brand, Schmidt und Schneeloch (2006) erste empirische Belege vorgelegt, die zeigen, dass Basketball-Schiedsrichter situationsabhängig im Sinne eines Game-Management Foulentscheidungen treffen. Da im vorliegenden Experiment das Game-Management konsequent und bewusst ausgeschlossen wurde, könnte dieses gezielt im Mittelpunkt von Folgestudien stehen.

Mit dem Konsistenzmodell können nun auch die vielen gelben Karten im Achtelfinalspiel der Fußball-Weltmeisterschaft 2006 zwischen Portugal und den Niederlanden erklärt werden. Schiedsrichter Valentin Ivanov verwarnte bereits in der 2. und 7. Minute Mark van Bommel und Khalid Boulahrouz, womit die Kategorie „gelbe Karte" bereits sehr früh verwendet wurde und folglich alle Fouls gleicher (oder schwererer) Härte ebenfalls mit Gelb geahndet werden mussten. Hinzu kam, dass Mitte der ersten Halbzeit ein relativ hartes Foul an Cristiano Ronaldo ebenfalls „nur" mit Gelb geahndet wurde. Damit war die Kategorie „gelbe Karte" unverhältnismäßig breit gefasst, sowohl nach Härte als auch nach Geringfügigkeit des Fouls. Dies hätte sich nur vermeiden lassen, wenn dieses harte Foul direkt mit Rot bestraft worden wäre und sich damit die gesamte Skala (kein Gelb, Gelb, Rot) nach unten verschoben hätte. Valentin Ivanov tat dies nicht und verwendete die sehr breite Kategorie (Gelb für schwache und sehr harte Fouls) konsistent über die gesamte Spielzeit, was im Endeffekt zu 16 gelben Karten (darin enthalten 4 gelb-rote Karten), jedoch keiner einzigen direkt roten Karte führte.

Welche Implikationen haben nun die dargestellten Befunde für die Urteilssituation von Schiedsrichterinnen und Schiedsrichtern? Zunächst muss klar herausgestrichen werden, dass die Unparteiischen in der Art und Weise handeln wie auch andere Menschen, z. B. Deutsch-Lehrer in Bezug auf Deutsch-Aufsätze oder Teilnehmer in einer Studie, bei der die Größe von Vierecken beurteilt werden muss (Parducci, 1965); in solchen Situationen existieren keine eindeutigen „Wenn-dann"-Regeln für die Zuordnung von Stimuli zu Kategorien. Lehrer beispielsweise versuchen durch eine größere Anzahl an Situationen (Korrigieren von mehreren Aufsätzen zu Beginn), ihr eigenes Handeln (Vergabe der Note 2 oder 3) zu kalibrieren. Um diesen Prozess abzukürzen wäre es unbestritten sinnvoll, dass die Schiedsrichter gleich zu Beginn in stärkerem Maße erworbene gedächtnis- bzw. regelbasierte Skalen einsetzten. Dies bedeutet, dass die Unparteiischen ihre „absolute", gedächtnis- und regelbasierte Kalibrierung (vgl. Plessner & Betsch, 2001, 2002) nach dem DFB-Regelbuch *kontextfrei* verwenden sollten. Kontextfrei bedeutet in diesem Zusammenhang, dass bestimmte Fouls immer eine gelbe Karte nach sich ziehen, andere nicht. Ob die Anwendung dieser Maßnahme über geeignete Trainings- und Schulungsprogramme überhaupt vermittelbar ist, werden die Praxis und weitere Treatmentstudien zeigen müssen.

Literatur

Ansorge, C. J., Scheer, J. K., Laub, J. & Howard, J. (1978). Bias in judging women's gymnastics induced by expectations of within-team order. *Research Quarterly, 49*, 399–405.

Bar-Eli, M. & Raab, M. (2006). Judgment and decision making in sport and exercise: Rediscovery and new visions. *Psychology of Sport and Exercise, 7*, 519–524.

Brand, R., Schmidt, G. & Schneeloch, Y. (2006). Sequential effects in elite basketball referees' foul decisions. *Journal of Sport & Exercise Psychology, 28*, 93–99.

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

Dawson, P. (in press). Crime and punishment in professional football: an economic perspective. In P. Andersson, P. Ayton & C. Schmidt (Ed.), *Myths and facts about football: The economics and psychology of the world's greatest sport.*

Deutscher Fußballbund (2006). *Fußball-Regeln 2006/2007* (Schulungsmaterialien). Frankfurt: Deutscher Fußballbund.

Haubensak, G. (1992a). The Consistency Model: A process model for absolute judgments. *Journal of Experimental Psychology: Human Perception and Performance, 18*, 303–309.

Haubensak, G. (1992b). The Consistency Model: A reply to Parducci. *Journal of Experimental Psychology: Human Perception and Performance, 18*, 314–315.

Lunney, G. H. (1970). Using analysis of variance with a dichotomous dependent variable: an empirical study. *Journal of Educational Measurement, 7*, 263–269.

Mascarenhas, D. R. D., Collins, D. & Mortimer, P. (2002). The art of reason versus the exactness of science in elite refereeing: Comments on Plessner and Betsch (2001).

Journal of Sport and Exercise Psychology, 24, 328–333.

Mascarenhas, D. R. D., O'Hare, D. & Plessner, H. (2006). The psychological and performance demands of association soccer refereeing. *International Journal of Sport Psychology, 37*, 1–22.

Mohr, P. B. & Larsen, K. (1998). Ingroup favoritism in umpiring decisions in Australian Football. *The Journal of Social Psychology, 138*, 495–504.

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72*, 407–418.

Parducci, A. (1992). Comment on Haubensak's associative theory of judgment. *Journal of Experimental Psychology: Human Perception & Performance, 18*, 310–313.

Parducci, A. & Wedell, D. H. (1986). The category effect with rating scales: Number of categories, number of stimuli, and method of presentation. *Journal of Experimental Psychology: Human Perception and Performance, 12*, 496–516.

Plessner, H. & Betsch, T. (2001). Sequential effects in important referee decisions. *Journal of Sport & Exercise Psychology, 23*. 254–339.

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 & Exercise Psychology, 24*, 334–337.

Plessner, H. & Haar, T. (2006). Sports performance judgments from a social cognitive perspective. *Psychology of Sport and Exercise, 7*, 555–575.

Spranca, M., Minsk, E. & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology, 27*, 76–105.

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

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

Daniel Memmert

Ruprecht-Karls-Universität Heidelberg
Institut für Sport und Sportwissenschaft
Im Neuenheimer Feld 720
69120 Heidelberg
E-Mail: Daniel.Memmert@urz.uni-heidelberg.de