Sportwiss 2009 · 39:347–358
DOI 10.1007/s12662-009-0086-3

Die Drei-Punkte-Regel in der Fußballbundesliga

Diskussion zum Beitrag

Strauß, B., Hagemann, N. & Loffing, F. (2009). Die Drei-Punkte-Regel in der deutschen 1. Fußballbundesliga und der Anteil unentschiedener Spiele. Eine Replik auf den Beitrag von Dilger und Geyer, 2007. *Sportwissenschaft, 39,* 16–22.

Leserbrief

A. Dilger, H. Geyer
Institut für Ökonomische Bildung und Centrum für Management, Westfälische Wilhelms-Universität Münster

Unentschieden sind nicht unabhängig von der Drei-Punkte-Regel

Strauß, Hagemann und Loffing (2009) kritisieren unseren Beitrag aus *Sport und Gesellschaft* (Dilger & Geyer, 2007), in dem wir aus theoretischen Gründen insbesondere die Hypothese herleiteten und dann mit empirischer Evidenz stützen, dass die Einführung der Drei-Punkte-Regel den Anteil der dadurch unattraktiveren Unentschieden reduziert. Es lassen sich alle ihre Einwände entkräften, da sie insbesondere nicht den korrekten statistischen Test auf Unabhängigkeit und keinerlei Theorie verwenden.

Strauß, Hagemann und Loffing (2009, im Folgenden zitiert als Strauß et al., worauf sich auch alle folgenden Seitenangaben beziehen, wenn nicht anders angegeben) greifen unseren (Dilger & Geyer, 2007) in *Sport und Gesellschaft* veröffentlichten Beitrag in dieser Zeitschrift an, weil sie ihre Replik dort, in *Sport und Gesellschaft*, nicht veröffentlichen konnten. Dies könnte daran liegen, dass ihre Kritik nicht hinreichend überzeugend ist, wie wir im Folgenden zeigen werden. Natürlich sind in der Sache unterschiedliche Einschätzungen möglich, doch „ausdrücklich als Replik" (S. 21) zu verstehende Ausführungen sollten stichhaltige Argumente enthalten, die deutlich stärker sind als die angegriffenen Thesen. Ob dieser Maßstab auch an unsere Antwort auf die vorgebrachten Gegenargumente anzulegen ist und sie diesem ggf. genügt, mag jeder Leser selbst beurteilen.

Ad Gegenargument 1: Geringe praktische Bedeutsamkeit

Strauß et al. werfen uns vor, dass der von uns gefundene Unterschied im Anteil der Unentschieden 10 Jahre vor gegenüber 10 Jahre nach Einführung der Drei-Punkte-Regel von geringer bis gar keiner praktischen Bedeutung sei. Es ist jedoch nicht zu sehen, wie die gewünschten „Verteilungskennwerte wie χ^2-Werte" (S. 18) diese praktische Bedeutsamkeit im Gegensatz zur statistischen Signifikanz, die wir explizit über χ^2-Tests berechneten, nachweisen können sollen. Jedenfalls geben wir entsprechende χ^2-Werte weiter unten in ◘ **Tab. 1** an, während Strauß et al. „zusätzlich das Effektmaß w (Cohen, 1988) zur Abschätzung der praktischen Bedeutsamkeit" anführen (S. 18). Dabei „ergibt sich mit w=0,063 ein sehr kleiner Wert nahe bei Null, der anzeigt, dass sich die Häufigkeiten nicht praktisch bedeutsam unterscheiden" (S. 18 f.).

Nun müssen wir zugeben, dass wir von diesem Maß w und Cohen (1988) noch nie etwas gehört hatten (wobei es sich, was Strauß et al. nicht angeben, um die 2. Auflage handelt, während die 1. Auflage bereits von 1969 stammt) und auch jetzt keine weiteren wissenschaftlichen Arbeiten finden konnten, die dieses Maß verwenden, weshalb dessen praktische Bedeutsamkeit zumindest fraglich erscheint. Das könnte auch erklären, warum Strauß et al. w falsch berechnen, wenn auch zu unseren Gunsten, nämlich mit 0,063 zu hoch. Korrekt berechnet beträgt w 0,039, wie sich sowohl aus Formel (7.2.1) mit 4 Zellen (unterteilt einerseits nach entschiedenen und unentschiedenen Spielen, andererseits vor und nach der Regeländerung) als auch Formel (7.2.5) ergibt (S. 216 bzw. S. 223 jeweils bei Cohen, 1988). Strauß et al. scheinen hingegen ihre unter der Gleichverteilungsannahme gewonnenen χ^2-Werte (siehe dazu unten zum Gegenargument 2) in Formel (7.2.5) einzusetzen (sie erklären nicht explizit ihre Art der Berechnung, doch bei dem genannten Vorgehen ergeben sich die von ihnen angegebenen w-Werte), obwohl diese explizit nur für „2×2 tables" (Cohen, 1988, S. 223) gilt, und außerdem für die Fallzahl n nur die Zahl der Unentschieden, obwohl n „the *total* number of cases in the comparison" (Cohen, 1988, S. 227, mit Hervorhebung im Original) symbolisiert, also einschließlich der entschiedenen Spiele. Unter der Gleichverteilungsannahme wäre richtigerweise Formel (7.2.1) mit m=2 zu verwenden, was 0,046 ergibt.

Damit stellt sich die Frage, ob die noch niedrigeren w-Werte die fehlende Bedeutung des Effekts nachweisen. Cohen (1988, S. 224) räumt selbst ein, dass die Größeneinschätzung „ad hoc" erfolgen sollte und ihr „use requires particular caution" mit „possible inaptness in any given substantive context". In unserem Kontext lässt sich die Größe des Effektes jedenfalls unmittelbar aus der deskriptiven Statistik ablesen: Der Anteil der unentschiedenen Spiele ist von durchschnittlich 29,23% in den 10 Jahren vor Änderung der Bepunktung auf 25,75% in den 10 Jahren danach gefallen. Das entspricht 3,48 Prozentpunkten bzw. einer Reduktion um 11,89%. Das mag man für viel oder wenig halten, es lässt sich auf jeden Fall deutlich leichter beurteilen als w. Wenn der Unentschiedenanteil um die Hälfte gefallen wäre, läge w bei 0,176, was angeblich auch nur ein kleiner Effekt ist (S. 19 unter Verweis auf Cohen

Tab. 1 χ^2-Unabhängigkeitstest für unterschiedliche Untersuchungsintervalle

Intervall	Unentschieden Zwei-Punkte-Regel	Unentschieden Drei-Punkte-Regel	χ^2 (1)	p<0,01	p<0,05	p<0,1
1	28,10%	35,29%	3,653	Ja	Ja	Ja
2	27,61%	29,08%	0,326	Nein	Nein	Nein
3	28,43%	28,65%	0,011	Nein	Nein	Nein
4	29,66%	28,59%	0,347	Nein	Nein	Nein
5	30,61%	28,56%	1,577	Nein	Nein	Nein
6	30,42%	27,56%	3,715	Nein	Nein	Ja
7	30,46%	26,80%	7,147	Ja	Ja	Ja
8	30,10%	26,59%	7,497	Ja	Ja	Ja
9	29,70%	26,25%	8,239	Ja	Ja	Ja
10	29,23%	25,75%	9,382	Ja	Ja	Ja
11	28,92%	26,26%	6,0325	nein	Ja	Ja
12	28,32%	26,23%	4,1131	Nein	Ja	Ja
13	28,04%	26,17%	3,5395	Nein	Nein	Ja
14	27,63%	26,03%	1,6261	Nein	Nein	Nein

1988), während selbst die maximal mögliche Reduktion auf gar kein Unentschieden nach der Regeländerung mit w=0,412 „unterhalb der Grenze für einen" großen Effekt von w=0,500 (ebendort) läge.

Schließlich ist darauf hinzuweisen, dass wir nicht willkürlich irgendwelche Fluktuationen in den Daten suchen, sondern aus guten theoretischen Gründen (s. auch Geyer, 2009) einen Rückgang des Anteils an unentschiedenen Spielen erwarten und diese Hypothese statistisch überprüfen. Im Falle eines solchen Hypothesentests ist jedoch die statistische Signifikanz deutlich wichtiger als die genaue Größe des von null verschiedenen Effekts, was zu Gegenargument 2 führt.

Ad Gegenargument 2: Selektive Auswahl des Beobachtungszeitraums

Strauß et al. behaupten (S. 19): „Es ist inhaltlich nicht zu begründen, und wird von Dilger und Geyer (2007) auch nicht begründet, warum gerade zwei Zehnjahresintervalle (und ausschließlich nur diese), jeweils eines vor und nach der Einführung der Drei-Punkte-Regel, verglichen wurden." Dabei gibt es eine ganz einfache pragmatische Erklärung für die Wahl dieses Zeitraums: Zum Zeitpunkt der Untersuchung lagen Daten von genau 10 Spielzeiten seit der Regeländerung vor, während die Wahl eines symmetrischen Intervalls davor aus statistischen und inhaltlichen (s. unten) Gründen sinnvoll erscheint. Entsprechend werten wir hier jetzt jeweils 14 Saisons davor und danach aus, während die Beschränkung auf 12 Saisons durch Strauß et al. sich nicht auf diese Weise erklären lässt (sie hätten Daten von 13 Saisons danach verwenden können).

Jedenfalls suggerieren Strauß et al. (S. 19 f., insbesondere ◘ **Tab. 1**), dass wir genau dieses Intervall gewählt hätten, weil die von uns behauptete statistische Signifikanz sonst nie auftreten würde. Auch wenn Strauß et al. es nicht explizit äußern, wäre eine bewusste Selektion höchst irreführend bis intellektuell unredlich, zumal eine Auswahl nach der statistischen Signifikanz deren Aussagekraft aufheben würde (sie gilt nur für den einmaligen Test und gibt dafür eine Irrtumswahrscheinlichkeit an, die bei hinreichend vielen Wiederholungen beliebig gesenkt werden kann, so dass potenziell auch ohne irgendeinen tatsächlichen Effekt jedes Signifikanzniveau erreichbar ist). Doch erstens haben wir das Intervall nicht auf diese manipulative Weise ausgewählt, sondern aus dem genannten pragmatischen Grund der Datenverfügbarkeit. Zweitens finden selbst Strauß et al. auf ihre Weise noch 4 weitere Intervalle, die immerhin auf dem 5-Prozent-Niveau signifikant sind (S. 19, Fußnote 7). Drittens verwenden sie nicht den angemessenen Test.

Strauß et al. führen χ^2-Tests auf Gleichverteilung durch, also ob der Anteil der unentschiedenen Spiele vor und nach Einführung der Drei-Punkte-Regel gleich ist. Sie meinen, dass wir denselben Test durchgeführt hätten, obwohl die von uns ausgewiesene Signifikanz von 0,002 deutlich unter ihrem Wert für die zwei Zehnjahresintervalle liegt. Im Übrigen führen sie 3 verschiedene Tests mit unterschiedlichen (bzw. keinen) Adjustierungen durch (S. 18 u. S. 20, wobei sie auf S. 20 auf unser Vorgehen verweisen, obwohl sich für unseren χ^2-Test die Frage nach Adjustierung überhaupt nicht stellt, sondern nur für die von Strauß et al. gar nicht behandelten t-Tests). Dabei erscheint der Verzicht auf Adjustierungen das richtige Vorgehen (S. 18, Fußnote 4), wobei es sich immer noch um einen Test auf Gleichverteilung handelt (entgegen ebendort), jedoch ohne Annahme von genauer Gleichheit der Zahl der Unentschieden vor und nach der Regeländerung, die wegen der abweichenden Anzahl an Spielen in der Saison 1991/1992 eine Ungleichverteilung implizieren würde. Die Adjustierungen erscheinen dagegen willkürlich und verändern über die Fallzahl das Signifikanzniveau, wobei wir die Anpassung im Text (S. 18) und Hauptteil von ◘ **Tab. 1** (S. 20) mit n=1682 nicht nachvollziehen können, den niedrigeren und weniger signifikanten χ^2-Wert mit n=1680 hingegen schon (S. 20, wofür offensichtlich die Zahl der Unentschieden von 1991/1992 mit 306/380 multipliziert wurde).

Ohne irgendwelche Adjustierungsprobleme haben wir den für diese Fragestellung adäquateren χ^2-Test auf Unabhängigkeit statt auf Gleichverteilung durchgeführt, der auch sonst weit gebräuchlicher ist, weshalb wir darauf nicht explizit hinwiesen. Der Unterschied zwischen beiden Tests besteht darin, dass der Test auf Gleichverteilung als Nullhypothese die gleiche Verteilung der Unentschiedenanteile über beide Intervalle annimmt (bei Adjustierung der Spieleanzahl wird der Durchschnittsanteil beider Intervalle zusammen betrachtet), während der Unabhängigkeitstest fragt, ob die Anteile und Intervalle voneinander stochastisch unabhängig sind. Letzterer Test ist strenger, stochastische Unabhängigkeit kann also leichter verworfen werden als die Annahme der Gleichverteilung. Wenn Verteilungen tatsächlich gleich sind, sind sie zwar auch unabhängig voneinander, doch der Test auf Gleichverteilung orientiert sich an den Unterschieden der Inter-

valle vom intervallübergreifenden Durchschnitt, während der Test auf Unabhängigkeit direkt die (bei gleich großen Fallzahlen) doppelt so große Differenz zwischen den Durchschnittswerten der beiden Intervalle betrachtet. Allein letzteres entspricht der vorliegenden Problemstellung und auch der von uns formulierten Hypothese. Zur Beurteilung der Frage, ob sich vor und nach Einführung der Drei-Punkte-Regel der Anteil der unentschiedenen Spiele verändert hat, kommt es auf Unterschiede dieses Anteils vor und nach der Einführung an, nicht auf Abweichungen dieser Durchschnitte in den jeweiligen Intervallen vom intervallübergreifenden Durchschnitt, dem hier keine weitere Bedeutung zukommt. Wir hatten in unserer Hypothese 1 vermutet, dass der Anteil der Unentschieden nach Einführung der Drei-Punkte-Regel geringer ist als vorher, was nicht impliziert, dass er unbedingt auch statistisch signifikant geringer ist als der intervallübergreifende Durchschnittsanteil.

Tabelle 1 weist die Ergebnisse des χ^2-Tests auf Unabhängigkeit für alle 14 gegenwärtig symmetrisch untersuchbaren Doppelintervalle aus. Dabei zeigt sich, dass 5-mal die Nullhypothese der Unabhängigkeit auf einem Signifikanzniveau von 1% verworfen werden kann, 7-mal auf dem 5%-Signifikanzniveau und 9-mal auf dem schwachen Niveau von 10%. Die Richtung der Abhängigkeit widerspricht jedoch bei der kürzesten Intervalllänge, nur einem Jahr vor und nach Änderung der Bepunktung, unserer aus theoretischen Gründen abgeleiteten Vorhersage, da der Anteil der Unentschieden signifikant gestiegen und nicht gesunken ist. Der hohe Anteil an Unentschieden in der Saison 1995/1996 ist auch für die insignifikanten Ergebnisse in den etwas längeren Betrachtungsintervallen verantwortlich. In späteren Saisons ist der Anteil der Unentschieden zwar niedriger, doch der Durchschnitt wird deshalb erst bei Betrachtung von mindestens 4 Saisons nach der Regeländerung niedriger als vorher und (schwach) signifikant so erst nach (6) 7 Saisons.

Bei einer Intervalllänge von 10 Saisons vor und nach der Regeländerung ist tatsächlich (zufällig) die Differenz der Unentschieden am größten, bei noch längeren Intervallen wird sie wieder geringer und ist nach 14 Saisons nicht einmal mehr schwach auf dem 10%-Niveau signifikant. Wirklich überraschend ist es jedoch nicht, dass bei immer weiterer Ausdehnung des Betrachtungszeitraums der Unterschied zwischen den Intervallen vor und nach der Einführung der Drei-Punkte-Regel nicht weiter zunimmt. So ist die Drei-Punkte-Regel doch v. a. deswegen eingeführt worden, weil der Anteil der Unentschieden auf längere Sicht anstieg, was bei Berücksichtigung von vielen Perioden zu einem fallenden Wert im Intervall vor der Regeländerung führen muss, wie in ◘ **Tab. 1** zu sehen ist. Umgekehrt könnte, was bislang eher weniger zu beobachten ist, im Intervall danach der Wert irgendwann wieder ansteigen, wenn die eigentlichen Ursachen für den langfristigen Anstieg fortbestehen, während die Umstellung von der Zweiauf die Drei-Punkte-Regel einer einmaligen Niveauveränderung entspricht. Unentschieden sind unter der Drei-Punkte-Regel stets gleichermaßen weniger attraktiv als unter der Zwei-Punkte-Regel, nicht mit einer zu- oder abnehmenden Rate.

Wirklich erklärungsbedürftig bleibt damit, warum in der Saison 1995/1996 der Anteil der Unentschieden entgegen unserer Hypothese angestiegen statt gefallen ist und das sogar in statistisch signifikanter (von null verschiedener) Höhe. Eine empirische Analyse weiterer Saisons hilft hier offensichtlich nicht weiter, doch auch eine genauere Betrachtung der innersaisonalen Verteilung der Unentschieden führt zu keinem Ergebnis. Am plausibelsten erscheint uns die Übertragung des folgenden Befundes aus vielen spieltheoretischen Experimenten: Reale Akteure verhalten sich nicht immer und sofort vollständig rational, sondern benötigen eine gewisse Zeit des Lernens und Begreifens neuer Spiele bzw. Spielregeln. Die Vereine hatten zwar seit 1995/1996 einen Anreiz, Unentschieden zu meiden und beim Streben nach einem Sieg stärker eine Niederlage zu riskieren (eine 50:50-Chance auf Sieg oder Niederlage ist seither im Erwartungswert einen halben Punkt mehr wert als ein Unentschieden), haben aber am Anfang die Situation nicht unbedingt so gesehen, sondern z. B. dem jeweiligen Gegner 2 Punkte mehr als vorher nur einen durch einen Führungstreffer nicht gegönnt oder ganz rational einen frühen Rückstand im Spiel stärker gefürchtet als unter der Zwei-Punkte-Regel (vgl. Brocas & Carrillo, 2004; s. dagegen Geyer, 2009), doch die Strategie im weiteren Spielverlauf nicht in optimaler Weise auf eine stärkere Offensive umgestellt. Über die genauen Gründe lässt sich nur spekulieren, doch suboptimales Verhalten kurz nach einer Regeländerung ist kein ungewöhnliches Phänomen und widerspricht nicht unserer grundsätzlichen These, dass sich professionelle Fußballmannschaften in einem harten Wettbewerb auf mittlere bis längere Frist den bestehenden Anreizen entsprechend verhalten werden, also konkret den Verlust von einem Punkt für beide spielenden Mannschaften zusammen durch ein Unentschieden möglichst meiden werden. Auf lange Sicht mag das nicht beobachtbar sein, weil dieser Anreiz gegen Unentschieden durch andere Regeländerungen und Entwicklungen im Endergebnis überlagert werden kann. Für den unwahrscheinlichen Fall einer Wiedereinführung der Zwei-Punkte-Regel sagen wir jedoch einen klaren Anstieg des Anteils unentschiedener Spiele voraus, höchstens um ein oder zwei Spielzeiten verzögert, während umgekehrt der wohl noch unwahrscheinlichere Übergang zu einer Vier- oder Fünf-Punkte-Regel oder sogar einer Wertung von Unentschieden mit null Punkten wie eine Niederlage den Anteil der Unentschieden nochmals signifikant reduzieren würde.

Wenn die Argumentation für abweichendes Verhalten in der ersten Saison nach der Regeländerung akzeptiert wird, ist eine ansonsten ad hoc wirkende Herausnahme dieser Saison bzw. des entsprechenden Saisonpaares (also auch der Saison 1994/1995 direkt vor der Regeländerung) vertretbar oder sogar geboten. So zeigt ◘ **Tab. 2** die Anwendung des χ^2-Tests auf Unabhängigkeit für die verschiedenen Untersuchungszeiträume jeweils ohne die letzte Saison vor und erste nach der Regeländerung. Für die in der Spalte Intervall mit „1" bezeichnete Zeile bedeutet dies also, dass die Saison 1993/1994 und die Saison 1996/1997 untersucht werden, in der Zeile „2" zusätzlich die Saisons 1992/1993 und 1997/1998 etc. Die Zahl der signifikanten Intervalle steigt dadurch deutlich gegenüber ◘ **Tab. 1**.

Tab. 2 Unabhängigkeitstest ohne letzte Saison vor und erste nach Regeländerung

Intervall	Unentschieden Zwei-Punkte-Regel	Unentschieden Drei-Punkte-Regel	χ^2 (1)	p<0,01	p<0,05	p<0,1
1	27,12%	22,88%	1,473	Nein	Nein	Nein
2	28,59%	25,33%	1,66	Nein	Nein	Nein
3	30,14%	26,36%	3,36	Nein	Nein	Ja
4	31,20%	26,88%	5,71	Nein	Ja	Ja
5	30,86%	26,01%	9,03	Ja	Ja	Ja
6	30,84%	25,38%	13,776	Ja	Ja	Ja
7	30,37%	25,35%	13,644	Ja	Ja	Ja
8	29,90%	25,12%	14,188	Ja	Ja	Ja
9	29,35%	24,69%	15,341	Ja	Ja	Ja
10	29,00%	25,36%	10,386	Ja	Ja	Ja
11	28,34%	25,40%	7,49	Ja	Ja	Ja
12	28,03%	25,41%	5,9254	Nein	Ja	Ja
13	27,59%	25,31%	5,3488	Nein	Ja	Ja

Zusammenfassend lässt sich also festhalten, dass sehr wohl statistisch signifikante Effekte der Einführung der Drei-Punkte-Regel erkennbar sind. Dies gilt nicht nur für den ursprünglich von uns untersuchten Zeitraum von 10 Saisons vor und 10 Saisons seit der Regeländerung, sondern auch für zahlreiche weitere Untersuchungszeiträume, wenn auch der Effekt anfangs durch einen höheren Anteil der Unentschieden in der Saison 1995/1996 das falsche Vorzeichen aufweist und sich bei mehr als 10 Saisons danach und v. a. davor wieder abschwächt.

Ad Gegenargument 3: Betrachtung des gesamten Zeitraums

Strauß et al. fordern eine Betrachtung des gesamten Zeitraums seit Einführung der Fußballbundesliga, die, entgegen ihrer Abb. 1 (S. 21) bereits 1963 begann und inzwischen 46 statt 44 Saisons besteht. Dabei ergibt sich, dass der Unentschiedenanteil in den 32 Saisons unter der Zwei-Punkte-Regel leicht geringer war als in den 14 Saisons danach. Strauß et al. fassen diesen deskriptiven Befund als Argument gegen unsere Hypothese und gegen die von uns präsentierte Evidenz für einen Effekt der Einführung der Drei-Punkte-Regel auf.

Wie bereits im vorhergehenden Abschnitt ausgeführt, wird jedoch der Einfluss dieser einmaligen Regeländerung mit zunehmendem Zeitablauf von immer mehr und ggf. andauernd und nicht nur einmalig wirkenden anderen Effekten überlagert. In **Tab. 1** wird deutlich aufgezeigt, dass bereits bei Betrachtung von 2-mal 14 Saisons kein statistisch signifikanter Unterschied vor und nach der Regeländerung mehr besteht. Dementsprechend ist er auch nicht bei Betrachtung von 32 Saisons davor und 14 oder irgendwann 32 oder noch mehr Saisons danach zu erwarten. Im Jahr 1963 wurde ganz anders und unter anderen wirtschaftlichen und sportlichen Bedingungen einschließlich Spielregeln Fußball gespielt als in den 1990er Jahren oder heute.

Hinzu kommt, wie schon erwähnt, dass die Drei-Punkte-Regel gerade wegen eines Anstiegs des Unentschiedenanteils im Zeitablauf eingeführt wurde, also bei Berücksichtigung hinreichend vieler Jahre davor nicht zufällig ein niedrigerer Anteil an Unentschieden gefunden wird. Schließlich wird auch bei anderen Eventstudien, z. B. der Analyse von Aktienkursbewegungen auf Grund theoretisch möglicherweise relevanter Ereignisse, nicht ein beliebiges oder sogar maximal großes Zeitfenster gewählt, sondern ein plausibler Zeitraum um das Ereignis herum, in diesem Fall oft von wenigen Tagen oder sogar Minuten statt Jahren oder Jahrzehnten.

Ad Gegenargument 4: Abnahme des Heimvorteils

Strauß et al. stellen fest, dass der Heimvorteil über den gesamten Zeitraum der Bundesliga abnimmt. Dieser Effekt würde besser unsere nicht verworfene Hypothese 2 erklären, dass Siege nach Einführung der Drei-Punkte-Regel knapper ausfallen. Darauf ist zu erwidern, dass die Abnahme des Heimvorteils ihrerseits gleich doppelt erklärungsbedürftig ist, von Strauß et al. jedoch in keiner Weise erklärt wird, während wir eine theoretische Erklärung für den Einfluss der Änderung der Drei-Punkte-Regel unterbreitet haben, der Strauß et al. ebenfalls keinerlei andere theoretische Erklärung entgegensetzten konnten. Beim Heimvorteil ist erstens dessen Existenz zu erklären, also warum die Heimmannschaft signifikant häufiger gewinnt als die Gastmannschaft, und zweitens der Rückgang dieses Vorteils im Zeitablauf.

Schließlich mag dieser Rückgang des Heimvorteils zwar auch zu knapperen Siegen führen, v. a. sollte er aber die Zahl der Unentschieden erhöhen, da jeder verhinderte Heimsieg zumindest im Spielverlauf erst einen unentschiedenen Spielstand erfordert, bevor die Gastmannschaft gewinnen kann. Unsere Hypothese 1 behauptet das Gegenteil, so dass die empirische Evidenz für sie als noch stärker, einen Gegeneffekt überwiegend anzusehen ist. Hypothese 2 ist schließlich im Verhältnis zu Hypothese 1 zu sehen: Prima facie deuten weniger Unentschieden auf klarer entschiedene Spiele hin, während wir aus theoretischen Gründen weniger deutliche Siege erwarten und letzteres tatsächlich statistisch signifikant ist. Der abnehmende Heimvorteil mag ebenfalls zu einer Vorhersage knapperer Siege führen, jedoch nur zusammen mit mehr Unentschieden, erklärt die Befunde also weniger differenziert und nicht so gut.

Ad Schlussbemerkung: Die „Schande von Gijon" und weitere Evidenz

Strauß et al. wiederholen in ihrer Schlussbemerkung nicht nur das bereits entkräftete Gegenargument 1, sondern bauen Gegenargument 3 weiter aus, wonach sich seit „Bestehen der Fußballbundesliga auch unabhängig von der Einführung der Drei-Punkte-Regel Schwankungen in den Anteilen ergeben. Wenn man in einem solchen Schwankungsgeschehen über einen solch langen Zeitraum relativ beliebig einen Punkt festlegt, wird man bei genügend großen Fallzahlen sehr schnell signifikante Unterschiede erhalten, die ohne praktische Bedeutsamkeit und damit bedeutungslos sind" (S. 21).

Das ist zweifellos richtig, doch wir bestreiten, dass wir „relativ beliebig einen Punkt“ festlegten. Dazu wäre eine Saison nach der Änderung der Drei-Punkte-Regel z. B. viel besser geeignet gewesen, doch es gibt dann kein theoretisch bedeutsames Ereignis. Empirie ohne Theorie kann immer irgendwelche (scheinbar) signifikanten Effekte finden und deshalb umgekehrt auch stets die Relevanz der gefundenen Signifikanzen bestreiten. Strauß et al. geben nicht an, welche Art von empirischer Evidenz sie überhaupt überzeugen könnte (ein hoher w-Wert?). Wir meinen dagegen, dass es bei empirischer Forschung keine harten Beweise geben kann, diese deshalb aber auch nicht verlangt werden dürfen, sondern theoretisch sinnvoll abgeleitete Hypothesen nicht zu verwerfen sind, wenn die korrespondierende Nullhypothese sich als hinreichend unwahrscheinlich erweist. Wird hingegen mittels Datamining irgendeine beliebige Hypothese gesucht, die dann statistisch signifikant erscheint, so ist sie es in der Regel nicht wirklich, weil in diesem Fall ganz andere Tests erforderlich wären (s. oben unsere Ausführungen zum Gegenargument 2).

Dies lässt sich an dem Beispiel, welches Strauß et al. eigentlich zur Widerlegung unserer Position anführen, gut verdeutlichen. „1982 kam es in Gijon bei der WM in Spanien zwischen Deutschland und Österreich zu einem Nichtangriffspakt, nachdem Deutschland schnell 1:0 führte“ (S. 21, wobei das explizite Vorliegen eines solchen Paktes nie nachgewiesen wurde und auch nicht nötig erscheint zur Erklärung des äußerst defensiven Verhaltens zum beiderseitigen Vorteil). „Konnte man (…) Auswirkungen auf die nachfolgende Unentschiedenquote in der Bundesliga finden – quasi als Ergebnis eines typischen, ‚vorbildhaften‘ Ergebnisses eines Nichtangriffspakts? Betrachtet man den Anteil der Unentschieden in der 1. Fußballbundesliga in einem 10-Jahres-Zeitraum jeweils vor und nach diesem Spiel im Jahr 1982, so stellt man Folgendes fest: Es ergaben sich 23,82% Unentschieden vor der WM 1982 und 27,92% Unentschieden 10 Jahre nach der WM 1982. Eine Prüfung auf Gleichverteilung zwischen den beiden Zeitintervallen ergibt ein hochsignifikantes Ergebnis (…).

Tab. 3 Chow-Test auf Strukturbruch durch die Regeländerung

	Summe quadrierter Residuen				
Anzahl Saisons	Gesamt	Zwei-Punkte-Regel	Drei-Punkte-Regel	t-Wert	Strukturbruch
10	2158,85	651,588	833,952	15,25	Ja
11	2732,385	664,000	1597,600	12,75	Ja
12	3246,636	815,055	1634,710	16,51	Ja
13	3376,541	815,247	1654,874	19,07	Ja
14	3690,386	832,418	1655,358	23,6	Ja

Wir glauben nicht, dass jemand ernsthaft die ‚Schande von Gijon‘ für den signifikanten, aber praktisch irrelevanten Anstieg der Unentschieden verantwortlich machen würde“ (S. 21).

Das glauben wir auch nicht, aber nicht wegen der statistischen Maßzahlen (das ausgewiesene w beträgt 0,079, ist jedoch vermutlich wieder, s. oben zum Gegenargument 1, falsch berechnet worden, da n mit 1604 zu niedrig angegeben wird), sondern weil die „theoretische Erklärung“ nicht überzeugt bzw. offensichtlich falsch ist. Das betreffende WM-Spiel ging gerade nicht unentschieden aus, während umgekehrt unter der Zwei-Punkte-Regel ein Unentschieden in der Regel keinen taktischen Vorteil (doch im Gegensatz zur Drei-Punkte-Regel auch keinen Nachteil) für beide Mannschaften bringt.

Strauß et al. verweisen auf die Möglichkeit, mit ausgefeilteren statistischen Verfahren weitere Evidenz für oder gegen unsere zentrale Hypothese zu finden, z. B. mit Längsschnittuntersuchungen, die sie jedoch nicht unternehmen, da das „den methodischen Rahmen von Dilger und Geyer (2007) deutlich verlassen“ würde (S. 21). Sie gehen jedoch in ihrer Replik nicht auf ein anderes von uns gewähltes Verfahren ein, nämlich die Berechnung von Unentschieden je Mannschaft und Jahr, was zu realen Zahlen führt und damit stärkere statistische Testverfahren wie t-Tests erlaubt.

Ein weiteres statistisches Testverfahren lässt sich auf noch höherer Aggregationsebene verwenden, nämlich der Chow-Test (vgl. Chow, 1960) für einen Strukturbruch in den Unentschiedenanteilen jeweils für die ganzen Saisons. In ◘ **Tab. 3** werden die entsprechenden Ergebnisse ab 10 Saisons vor und nach der Regeländerung dargestellt. Es lässt sich jeweils signifikant ein Strukturbruch feststellen. Dabei ist nochmals darauf hinzuweisen, dass in den Daten allein natürlich auch an anderer Stelle Strukturbrüche auftreten und z. B. eine Saison später der Effekt noch ausgeprägter wäre, eine sinnvolle Untersuchung jedoch von theoretisch abgeleiteten Hypothesen ausgehen und diese testen sollte.

Weiterhin haben wir in einem anderen Aufsatz, der bereits im *Journal of Sports Economics* erschienen ist (Dilger & Geyer, 2009), die Pokalspiele als Vergleichsgröße herangezogen. Bei den Pokalspielen gilt die Drei-Punkte-Regel nicht und sie wäre ansonsten auch irrelevant, da der Verlierer stets ausscheidet. Da die meisten anderen Änderungen von Regeln sowie sportlichen, wirtschaftlichen und sonstigen Umständen jedoch im Pokal gleichermaßen wirksam sind wie in der Bundesliga, sollten sich auf diese Weise die Effekte der interessierenden Drei-Punkte-Regel isolieren lassen. Die empirischen Ergebnisse werden etwas schwächer, sind jedoch insgesamt noch statistisch signifikant, was die empirische Evidenz für unsere Hypothese weiter verstärkt.

Schließlich müssen wir tatsächlich einen Satz unseres Essays korrigieren (Dilger & Geyer, 2007, S. 268): „In einer ligenübergreifenden Studie mit Daten aus 7 Ländern zeigen Aylott und Aylott (2007), dass sowohl die durchschnittliche Anzahl der Tore pro Spiel (mit Ausnahme der deutschen Fußball-Bundesliga) als auch die Anzahl der Unentschieden mit Einführung der Drei-Punkte-Regel angestiegen ist.“ Korrekt muss es heißen, dass „die Anzahl der entschiedenen Spiele mit Einführung der Drei-Punkte-Regel angestiegen ist“. Im internationalen Vergleich findet sich ganz überwiegend empirische Evidenz für unsere Hypothese (die sich eben nicht nur auf die Bundesliga bezieht, was die Diskussion in Fußnote 5 auf S. 18 erledigt). Nur Deutschland schien vor unserer Untersuchung ein Sonderfall zu sein, was vermutlich mit dem frühen Zeitpunkt

anderer Studien zusammen mit dem Anstieg der Unentschieden in 1995/1996 erklärt werden kann.

Fazit

Strauß et al. kritisieren unseren Beitrag auf mehrfache Weise, die jedoch in keinem Fall überzeugend erscheint.[1] Sie verweisen erstens auf die geringe praktische Bedeutsamkeit unserer Befunde, wozu sie ein altes und ungebräuchliches statistisches Maß heranziehen und auch noch falsch berechnen, während sich die Bedeutung unmittelbar an der deskriptiven Änderung des Unentschiedenanteils um knapp 12% ablesen lässt. Zweitens halten sie unsere Auswahl der Zeitintervalle für selektiv, da in jedem anderen Intervall unsere Hypothese nicht mehr (in dem Ausmaß) signifikant wäre, was jedoch in ihrer Auswahl des falschen Tests auf Gleichverteilung statt Unabhängigkeit begründet liegt. Drittens plädieren sie für die Berücksichtigung aller Saisons seit Gründung der Bundesliga, wobei der Unentschiedenanteil vor der Regeländerung insgesamt kleiner wird als danach, was jedoch irrelevant ist, da der Zeitraum zu lang ist, zu viele andere Veränderungen beinhaltet und die Regel gerade wegen eines Anstiegs dieses Anteils geändert wurde. Viertens würde eine Abnahme des Heimvorteils besser die sinkende Tordifferenz erklären, bleibt jedoch ihrerseits unerklärt und passt nicht zum zentralen Befund des abnehmenden Anteils an Unentschieden. Schließlich führen sie die „Schande von Gijon" an, nach der der Unentschiedenanteil gestiegen sei, ohne dass jemand dies für eine sinnvolle Erklärung halten würde, was wir natürlich auch nicht tun, ohne dies als Widerspruch zur empirischen Evidenz für unsere theoretisch hergeleitete Hypothese aufzufassen, sondern als Bekräftigung dafür, dass Empirie ohne Theorie nicht sinnvoll ist. Die Sportwissenschaft muss nicht ökonomische Theorien verwenden, doch irgendwelche gehaltvollen Theorien sind nötig, von denen Strauß et al. leider keine anbieten.

Korrespondenzadresse

Prof. Dr. A. Dilger
Institut für Ökonomische Bildung und Centrum für Management
Westfälische Wilhelms-Universität Münster
Scharnhorststr. 100
48151 Münster
Alexander.Dilger@uni-muenster.de

Literatur

1. Aylott, M. & Aylott, N. (2007). A Meeting of Social Science and Football: Measuring the Effects of Three Points for a Win. *Sport in Society, 10,* 205–222.
2. Brocas, I. & Carrillo, J.D. (2004). Do the „Three-Point Victory" and „Golden Goal" Rules Make Soccer More Exciting? *Journal of Sports Economics, 5,* 169–185.
3. Chow G.C. (1960). Tests of Equality Between Sets of Coefficients in Two Linear Regressions. *Econometrica, 28,* S. 591–605.
4. Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd edn.). Hillsdale, NJ: Lawrence Erlbaum.
5. Dilger, A. & Geyer, H. (2007). Theoretische und empirische Analyse der Drei-Punkte-Regel. *Sport und Gesellschaft (Zeitschrift für Sportsoziologie, Sportphilosophie, Sportökonomie, Sportgeschichte), 4,* 265–277.
6. Dilger, A. & Geyer, H. (2009). Are Three Points for a Win Really Better Than Two? A Comparison of German Soccer League and Cup Games. *Journal of Sports Economics, 10,* 305–318..
7. Geyer, H. (2009). Warum Angriff nicht immer die beste Verteidigung ist: Strategienwahl im Fußball. *Sportwissenschaft, 39,* 7–15.
8. Strauß, B., Hagemann, N. & Loffing, F. (2009). Die Drei-Punkte-Regel in der deutschen 1. Fußballbundesliga und der Anteil unentschiedener Spiele: Eine Replik auf den Beitrag von Dilger und Geyer 2007. *Sportwissenschaft, 39,* 16–22.

Erwiderung

B. Strauß, N. Hagemann, F. Loffing
Institut für Sportwissenschaft, Westfälische Wilhelms-Universität Münster

Das Nullritual und die Beurteilung von Effekten: „Die Erde ist rund (p<?)"

„The earth is round (p<0.05)" von Jacob Cohen ist ein vielzitierter Beitrag aus dem Jahre 1994 und behandelt die Notwendigkeit, auch Effektgrößen in die Interpretation von Signifikanztests einzubeziehen. Dieser Beitrag kann als Schlusspunkt unter eine bis dahin über 60 Jahre anhaltende Debatte um den Gebrauch und die Interpretation von Signifikanztests betrachtet werden. Fisher hatte 1925 das sog. Nullhypothesentesten vorgeschlagen, in der die Annahme und Ablehnung der statistischen Nullhypothese (z. B. der Gleichheit von Häufigkeiten oder Mittelwerten) vom p-Wert abhängen. Bereits 1928 hatten Neyman und Pearson ein alternatives Modell vorgestellt, in dem die Konzepte der Alternativhypothese und der statistischen Power eingeführt werden. In den folgenden Jahrzehnten ist dann eine Debatte unter Statistikern entbrannt, die sich auch mit den häufig anzutreffenden Fehlinterpretationen von p und des Nullhypothesentestens beschäftigte.[2]

Dilger und Geyer (2007, 2009) und das Nullhypothesentesten

Dilger und Geyer (2009) erweisen sich mit ihrer grundsätzlichen Ablehnung von Effektgrößen und ihrer Favorisierung des Nullhypothesentestens als Anhänger des Testkonzepts von Fisher (1925). Sie ignorieren dabei offensichtlich eine jahrzehntelange methodische Debatte unter den Statistikern, die längst Eingang in die alltägliche empirische Forschung und in die universitäre Ausbildung gefunden hat. Ihr Plädoyer für den Verzicht auf Effektgrößen entspricht nicht dem aktuellen wissenschaftlichen Diskussionsstand.

Mit „The earth is round (p<0.05)" macht Cohen (1994) noch einmal deutlich, dass das alleinige Testen von Nullhypothesen letztlich keinen Erkenntnisgewinn bringt ["My work in power analysis led me to realize that the nil hypothesis is always false" (S. 1000)]. Es geht um den Zusammenhang, dass das Ergebnis eines Signifikanztests nicht nur von der Größe des α-Fehlers, sondern auch von der Größe der Stichprobe, der Power eines statistischen Tests (bzw. dem β-Fehler) und der Effektgröße abhängt. Je nach Konstellation der verschiedenen Stellgrößen können bei sehr großen Stichproben dann auch kleinste, inhaltlich bedeutungslose Un-

[1] Das gilt auch für die nachfolgende Erwiderung, deren massive Vorwürfe und damit wohl die ganze Debatte sich mit einem korrekten Verständnis des Unabhängigkeitstests erledigen lassen, der im einfachsten Fall, wie hier, auf vier Feldern mit einem Freiheitsgrad beruht.

[2] Zum Beispiel ist p weder Indikator für die Effektgröße noch für die Wahrscheinlichkeit über das Zutreffen einer Hypothese (vgl. zum Beispiel Cohen, 1994; Sedlmeier & Renkewitz, 2008, S. 397 f.; wie aber auch zahlreiche andere Statistiklehrbücher).

terschiede signifikant werden und bei extrem großen Stichproben ist zwangsläufig die Nullhypothese abzulehnen.[3]

Diese Zusammenhänge haben Eingang in die gängigen einführenden Statistiklehrbücher der Medizin, Psychologie, Sozialwissenschaften oder auch Sportwissenschaft gefunden (wie Bortz, 2005; Hays, 1994; Sedlmeier & Renkewitz, 2008; Strauß, Haag & Kolb, 1999; um nur einige wenige herauszugreifen) sowie in die entsprechenden Grundlagenveranstaltungen der Statistikausbildung in den verschiedenen universitären Fächern.[4] Gemeinsam ist allen Lehrbuchdarstellungen (unabhängig von der Auflage), dass die Berechnung von Effektgrößen notwendiges Element bei der Durchführung von Signifikanztests ist, entweder um a priori den notwendigen Stichprobenumfang insbesondere in Experimenten festzulegen oder a posteriori die Größe der Effekte abzuschätzen (z. B. mit Cohen, 1988; oder dem Computerprogramm G*Power von Faul, Erdfelder, Lang & Buchner, 2007). Die Angabe der Effektgröße (oder mindestens der zugrundeliegenden Teststatistik) ist in der Regel notwendig, damit empirische Studien unproblematisch Eingang in Metaanalysen finden können. Seit einigen Jahren fordern auch die Fachgesellschaften wie die APA (American Psychological Association) in ihren Publikationsstandards die Angabe von Effektgrößen (2001, S. 25 f.). Dies ist auch daher selbstverständlicher Standard in zahlreichen Peer-reviewed-Zeitschriften.

Die Verbreitung von Effektgrößen in der empirischen Forschung

In dem von uns zitierten Standardwerk[5] „Statistical power analysis for the behavioral sciences" zur Berechnung von Effektgrößen und zur Poweranalyse von Jacob Cohen (1988) werden für verschiedenste Teststatistiken (wie χ^2, t, F) und den Varianten von Signifikanztests (Korrelationen, Mittelwertsvergleiche, Häufigkeitsanalysen usw.) die entsprechenden Effektgrößen (wie d, f und eben auch w) mit dem entsprechenden Tabellenmaterial beschrieben. Eine Kurzform (ein sog. Primer) seines Lehrbuchs mit den wichtigsten Formeln zur Berechnung der Effektgrößen und der Power in verschiedenen Testsituationen findet sich übrigens bei Cohen (1990).

Effektgrößen, die verschiedenen Koeffizienten und ihre Anwendung finden sich mittlerweile in zahlreichen Statistiklehrbüchern (z. B. Bortz, 2005; Hays, 1994; Sedlmeier & Renkewitz, 2008 und zahlreichen anderen Werken) oder weiteren Büchern, die sich ausschließlich mit Effektgrößen beschäftigen (z. B. Grissom & Kim, 2005).

Effektgrößen sind relativ einfach ineinander überführbar und können leicht aus jeder Teststatistik abgeleitet werden. Wenn beispielsweise ein χ^2 berechnet wurde, existiert auch zwangsläufig ein w (vgl. z. B. Cohen, 1988, 1990; Sedlmeier & Renkewitz, 2008). Wenn ein t berechnet wurde, existiert zwangsläufig ein d oder r, und wenn ein F berechnet wurde, ein f, ω^2 oder η^2. Alle diese Effektgrößen können auch leicht unter Kenntnis weiterer Größen (wie der Stichprobengröße und ggf. Varianzkomponenten) aus der jeweiligen Teststatistik berechnet werden (vgl. z. B. Anhang 1 für die Effektgröße w).[6]

Prozentuale Veränderungen sind keine Effektgrößen

Prozentuale Veränderungen, wie sie Dilger und Geyer (2009) favorisieren, gehören nicht zu den Effektgrößen, weil sie keine Binnen- bzw. Fehlervariation sowie Stichprobengrößen einbeziehen und damit auch keine Konventionen für die Interpretation (wie bei α, β oder den Effektgrößen) vorliegen. Sie können daher nicht zur Beurteilung der Größe eines Effekts herangezogen werden (vgl. auch Fußnote 2).

Konventionen zur Beurteilung, wann ein Effekt klein, mittel oder hoch zu be-

[3] Ein Beispiel aus einem inhaltlichen Bereich, in dem aus wissenschaftlicher Sicht keine Zusammenhänge vermutet werden würden (außer von denjenigen, die daran glauben wollen), soll dies verdeutlichen: Gunter Sachs hat 1997 ein Buch veröffentlicht, in dem er den Zusammenhang von Sternzeichen und menschlichem Verhalten statistisch zeigen möchte. Er hat Millionen von Daten aus statistischen Ämtern gesammelt (neben den Geburtsdaten z. B. Eheschließungen, Studienplatzwahlen, Todesursachen und einiges mehr). Sachs führte zahlreiche χ^2-Anpassungstests und χ^2-Unabhängigkeitstests durch (die er am Beginn des Buches ausführlich beschreibt). Offenbar praktizierte Gunter Sachs das Nullhypothesentesten von Fisher (1925), denn seine *Beweisführung*, dass ein Zusammenhang zwischen Sternzeichen und Verhalten existiere, basierte alleinig auf dem ermittelten p und den *, **, ***. Er erhielt zahlreiche zum Teil hochsignifikante p's, die er als Beleg für den Zusammenhang von Sternzeichen und Verhalten interpretierte. So kann Gunter Sachs (1997, S. 178 f.) bspw. anhand der Daten einer britischen Versicherungsgesellschaft über n=25.003 Fahrunfälle aus dem Jahre 1996 zeigen, dass britische Autofahrer mit dem Sternzeichen *Stier* hochsignifikant häufiger als erwartet an solchen Unfällen beteiligt sind (p<0,001, ***). Sie sind in 13,72 % mehr Fahrunfällen verwickelt als statistisch erwartet werden würde. Wir zweifeln nicht das p an. Wir haben dies auch aus seinen berichteten Daten ermittelt. Nur: wenn die Effektgröße w (hier für das Sternzeichen Stier) errechnet wird (was wir nachgeholt haben), ergibt sich ein w=0,0398, also eines, das nach Cohen (1988) als praktisch bedeutungslos zu bezeichnen ist.

[4] Hinzuzufügen ist (was aber unerheblich für die weitere Argumentation ist), dass in den Lehrbüchern in der Regel nicht das pure Neyman- und Pearson-Konzept vorgestellt wird, sondern ein hybrides Vorgehen als Mischung aus den Ansätzen von Fisher sowie von Neyman und Pearson präsentiert wird, worauf z. B. Cohen (1994), Gigerenzer et al. (1989), Sedlmeier und Renkewitz (2008) oder auch bspw. in der Sportpsychologie Conzelmann und Raab (2009) hinweisen.

[5] Nicht nur Strauß et al. (2009) haben dieses Buch zur Kenntnis genommen und zitiert. Es handelt sich um ein außergewöhnlich vielzitiertes Werk. Die Datenbank Scopus (die vorwiegend naturwissenschaftliche, psychologische und empirische sozialwissenschaftliche Zeitschriftenbeiträge listet) weist (mit Zugriff am 31.8.2009) aus, dass das Werk alleine in wissenschaftlichen Zeitschriften 23.080-fach zitiert wurde. Der eingangs erwähnte Beitrag von Cohen (1994) übrigens 820-mal. Mit all seinen zahlreichen wissenschaftlichen Beiträgen gehört Jacob Cohen zu den meistzitierten und weltweit einflussreichsten Wissenschaftlern. Das Web of Science führt ihn in seiner sehr exklusiven Liste der „highly cited" Wissenschaftler (Zugriff am 31.8.2009).

[6] Dazu ist es natürlich grundsätzlich notwendig, dass Autoren sich an die üblichen Publikationskonventionen (z. B. der American Psychological Association, APA, oder auch der Deutschen Vereinigung für Sportwissenschaft, dvs, die die APA-Standards übernommen hat) halten und neben den Stichprobengrößen die konkreten Teststatistiken berichten. In Dilger und Geyer (2007) finden sich weder die Angabe der χ^2-Werte noch der t-Werte. Wie man trotzdem, unter bestimmten Voraussetzungen, auch bei unvollständigen Angaben die Effektgröße bestimmen kann, zeigen Sedlmeier und Renkewitz (2008, S. 644 ff.).

werten ist (wie z. B. w=0,10 oder ein f=0,10 in der Varianzanalyse einen kleinen Effekt darstellt), sind deshalb wichtig, damit die Werte intersubjektiv und unabhängig vom subjektiven Gefühl bewertbar sind. Die von Dilger und Geyer (2007, 2009) berichtete prozentuelle Differenz von 3,48% für den Vergleich der beiden Zehnjahresintervalle (was einer prozentualen Verminderung von 11,89% Unentschieden nach Einführung der Drei-Punkte-Regel entspricht) mag sich für manchen beachtlich anfühlen. Weniger beachtlich ist das Gefühl, wenn der gleiche Sachverhalt so umschrieben worden wäre: Die auf die zweite Nachkommastelle gerundete Anzahl der Unentschieden hat von 2,63 (10-Jahres-Intervall vor der Regeländerung) auf 2,32 (10-Jahres-Intervall nach der Regeländerung) Spiele pro Spieltag abgenommen. Beide Aussagen sind identisch. Und beide Aussagen entsprechen einer extrem kleinen Effektgröße nahe bei 0 von w=0,063, wie Strauß, Hagemann und Loffing (2009) korrekt berichtet haben.

Die Effektgröße w und die Berechnungen von Strauß et al. (2009)

Die Effektgröße w ist kein abseitiges, kaum bekanntes Maß, wie Dilger und Geyer (2009) suggerieren wollen, sondern *das* Standardmaß für Goodness-of-fit-Tests (oder auch Anpassungstests genannt) für den eindimensionalen Fall, etwa zur Prüfung auf Gleichverteilung oder für jedwede andere Verteilungsformen (wie der Normalverteiltheit oder andere, auch nicht symmetrische Verteilungen). Aber auch für den zweidimensionalen Fall, wenn der Zusammenhang von zwei nominalen Variablen wie beim χ^2-Unabhängigkeitstest geprüft wird, gibt w die Stärke des Effekts wieder (im Falle einer Vierfeldertafel ist w=φ, s. z. B. Sedlmeier & Renkewitz, 2008, S. 572).

Für Einsteiger, die bislang noch nicht mit Effektgrößen gerechnet haben und denen Cohen (1988) etwas zu unübersichtlich ist, empfiehlt sich z. B. das aktuelle Lehrbuch von Sedlmeier und Renkewitz (2008) mit vielen Rechenbeispielen, die auch ausführlich auf die Effektgröße w (S. 559 ff.) für den eindimensionalen und zweidimensionalen Fall eingehen, wie sie im Übrigen auch auf alle anderen gängigen Effektgrößen für andere Teststatistiken wie F und t eingehen. Im Folgenden werden wir uns der Einfachheit halber bei technischen Fragen zu den Effektmaßen auf Sedlmeier und Renkewitz (2008) sowie Cohen (1988) beschränken, um den Leserinnen und Lesern lange Literaturlisten von Statistiklehrbüchern zu ersparen, in denen Gleiches oder Ähnliches zu finden ist.

Wir haben hier für die interessierten Leser im Anhang 1 die (übersichtlichen und vielen Lesern bekannten) Formeln für den χ^2 und w notiert und ein ausführliches Rechenbeispiel für die Prüfung der Gleichverteiltheit für das 2-Jahres-Intervall (2 Jahre vor der Einführung unter der Drei-Punkte-Regel und 2 Jahre danach) in allen Einzelheiten angegeben. Das Resultat entspricht exakt den von uns in Strauß et al. (2009) berichteten Werten.

Wir haben im Übrigen natürlich nach dem in diesem Heft von Dilger und Geyer (2009) formulierten Vorwurf, wir hätten uns permanent bei der Berechnung von w verrechnet (und dies auch noch zu ihren Gunsten) und kein Wert sei richtig, alle Berechnungen noch einmal angeschaut. Wir haben keine Falschberechnungen in unserem Beitrag entdecken können. Mit den Formeln in Anhang 1 und den von uns berichteten Daten in Strauß et al. (2009) ist es jedem interessierten Leser möglich, dies zu prüfen.

Die Effektgröße w und die Berechnungen von Dilger und Geyer (2009)

Dilger und Geyer (2009) haben nun davon abweichende und fehlerhaft berechnete Werte für die Effektgröße w berichtet (in der Regel in der 2. Nachkommastelle abweichend). Wir haben versucht, die Gründe nachzuvollziehen, können diese aber aufgrund der fehlenden expliziten Beschreibung von Dilger und Geyer (2009) nur vermuten. Ein wesentlicher Grund dürfte darin liegen, dass Dilger und Geyer (2009) eine falsche Stichprobengröße in die Formel für die Effektgröße einbeziehen. Für den Vergleich zweier Häufigkeiten bzw. Anteile von Unentschieden (vorher und nachher) dividieren sie zur Berechnung von w auch durch die Anzahl der entschiedenen Spiele in dem Zeitraum (neben der Anzahl der unentschiedenen Spiele). Dies ist natürlich falsch und führt zu der von ihnen berichteten Unterschätzung der Effektgröße, da sich das n im Nenner immer auf die Anzahl bezieht, die im Zähler der Formel Eingang findet (siehe die Formeln in Anhang 1).

Es sollte aber abseits aller Diskussion über die richtige Anwendung der einfachen Formeln festgehalten werden, dass selbst Dilger und Geyer (2009) mit ihrer Art der Berechnung Effektgrößen erhalten, die noch näher an 0 liegen als die von Strauß et al. (2009) berichteten und damit im Übrigen noch weiter weg von Cohens (1988) eingeführter Grenze von w=0,10 für einen kleinen Effekt.

Die maximale Größe von w

Dilger und Geyer (2009) behaupten, das Maß w könne bei diesen Daten gar nicht den maximalen Wert von 1 erreichen, sondern nur den Wert w=0,412 und sei damit gar nicht geeignet. Dies ist falsch. In Anhang 2 ist ein Rechenbeispiel für das 10-Jahres-Intervall angegeben. Bei korrekter Anwendung der Formeln ergibt sich eine maximal mögliche Effektgröße w von 1 wie von Cohen (1988) auch beschrieben. Gleiches gilt natürlich für alle Jahresintervalle wie für alle anderen von uns berichteten χ^2-Tests.

Adjustierungen

Wenn die Saison 1991/1992, wie beispielsweise im 10-Jahres-Intervall, einbezogen werden muss, sind wegen der höheren Spielanzahl (380 Spiele, ansonsten immer 306 Spiele) marginale Adjustierungen für den χ^2-Test notwendig. Dies kann man auf unterschiedliche Arten lösen: Zum einen kann dies in die Bildung der beiden Erwartungswerte p_e (s. Anhang 1) einfließen (für den Vergleich der 10-Jahres-Intervalle vorher und nachher wären es dann die Erwartungswerte 0,50597 und 0,49403 bei n=1704 unentschiedenen Spielen). Es wäre dann immer noch ein Anpassungstest an eine gegebene Verteilung, nur ganz genau genommen ist dies dann kein Gleichverteilungstest (was aber für die Aussagekraft von w und χ^2 völlig unerheblich ist). Zum anderen kann die Adjustierung bei Erhalt der beiden Erwartungswerte von jeweils 0,5 (vor bzw. nach der Einführung

der Drei-Punkte-Regel) über eine Anpassung der Anzahl der Unentschieden in der Saison 1991/92 auf eine Spielanzahlgröße von 306 Spiele[7] erfolgen. Alle Adjustierungsverfahren führen zu nahezu identischen Ergebnissen (s. auch bereits ausführlich in Strauß et al., 2009).

Der von Dilger und Geyer (2007, 2009) favorisierte χ^2-Unabhängigkeitstest

Dilger und Geyer (2009) berichten nunmehr, dass sie gar nicht in Dilger und Geyer (2007) einen χ^2-Anpassungstest vorgenommen hätten (sie hatten es dort nicht berichtet), sondern den zweidimensionalen Fall mit einer Vierfeldertafel (vorher, nachher) x (unentschiedene Spiele, entschiedene Spiele) zur Ermittlung des Zusammenhangs geprüft hätten. Sie heben nunmehr die Vorzüge dieser Testvariante hervor. Die Einbeziehung der entschiedenen Spiele, neben den unentschiedenen Spielen, führt aber zu verzerrten Ergebnissen, da die Voraussetzung der Unabhängigkeit der Zellenbesetzungen des χ^2-Tests damit (erheblich) verletzt ist (siehe z. B. Sedlmeier & Renkewitz, 2008, S. 574 ff.). Der Grund ist einfach. Die Anzahl der entschiedenen Spiele ist bei Kenntnis der Anzahl der Spiele (dies sind pro Saison üblicherweise 306, außer 1991/92) und der Kenntnis der Anzahl der Unentschieden in der jeweiligen Saison zu 100% determiniert und damit völlig redundant. Dies führt bei dieser fehlerhaften Anwendung des χ^2-Tests zu einer Erhöhung des χ^2-Werts (wegen der Addition zweier neuer Elemente zu insgesamt 4 Summanden, vgl. z. B. Anhang 1 für den Gleichverteilungstest mit 2 Summanden) und damit zu einer Erniedrigung der Wahrscheinlichkeit p (für die Signifikanzprüfung) sowie gleichzeitig zur Erniedrigung der Effektgröße w, weil sich die Stichprobenanzahl deutlich erhöht hat und diese Größe in den Nenner einfließt. Diese Veränderung von p und w kann jeder Leser beim Vergleich der ◘ **Tab. 1** aus Dilger und Geyer (2009) mit der von Strauß et al. (2009) feststellen. Beispielsweise berichten Strauß et al. (2009) für das 10-Jahres-Intervall bei Verwendung des χ^2-Anpassungstests ein $\chi^2(1)=6{,}73$ mit einem $w=0{,}063$. Dilger und Geyer (2009, vgl. deren ◘ **Tab. 1**) errechnen mit Hilfe des χ^2-Unabhängigkeitstests ein $\chi^2(1)$ von 9,382. Das w dazu beträgt $w=0{,}039$, mithin auch hier eine extrem kleine Effektgröße nahe bei 0, kleiner noch als die von Strauß et al. (2009) für den χ^2-Anpassungstest berichtete. Diese Unterschiede finden ihre simple Erklärung in der Anwendung des zweidimensionalen χ^2-Tests, der in dieser Testsituation nicht angemessen ist.

p<0,01; p<0,05; p<0,10; p<?

Dilger und Geyer (2009) steigern mit der Erhöhung des Signifikanzniveaus von $p<0{,}01$, über $p<0{,}05$ auf $p<0{,}10$ natürlich die Anzahl der signifikanten Unterschiede vor und nach der Einführung der Drei-Punkte-Regel. Als signifikant wird allgemein ein Unterschied dann betrachtet, wenn er kleiner ist als eine selbst benannte Schranke, die in aller Regel auf der Anwendung von Konventionen beruht. Was Dilger und Geyer (2009) hier nicht beachten: Mit der Erhöhung des p auf 0,10 werden dann auch noch kleinere, und damit noch unbedeutsamere Unterschiede als signifikant betrachtet als sie sowieso schon mit einem $p<0{,}01$ berichten.

Wir haben zu Beginn unseres Beitrags auf den Zusammenhang von Stichprobengröße, Power, Effektgröße und p hingewiesen, so wie er in allen Statistiklehrbüchern zu finden ist. Dies bedeutet beispielsweise für den 12-Jahres-Zeitraum konkret, dass für ein $p<0{,}10$ ein $\chi^2(1)$ von mindestens 2,7055 und damit ein $w=0{,}037$ (also noch näher an 0 liegend) ausreicht, damit Unterschiede als signifikant eingestuft werden. Für $p<0{,}05$ ergibt sich ein $\chi^2(1)=3{,}84$ als Grenze mit einer Mindesteffektgröße von $w=0{,}044$ (vgl. die Berechnung in Anhang 1).

Um es noch einmal zu betonen: p spiegelt wegen der Abhängigkeit von der Größe der Stichprobe nicht die praktische Bedeutsamkeit des Effekts und die Effektgröße wider.

Für die Höhe der a posteriori berechneten w's und deren Beurteilung ändert das *Spielen* mit p, wie es Dilger und Geyer (2009) in ihrer ◘ **Tab. 1** machen, übrigens nichts. Die empirisch berechneten Effektgrößen bleiben so klein und unbedeutend wie sie sind, egal ob $p<0{,}01$; $p<0{,}05$; $p<0{,}10$ oder p<? verwendet wird.

t-Tests

Schließlich erwähnen Dilger und Geyer (2009), dass ein t-Test wesentlich aussagekräftiger sei. Auch t-Statistiken und damit die p-Werte sind von der Stichprobengröße abhängig. Die gleichen eingangs gemachten Argumente gelten auch hier. Auch für t-Tests gibt es Effektgrößen (d, oder auch r nach Cohen, 1988).

Diese erbringen keine anderen Resultate und keine Veränderung in der Bewertung. Als Beispiel sei das von Dilger und Geyer (2007, S. 275, ◘ **Tab. 2**) genannte Ergebnis des t-Tests für die Prüfung des Unterschieds der Anzahl der Unentschieden pro Saison und Team 10 Jahre vor und nach der Einführung der Drei-Punkte-Regel herangezogen. Obwohl Dilger und Geyer (2009) keinen t-Wert und keine Effektgröße berichten, ist es möglich (selbst wenn die Rohdaten nicht vorliegen würden), mit Hilfe der feststehenden Stichprobengröße von n=1704 und der Angabe p den t-Wert und damit r bzw. die Varianzaufklärung r^2 (über ein bei Sedlmeier & Renkewitz, 2008, S. 644 ff. berichtetes Verfahren) zu errechnen.

Dilger und Geyer (2007, S. 275) berichten ein zweiseitiges p von 0,004. Dies ergibt nach dem Verfahren von Sedlmeier und Renkewitz (2008, S. 645) eine Effektgröße von $r=0{,}0697$ und damit eine Varianzaufklärung von r^2 von 0,486%. Diese sehr niedrige Effektgröße r (die Mindestgröße für einen kleinen Effekt wäre $r=0{,}10$ nach Cohen, 1990) korrespondiert mit der ebenso niedrigen Effektgröße $w=0{,}063$ aus dem χ^2-Anpassungstest (vgl. Strauß et al., 2009).

[7] Möglichkeit 1: Für die Saison 91/92 Adjustierung der 124 unentschiedenen Spiele auf 100 (124x306/380). Dies führt zu adjustierten 892 Unentschieden vorher und mit den 788 unentschiedenen Spielen nachher zu n=1680 adjustierten Unentschieden für 2x10 Jahre. Möglichkeit 2: Es ergeben sich adjustiert 894 Unentschieden vorher und damit n=1682 Unentschieden insgesamt, wenn der gesamte 10-Jahreszeitraum bereits am Beginn der Rechnung in die Adjustierung einbezogen wird (916×3060/3134). Wie in Strauß et al. (2009, ◘ Tab. 1) haben wir uns für den ausführlichen Bericht nach Möglichkeit 2 entschieden. Beide Möglichkeiten sind logisch ebenbürtig und führen im Übrigen nur im zu vernachlässigenden Nachkommabereich zu unterschiedlichen Ergebnissen. Die Interpretation wird davon nicht berührt.

Es ist festzuhalten: Unabhängig von der Art des Signifikanztests (ob t, χ^2 oder F) sind Effektgrößen immer notwendig, um die Ergebnisse korrekt zu bewerten.

Der Chow-Test

Die Notwendigkeit, Effektgrößen zu betrachten, gilt im Übrigen natürlich auch für den Chow-Test (1960).

Der Chow-Test ist aber aus ganz anderen Gründen zur Prüfung der Hypothesen ungeeignet. Vereinfacht gesehen werden hier die beiden Regressionskoeffizienten zweier Intervalle (vor und nach der Einführung der Drei-Punkte-Regel) miteinander verglichen. Wenn sich die Koeffizienten nicht unterscheiden, belässt man es bei einer Regressionsgeraden über den gesamten Zeitraum, also über die beiden Zeiträume vorher und nachher (also kein Strukturbruch). Diese Prüfung setzt aber die Annahme von Regressionskoeffizienten, also einer Steigung von ungleich 0 voraus. Inhaltlich hieße dies, man müsste annehmen, dass es in jedem Intervall zu einer linearen Zunahme oder Abnahme innerhalb (!!) des jeweiligen Intervalls unter einer Regel gekommen ist. Diese inhaltliche Hypothese wurde bislang nicht formuliert und kann auch kaum begründet werden. Vielmehr würde man ja eher annehmen, dass die beiden Regressionsgeraden eine Steigung von 0 aufweisen, aber – wenn die Einführung der Drei-Punkte-Regel Einfluss hätte – eine parallele Verschiebung zur x-Achse im Nachfolgeintervall verursachen würde (und die Schwankungen innerhalb der Intervalle auf Zufallsschwankungen zurückgehen). Ein Chow-Test könnte diese einzig sinnvolle Hypothese gar nicht prüfen, denn er könnte unter dieser inhaltlichen Annahme keinen Strukturbruch feststellen, da die Regressionskoeffizienten identisch wären. Der Chow-Test ist gänzlich ungeeignet.[8]

Zur Immunisierungsstrategie von Dilger und Geyer (2009)

Der Beitrag von Dilger und Geyer (2009) dient offenbar dazu, ihre Behauptung, die Einführung der Drei-Punkte-Regel würde zu einem geringeren Anteil an Unentschieden führen, gegen Gegenargumente zu immunisieren. Ganz besonders augenfällig wird dies in der Nichtberücksichtigung von einzelnen Intervallen kurz nach der Einführung (Herausnahme des Intervalls ein Jahr danach) oder mit größerem Abstand zur Einführung der Drei-Punkte-Regel. Sicherlich wird man nicht im Ernst annehmen wollen, dass die FIFA eine derartig weitreichende Maßnahme einführt, die nur für 10-Jahres-Zeiträume (±einige Jahre) nachweisbar ist. Post hoc glauben sie, damit ihre Ergebnisse rechtfertigen zu können. Mit empirischer hypothesenorientierter Forschung hat dies nichts zu tun, wenn die Ergebnisse, die nicht zur Hypothese passen im Nachhinein (mit außerdem noch spekulativen Begründungen ohne empirische Basis, die sich bislang auch in keiner der internationalen Studien finden lassen[9]) eliminiert werden. Wenn Dilger und Geyer (2009) argumentieren, dass es mit Anwachsen des Zeitraums Überlagerungseffekte gibt und die Drei-Punkte-Regel trotzdem Einfluss hat, sind sie es, die die Überlagerung empirisch nachweisen müssen. Glauben allein reicht nicht.

Es geht um die empirischen Belege von Dilger und Geyer (2007), nicht um deren Theorie

Es ist nicht notwendig, eine neue Theorie vorzulegen, um die Analysen und Interpretationen von Dilger und Geyer (2007) zu kritisieren. Es geht hier allein um die empirische Frage, ob die Einführung der Drei-Punkte-Regel eine substanzielle Veränderung in entscheidenden Ergebnisvariablen, wie dem Anteil der Unentschieden oder der Anzahl der geschossenen Tore erbracht hat. Dies haben Dilger und Geyer (2007) behauptet. Wir argumentieren in Strauß et al. (2009), dass die von Dilger und Geyer (2007) vorgetragenen empirischen Belege und methodischen Analysen unzureichend sind und ihre Schlussfolgerungen nicht zulassen, und sie ihre vorgetragenen Ergebnisse überinterpretieren. Zu unserer Methodenkritik bedarf es keiner neuen Theorie.

Natürlich ist es in späteren empirischen Untersuchungen gewinnbringend, darüber nachzudenken, warum die Einführung der Drei-Punkte-Regel keine substanziellen Veränderungen in den Unentschiedenanteilen bewirkt und da-

[8] Hier ist eine gute Gelegenheit, auf die sich außerhalb jedes akademischen Diskurses befindliche Einleitung von Dilger und Geyer (2009) einzugehen. In der Tat hatten wir den ursprünglichen Beitrag bei *Sport und Gesellschaft*, naheliegender Weise, eingereicht. Ein Gutachter empfahl die zusätzliche Analyse von Longitudinaltests zur Ermittlung von Strukturbrüchen und dabei insbesondere den schon erwähnten Chow-Test (1960). Da wir bei unserer Replik im gleichen methodischen Rahmen wie Dilger und Geyer (2007) bleiben wollten und im Übrigen den Chow-Test für die Beantwortung der Fragestellung gänzlich ungeeignet halten, sind wir dem Gutachter in diesem Teil seines Gutachtens nicht gefolgt, was die Ablehnung des Beitrags für *Sport und Gesellschaft* zur Folge hatte. Dies ist ein normaler Vorgang in Peer-reviewed-Zeitschriften: Autoren entscheiden, was sie für sinnvoll und vertretbar halten, Editoren von Zeitschriften entscheiden mit Hilfe von Gutachtern, ob der Beitrag für die Zeitschrift passend ist oder nicht. Autoren stehen im Wettbewerb zueinander, wie dies für die Zeitschriften untereinander auch gilt (vgl. Strauß & Tietjens, 2002). Wir halten es dabei mit Robert Sternberg (1994), einem der führenden Kognitionspsychologen, in seinem *Begleiter* für jeden Psychologen *The psychologist's companion, Edition III*: „…even if one journal flatly rejects your article, another may love it. I'm not alone in having been (…) rejected by one journal, only to be welcomed with open arms by another." (S. 180).

[9] Beispielsweise versucht der Ökonom Moschini (2008), den Einfluss der Einführung der Drei-Punkte-Regel auf die Unentschiedenanteile in 35 Ländern über einen Zeitraum von 30 Jahren zu zeigen. Ein wenigstens plausibler Argumentationshintergrund für die Bemerkungen von Dilger und Geyer (2009), dass Effekte sich nur in einem nicht zu kurzen und nicht zu langen Intervall zeigen können, würde sich hier nicht ergeben – im Gegenteil. Im Übrigen gibt es aber bei der Moschini–Untersuchung (2008) das gleiche Problem bzgl. der Beurteilung der Stärke der Effekte. Moschini (2008) meint in den meisten Ländern langfristige Effekte mit Hilfe von Regressionsanalysen nachweisen zu können. Sein einziges Beurteilungskriterium für Unterschiede vor und nach der Einführung ist wie bei Dilger und Geyer (2007, 2009) das p des Signifikanztests, wobei er ein sehr hohes $p<0{,}10$ (s. auch oben unsere Bemerkungen zu Dilger & Geyer, 2009) als Mindestgröße festlegt. Damit können auch hier extrem kleine, und damit praktisch nicht bedeutsame Effekte sehr leicht signifikant werden, zumal die Stichprobengrößen teilweise deutlich größer sind als bei Dilger und Geyer (2007, 2009). Es fehlt hier Raum, auf diese sehr umfangreiche Untersuchung von Moschini (2008), deren Ergebnisse und Überinterpretationen einzugehen. Dies wird an anderer Stelle nachgeholt (Strauß, Loffing & Hagemann, in Vorbereitung).

mit nicht verhaltens- oder mindestens ergebniswirksam ist, obwohl vordergründig die Anreize eines Sieges erhöht wurden. Dies wäre aber Gegenstand insbesondere von Experimenten und anderen empirischen Untersuchungen, die sich nicht auf Archivdaten beschränken und das Verhalten und die Kognitionen der Athleten und Athletinnen in solchen asymmetrischen Situationen mit materiellen sowie aber auch immateriellen Anreizen in den Mittelpunkt stellen.

Fazit

Insgesamt weisen wir die Kritik von Dilger und Geyer (2009) zurück. Wir können auf dem Hintergrund von Dilger und Geyer (2007, 2009) nicht erkennen, dass in der Fußballbundesliga die Einführung der Drei-Punkte-Regel zu substanziellen Veränderungen in den Anteilen der Unentschieden geführt hat und bleiben bei unseren Aussagen und Ergebnissen in Strauß et al. (2009). Die bislang vorgetragene Datenlage von Dilger und Geyer (2007, 2009), Strauß et al. (2009) sowie Hundsdörfer (2004; s. Strauß et al., 2009) spricht dafür, dass die von der FIFA erhofften Wirkungen der Einführung der Drei-Punkte-Regel in der Fußballbundesliga nicht eingetreten sind.
Wir haben in Strauß et al. (2009) berichtet, dass sich international ein inkonsistentes Bild ergibt. Während es sogar marginale Erhöhungen des Unentschiedenanteils in England (Palacios-Huerta, 2004, der 14 Saisons nach der Einführung 1982 in England betrachtet) oder in Italien (Moschini, 2008) zu geben scheint, berichtet Moschini (2008) von zahlreichen Ländern mit signifikantem Absinken (teilweise mit einem $p<0,10$ bei extrem hoher Stichprobengröße) des Unentschiedenanteils nach Einführung der Drei-Punkte-Regel. Für Spanien beispielsweise berichtet Moschini (2008) kein signifikantes Absinken, wohingegen Garciano und Palacios-Huerta (2006) beim Vergleich zweier spanischer Saisons durch die Analyse weiterer Spielindikatoren vermuten, dass sich durch die Einführung der Drei-Punkte-Regel das Fußballspiel zu seinem Nachteil verändert und die *Schönheit* gelitten habe.

Allerdings haben alle diese Studien bislang wie Dilger und Geyer (2007, 2009) lediglich das Nullhypothesentesten praktiziert und keine Effektgrößenanalysen vorgenommen. Eine solche Analyse wird zur Zeit vorbereitet (Strauß, Loffing & Hagemann, in Vorbereitung).
Abseits aller Fragen um die richtige Anwendung von Formeln soll unser Beitrag ein klares Plädoyer für die Berücksichtigung von Effektgrößen in allen Untersuchungen sein. Dies gilt natürlich a priori, wenn möglich, den Stichprobenumfang vor der Planung eines Experiments festzulegen. Dies gilt aber auch a posteriori, wenn die Daten wie hier vorliegen (und dabei natürlich insbesondere, wenn große Stichproben analysiert werden). Gerade in solchen Situationen bei alleiniger Betrachtung der Irrtumswahrscheinlichkeit p und der Favorisierung des Nullhpothesentestens (Fisher, 1925), dessen Anhänger Dilger und Geyer (2007, 2009) ganz offensichtlich sind, werden sonst bedeutungslose Unterschiede leicht überinterpretiert und führen zu falschen Beurteilungen. Es werden praktische Signifikanz und Relevanz suggeriert, die aber tatsächlich gar nicht vorhanden sind. Die Gefahr ist groß, dass dabei auch falsche Theorien länger aufrechterhalten werden als eigentlich notwendig. Oder um auf das Zitat von Jacob Cohen (1994, S. 1000) zurückzukommen: „My work in power analysis led me to realize that the nil hypothesis is always false".

Korrespondenzadresse

Prof. Dr. B. Strauß
Institut für Sportwissenschaft
Westfälische Wilhelms-Universität Münster
Horstmarer Landweg 62b
48149 Münster
bstrauss@uni-muenster.de

Literatur

1. APA (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: apa.
2. Bortz, J. (2005). *Statistik für Sozialwissenschaftler* (6. Auflage). Heidelberg: Springer.
3. Chow, G.C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica, 28,* 591–605.
4. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
5. Cohen, J. (1990). A Power Primer. *Psychological Bulletin, 112,* 155–159.
6. Cohen, J. (1994). The earth is round (p<.05). *American Psychologist, 49,* 997–1003.
7. Conzelmann, A. & Raab, M. (2009). Datenanalyse: Das Null-Ritual und der Umgang mit Effekten in der Zeitschrift für Sportpsychologie. *Zeitschrift für Sportpsychologie, 16,* 41–54.
8. Dilger, A. & Geyer, H. (2007). Theoretische und empirische Analyse der Drei-Punkte-Regel. *Sport und Gesellschaft, 4,* 265–277.
9. Dilger, A. & Geyer, H. (2009). Unentschieden sind nicht unabhängig von der Drei-Punkte-Regel: Eine Antwort auf Strauß, Hagemann und Loffing. *Sportwissenschaft*, *39*, 347–352.
10. Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191.
11. Fisher, R.A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.
12. Garciano, L. & Palacios-Huerta, I. (2006). Sabotage in Tournaments: Making the beautiful game a little bit less beautiful. *CEPR Discussion Paper 5231.*
13. Gigerenzer, G. et al. (1989). *The empire of chance: How probability changed science and everyday life*. New York: Cambridge University Press. (dt. 1999, Das Reich des Zufalls, Heidelberg: Spektrum).
14. Grissom, R J. & Kim, J.J. (2005). Effect sizes for research. Majwah, NJ: Erlbaum.
15. Hays, W.L. (1994). *Statistics* (5. Aufl.). Belmont, CA: Wadsworth.
16. Hundsdoerfer, J. (2004). Fördert die 3-Punkte-Regel den offensiven Fußball? In P. Hammann, L. Schmidt & M. Welling (Hrsg.), *Ökonomie des Fußballs* (S. 105–129). Wiesbaden: DVU.
17. Moschini, G.C. (2008). *Incentives and outcomes in a strategic setting: The 3-points-for-a-win system in soccer* (unpublished working paper). Ames, IA: Department of Economics.
18. Neyman, J. & Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika, 20A,* 175–240, 263–294.
19. Palacios-Huerta, I. (2004). Structural changes during a century of the world's most popular sport. *Statistical Methods and Applications, 13,* 241–258.
20. Sachs, G. (1997). *Die Akte Astrologie*. München: Goldmann.
21. Sedlmeier, P. & Renkewitz, F. (2008). *Forschungsmethoden und Statistik in der Psychologie*. München: Pearson.
22. Sternberg, R. (1994). *The psychologist's companion* (Edition III). Cambridge, MA: Cambridge University Press.
23. Strauß, B., Haag, H. & Kolb, M. (Hrsg.). (1999). *Datenanalyse in der Sportwissenschaft*. Schorndorf: Hofmann.
24. Strauß, B., Hagemann, F. & Loffing, F. (2009). Die Drei-Punkte-Regel in der deutschen 1. Fußballbundesliga und der Anteil unentschiedener Spiele: Eine Replik auf den Beitrag von Dilger und Geyer 2007. *Sportwissenschaft, 39,* 16–22.
25. Strauß, B., Loffing, F. & Hagemann, N. (i. V.). Research on the impact of the three-point-rule in soccer and the overestimation of its effects by misinterpretations of statistics.
26. Strauß, B. & Tietjens, M. (2002). Wissenschaft: Wettbewerb der Ideen – Wettbewerb der Zeitschriften. Ein Plädoyer für das Publizieren in wissenschaftlichen Zeitschriften. *dvs-Informationen, 17*(1), 15–19.

Anhang 1: Beispielrechnung

Vergleich des Unentschiedenanteils 2 Jahre vor und nach der Einführung der Drei-Punkte-Regel

Im Folgenden erläutern wir exemplarisch und detailliert die Berechnung von χ^2 sowie der Effektgröße w (Cohen, 1988). Da in der Saison 1991/92 20 Mannschaften am Erstligaspielbetrieb teilnahmen und daher vorab eine Adjustierung der Vergleichswerte vorgenommen werden müsste (vgl. Text), wählen wir für den Einstieg der Einfachheit halber das symmetrische 2-Jahres-Intervall um die Regeländerung, in dem keine Adjustierung notwendig ist. Das skizzierte Vorgehen ist aber auf jedes beliebige Intervall (ggf. unter Berücksichtigung einer vorherigen Adjustierung) anwendbar:

Insgesamt fanden in den 2 Jahren vor und nach der Regeländerung n=1224 Spiele (je n=612 Spiele davor und danach) statt, wobei davon insgesamt n_{unent}=347 Spiele unentschieden endeten (n_{vorher}=169 und $n_{nachher}$=178).

Da wir daran interessiert sind herauszufinden, ob und wenn ja wie sehr sich die Unentschiedenanteile mit der Regeländerung verschoben haben, berücksichtigen wir – anders als Dilger und Geyer (2007) – für die Berechnung des χ^2- und des w-Werts ausschließlich die unentschiedenen Spiele. Dieses Vorgehen resultiert in einer 1×2-Tabelle, wobei „the *total* number of cases in the comparison" (Cohen, 1988, S. 272, Hervorhebung im Original) hier gleich 347 ist. In diesem Zusammenhang ist anzumerken, dass „the *total* number of cases" nicht meint, dass alle möglichen Fälle – also sowohl unentschiedene als auch entschiedene Spiele – einbezogen werden müssen (wie offenbar Dilger & Geyer, 2009, glauben).

Basierend auf diesen Annahmen ergibt sich mit

$$\chi^2 = \sum_{n=1}^{2} \frac{(f_n - f_e)^2}{f_e}$$

und f_1=169, f_2=178 sowie f_e=173,5 ein χ^2 von 0,23. Die sich nun anschließende Berechnung der Effektgröße w ist auf 2 Wegen möglich, wobei beide Varianten zu demselben Wert führen (für eine ausführliche Diskussion, siehe Sedlmeier & Renkewitz, 2008, S. 559 ff.).

Variante A

(Vgl. Cohen, 1988, S. 216, Formel 7.2.1 mit m=2; vgl. auch Sedlmeier & Renkewitz, 2008, S. 559).

Wir prüfen den Einfluss der Regeländerung, indem wir die Unentschiedenanteile vorher und nachher auf Gleichverteilung untersuchen. Die Annahme, dass die Anteile gleich seien, führt zu einem erwarteten Wahrscheinlichkeitswert p_e=0,5. Für die real gemessenen prozentualen Anteile ergeben sich p_1=0,487 (Unentschieden vorher) und p_2=0,513. Die Effektgröße w berechnet sich nun wie folgt:

$$w = \sqrt{\sum_{n=1}^{2} \frac{(p_n - p_e)^2}{p_e}} = 0.026.$$

Variante B

(Vgl. Cohen, 1988, S. 223, Formel 7.2.5, vgl. auch Sedlmeier & Renkewitz, 2008, S. 560).

Aus den vorherigen Berechnungen ergab sich ein χ^2 von 0,23 (gerundet auf die 2. Nachkommastelle) und die Gesamtzahl der Unentschieden in dem betrachteten Zeitraum lautete N_{unent}=347. Diese beiden Größen reichen bereits zur Berechnung von w, nämlich:

$$w = \sqrt{\frac{\chi^2}{N_{unent}}} = 0.026.$$

Genau diesen Wert haben wir in Strauß et al. (2009) berichtet, und es ist unerheblich, ob Sie es aus Variante A oder Variante B berechnen. Wenn ein χ^2 vorhanden ist, ist auch unmittelbar die Effektgröße w durch einfache Umrechnung mit der Division an der Stichprobengröße vorhanden.

Anhang 2

Das maximale w

Dilger und Geyer (2009) führen in ihrer Replik aus, dass der Wert der Effektgröße w für „die maximal mögliche Reduktion auf gar kein Unentschieden nach der Regeländerung" auf w=0,412 für den Vergleich der beiden 10-Jahres-Intervalle vor und nach der Regeländerung zu beziffern sei und damit sogar noch unterhalb des von Cohen (1988) vorgeschlagenen großen Effekts von w=0,50 liege. Dies ist falsch, wie folgende Rechnung zeigt.

Tatsächlich ergibt sich für den Fall des dramatischen Absinkens der Unentschiedenquote auf 0% ein maximales w mit Wert 1. Dies entspricht im Übrigen auch dem bei Cohen (1988, vgl. S. 218) hergeleiteten maximalen w für eine 1×2-Tabelle:

$$w = \sqrt{2-1} = 1.$$

Hier unsere Berechnung für w – wie zuvor unter der Annahme einer Gleichverteilung der Unentschieden vor und nach der Regeländerung (p_e=0,5) für das 10-Jahres-Intervall.

- Unentschieden vorher: 29,23% (=894 Unentschieden unter Berücksichtigung der Adjustierung aufgrund der Saison 1991/92, vgl. Fußnote 6, Möglichkeit 2).
- Hypothetischer Anteil nachher: 0% (=0 Unentschieden).

Daraus ergeben sich wiederum p_1=1 (100% Unentschieden vorher) und p_2=0 (0% Unentschieden nachher) und nach Variante A (vgl. Anhang 1):

$$w = \sqrt{\sum_{n=1}^{2} \frac{(p_n - p_e)^2}{p_e}} = \sqrt{2 \cdot \frac{(1-0.5)^2}{0.5}} = 1,$$

bzw. nach Variante B (vgl. Anhang 1), mit

$$\chi^2 = \sqrt{\frac{(894-447)^2}{447}}.2 = 894$$

und N_{unent}=894 (gesamte Anzahl an Unentschieden im betrachteten Zeitraum):

$$w = \sqrt{\frac{\chi^2}{N_{unemt}}} = \sqrt{\frac{894}{894}} = 1.$$

Damit sind auch die von Cohen (1988, s. auch Sedlmeier & Renkewitz, 2008) formulierten Konventionen für die Beurteilung der Höhe der Effektgrößen für Dilger und Geyer (2007, 2009) anzuwenden. Ein w=0,10 entspricht einem kleinen Effekt, ein w=0,30 einem mittleren Effekt und ein w=0,50 einem großen Effekt.