DER REIHENFOLGE-EFFEKT BEI DER BEURTEILUNG VON SCHNELLEN BEWEGUNGEN AM BEISPIEL VON KAMPFRICHTERURTEILEN IM GERÄTTURNEN

Von der Fakultät Wirtschafts- und Sozialwissenschaften der Universität Stuttgart zur Erlangung der akademischen Würde einer Doktorin der Philosophie (Dr. phil.) genehmigte Abhandlung.

Vorgelegt von

Karla Graf

aus Neustadt

Hauptberichter:	Prof. Dr. Wolfgang Schlicht
Mitberichter:	Prof. Dr. Henning Plessner
Tag der mündlichen Prüfung:	29. Juli 2010

Institut für Sport- und Bewegungswissenschaft, September 2010

Inhaltsverzeichnis

ABBILDUNGSVERZEICHNIS **4**

TABELLENVERZEICHNIS **5**

ABKÜRZUNGSVERZEICHNIS **6**

ZUSAMMENFASSUNG **8**

ABSTRACT **10**

1 FORSCHUNGSPROBLEM **12**

2 LEISTUNGSURTEILE IM SPORT **16**

2.1 Die sportliche Leistungsbeurteilung und ihre Gütekriterien 18
2.2 Kampf- und Schiedsrichter und ihre Entscheidungen 28
2.3 Wertungsvorschriften und Kampfrichterlizenzen im Gerätturnen 31
2.4 Die komplexe Aufgabe der Kampfrichter 36

3 URTEILSFEHLER **41**

3.1 Urteilsbildung und die Entstehung von Urteilsfehlern 41
3.2 Verzerrende Einflüsse auf Kampfrichterurteile 48
3.3 Der Reihenfolge-Effekt 65
3.4 Vorhersage von Reihenfolge-Effekten 73

4 FORSCHUNGSFRAGEN **82**

5 METHODE UND MATERIAL **85**

5.1 Design 85
5.2 Untersuchungsvariablen 88
5.2.1 Die Geräteauswahl 89
5.2.2 Unabhängige Variablen 91
5.2.3 Abhängige Variable 93
5.2.4 Störvariablen und der Einfluss anderer Effekte 93
5.3 Videomaterial und Präsentation 98
5.3.1 Auswahl der Video-Personen 98
5.3.2 Erstellung der Videoversionen 100
5.3.3 Präsentation 104
5.4 Bewertungs- und Personenfragebogen 105

6 UNTERSUCHUNGSTEILNEHMER UND DURCHFÜHRUNG **107**

6.1 Untersuchungsteilnehmer 107
6.1.1 Rekrutierung der Studienteilnehmer 107
6.1.2 Beschreibung der Stichprobe 110
6.2 Durchführung der Untersuchung 114
6.3 Vorgenommene Änderungen im Zweitexperiment 118

7 HYPOTHESEN UND STATISTISCHE AUSWERTUNG **120**

7.1 HYPOTHESEN ... 120
7.2 STICHPROBENPLANUNG ... 121
7.3 DATENBEREINIGUNG UND -AUFBEREITUNG ... 124
7.4 STATISTISCHE DATENAUSWERTUNG ... 127

8 ERGEBNISSE ... 130

8.1 ERSTEXPERIMENT ... 130
8.1.1 Betrachtung der beiden Teil-Stichproben ... 130
8.1.2 Deskriptive Datenauswertung ... 131
8.1.3 Inferenzstatistische Auswertung ... 136
8.2 ZWEITEXPERIMENT ... 142
8.2.1 Deskriptive Datenauswertung ... 142
8.2.2 Inferenzstatistische Auswertung ... 150

9 INTERPRETATION UND RELEVANZ DER ERGEBNISSE ... 156

10 AUSBLICK ... 164

LITERATURVERZEICHNIS ... 166

ANHANG ... 176

Abbildungsverzeichnis

Abbildung 1: Die Stufen der sozialen Informationsverarbeitung (Fiedler & Bless, 2002, S. 133) 46
Abbildung 2: Übersicht möglicher verzerrender Einflüsse auf Kampfrichterurteile im Gerätturnen zugeordnet zu den Stufen der sozialen Informationsverarbeitung (eigene Darstellung, angelehnt an Fiedler & Bless, 2002, S. 133) 49
Abbildung 3: Geräte im männlichen Gerätturnen geordnet von langsam nach schnell (eigene Darstellung) 89
Abbildung 4: Videosequenz unterteilt in einzelne Abschnitte vor und nach dem Videoschnitt 104
Abbildung 5: Aktuelle Lizenzhöhe der VPn des Erstexperiments 111
Abbildung 6: Aktuelle Lizenzhöhe der VPn des Zweitexperiments 113
Abbildung 7: Schematische Darstellung des Versuchsablaufs 116
Abbildung 8: Interaktionsdiagramme der zweiten Übung am Reck (Reihenfolge - links und Bewertungstyp - rechts) 140
Abbildung 9: Konfidenzintervalle der Kontrollübungen im Erstexperiment (Ringe-Übung drei – links; Reck-Übung drei – rechts) 150
Abbildung 10: Konfidenzintervalle der Kontrollübungen im Zweitexperiment (Ringe-Übung vier – links; Reck-Übung vier – rechts) 150
Abbildung 11: Interaktionsdiagramme der sechsten Reck-Übung (Reihenfolge - links und Bewertungstyp - rechts) 152

Tabellenverzeichnis

TABELLE 1: VORHERSAGEN DES ‚BELIEF-ADJUSTMENT'-MODELLS (ANGELEHNT AN HOGARTH & EINHORN, 1992, P. 7) 78
TABELLE 2: KLASSIFIZIERUNG DER ERGEBNISSE VON UNTERSUCHUNGEN ZUM REIHENFOLGE-EFFEKT (HOGARTH & EINHORN, 1992, P. 5) 79
TABELLE 3: UNTERSUCHUNGSDESIGN DES ERSTEXPERIMENTS 87
TABELLE 4: UNTERSUCHUNGSDESIGN DES ZWEITEXPERIMENTS 88
TABELLE 5: DEMOGRAPHISCHE ANGABEN DER ERSTEXPERIMENT-TEILNEHMER 110
TABELLE 6: DEMOGRAPHISCHE ANGABEN DER ZWEITEXPERIMENT-TEILNEHMER 112
TABELLE 7: MÖGLICHE REIHENFOLGE-EFFEKTE BEI DER BEURTEILUNG VON TURNÜBUNGEN 121
TABELLE 8: DESKRIPTIVE STATISTIK DER ERSTEXPERIMENT-ÜBUNGEN (UNTERSUCHUNGSBEDINGUNG EINS UND ZWEI) 132
TABELLE 9: DESKRIPTIVE STATISTIK DER ERSTEXPERIMENT-ÜBUNGEN (UNTERSUCHUNGSBEDINGUNG DREI UND VIER) 133
TABELLE 10: ABZÜGE FÜR DIE EXPERIMENTALÜBUNG VIER AN DEN RINGEN 138
TABELLE 11: ABZÜGE FÜR DIE EXPERIMENTALÜBUNG ZWEI AM RECK 139
TABELLE 12: ABZÜGE FÜR DIE EXPERIMENTALÜBUNG VIER AM RECK 141
TABELLE 13: ERSTEXPERIMENT – ÜBERBLICK ZUM INFERENZSTATISTISCHEN VERGLEICH DER EXPERIMENTALÜBUNGEN (EINWERT- UND KONTROLLÜBUNGEN NICHT AUFGEFÜHRT; HE – HAUPTEFFEKT) 142
TABELLE 14: DESKRIPTIVE STATISTIK DER ZWEITEXPERIMENT-ÜBUNGEN (UNTERSUCHUNGSBEDINGUNG EINS UND ZWEI) 143
TABELLE 15: DESKRIPTIVE STATISTIK DER ZWEITEXPERIMENT-ÜBUNGEN (UNTERSUCHUNGSBEDINGUNG DREI UND VIER) 144
TABELLE 16: ABZÜGE FÜR DIE EXPERIMENTALÜBUNG SECHS AM RECK 152
TABELLE 17: ABZÜGE FÜR DIE EXPERIMENTALÜBUNG DREI AN DEN RINGEN 154
TABELLE 18: ZWEITEXPERIMENT – ÜBERBLICK ZUM INFERENZSTATISTISCHEN VERGLEICH DER EXPERIMENTALÜBUNGEN (EINWERT- UND KONTROLLÜBUNGEN NICHT AUFGEFÜHRT; HE – HAUPTEFFEKT) 155
TABELLE 19: DARSTELLUNG DER VARIANZANALYTISCHEN KENNWERTE DES ERSTEXPERIMENTS (EINWERTÜBUNGEN NICHT AUFGEFÜHRT) 180
TABELLE 20: DARSTELLUNG DER VARIANZANALYTISCHEN KENNWERTE DES ZWEITEXPERIMENTS (EINWERTÜBUNGEN NICHT AUFGEFÜHRT) 181

Abkürzungsverzeichnis

A-Note:	Schwierigkeitswert einer Gerätturn-Übung
AUT:	Österreich (offiziell verwendete Abkürzung)
B-Note:	Ausführungswert einer Gerätturn-Übung
CdP:	Code de Pointage (internationale Wertungsvorschriften im Gerätturnen)
E:	Einwertübung
EoS:	„End of sequence“-Prozess der Beurteilung (Abzüge werden nicht notiert)
FIG:	Fédération Internationale de Gymnastique (Internationaler Turnerbund)
GER:	Bundesrepublik Deutschland (offiziell verwendete Abkürzung)
H:	Experimentalübung mit einem deutlichen Fehler hinten
K:	Kontrollübung
KZS:	Kurzzeitspeicher des Gedächtnisses
LZS:	Langzeitspeicher des Gedächtnisses
M:	Mittelwert
N:	Anzahl der Versuchspersonen
SbS:	„Step by step“-Prozess der Beurteilung (Abzüge werden notiert)
SD:	Standard deviation (Standardabweichung)
SUI:	Schweiz (offiziell verwendete Abkürzung)
UKZS:	Ultrakurzzeitspeicher des Gedächtnisses
V:	Experimentalübung mit einem deutlichen Fehler vorne
VA:	Varianzanalyse
VPn:	Versuchsperson

Zusammenfassung

Im Gerätturnen entscheiden Kampfrichter[1] über Sieg oder Niederlage im sportlichen Wettkampf. Dabei werden von ihnen komplexe kognitive Fähigkeiten abverlangt, die teilweise sogar ihre Kapazitäten überschreiten (O´Brien, 1991; Plessner, 1997, 2004; Salmela, 1978). Die vorhandene Literatur auf dem Gebiet der Sozialpsychologie zeigt, dass auch Kampfrichterurteile den allgemeinen Prinzipien der sozialen Urteilsbildung unterliegen. Bereits in den 50er Jahren wurde die verzerrende Wirkung des Reihenfolge-Effekts auf Urteile in verschiedenen Kontexten eindrucksvoll belegt (Asch, 1946; Anderson, 1959). Beim Recency-Effekt, einer Ausprägung des Reihenfolge-Effekts, neigt die Person dazu, spätere Informationen als wichtiger zu bewerten als frühere Informationen. Der gegenläufige Primacy-Effekt zeigt sich dann, wenn die zuerst präsentierten Informationen stärker in das Urteil eingehen.

Bis heute wurde aber Reihenfolge-Effekten innerhalb einer Übung keine Beachtung geschenkt. Das besondere Interesse besteht darin, inwieweit die unterschiedliche Reihenfolge von schnellen Bewegungen, innerhalb einer Übung, beeinflussend auf das Kampfrichterurteil wirkt. Erhält man somit für denselben Fehler unterschiedliche Abzüge, abhängig davon, ob er am Anfang oder am Ende der Übung zu sehen ist? Vom Forschungsbereich der sozialen Informationsverarbeitung ausgehend und mittels des ‚belief-adjustment'-Modells (Hogarth & Einhorn, 1992) wird daher die folgende Vorhersage überprüft: Kampfrichter neigen bei der klassischen sequenziellen Bewertung von Turnübungen zu einem Reihenfolge-Effekt.

Die empirische Arbeit ist so angelegt, dass lizenzierte Kampfrichter verschiedene auf dem Laptop präsentierte Videos von Übungen im männlichen Gerätturnen bewerten. Dabei sollen sie nach den gültigen Wertungsvorschriften ihre Urteile bezüglich der Ausführung abgeben. Eine Hälfte der Kampfrichter notiert die vorgenommenen Abzüge simultan zur geturnten Übung. Die andere Hälfte darf sich während der Bewertung keine Aufschriebe machen, sondern gibt die Bewertung am Ende jeder Übung ab. Die Übungen werden an zwei Geräten gezeigt, wobei sich diese in einem grundsätzlichen Aspekt unterscheiden: Dem Zeitpunkt des Auftretens eines groben Fehlers in der Übung. Die Unterscheidung in schnelle und langsame Geräte wird aufgegriffen (Plessner, 1999) und untersucht, ob der Reihenfolge-Effekt gerätespezifische Besonderheiten aufzeigt.

[1] Zur besseren Lesbarkeit werden in dieser Arbeit hauptsächlich männliche Personenbezeichnungen verwendet. Frauen sollen dabei nicht ausgeschlossen werden.

In zwei Experimenten wird damit ein detaillierter Einblick in die Urteilsfindung von Kampfrichtern gegeben. Weiterhin wird das ‚belief-adjustment'-Modell getestet und Vorschläge zur Optimierung der Ausbildung von Kampfrichtern technisch-kompositorischer Sportarten gegeben.

Im Rahmen der Datenauswertung, wird mittels zweifaktorieller Varianzanalyse überprüft, ob die Annahmen der vorliegenden Arbeit im speziellen Untersuchungsfeld der Experimente Gültigkeit haben. Die Untersuchungen zeigen, dass ein überzufälliger Unterschied in den Kampfrichterwertungen bezüglich der präsentierten Fehlerposition nicht nachgewiesen werden konnte.

Vermutet werden kann, dass sich der von Hogarth und Einhorn (1992) vorhergesagte Recency-Effekt für komplexe Aufgaben durch einen bestehenden Primacy-Effekt aufgehoben wird und dadurch nicht zum Vorscheinen kommt. Der Primacy-Effekt entsteht häufig im Sportfeld (Greenless, Dicks, Thelwell & Holder, 2007) und lässt sich darauf zurückführen, dass der Unparteiische die Leistungsfähigkeit als stabiles Gebilde ansieht und aufgrund des ersten Eindrucks ein relativ festes Urteil bildet. Damit werden nachfolgende Informationen in geringerem Maße in das Leistungsurteil einbezogen bzw. in die entsprechende Richtung interpretiert. Weiterhin werden andere Faktoren zur Erklärung des Nichtauftretens des Reihenfolge-Effekts formuliert, wie etwa Besonderheiten der Studien oder statistisch begründbare Einschränkungen des Untersuchungsfeldes.

Die Art der Bewertung scheint im Gegensatz dazu, vor allem für weniger routinierte Kampfrichter, nicht unbedeutend für die Sportart Gerätturnen zu sein. Die Ergebnisse lassen vermuten, dass dieser Faktor überzufällige Unterschiede in den Wertungen hervorbringt und sollte daher in der Aus- und Fortbildung entsprechend thematisiert werden. Auch der Geräteeffekt ist erkennbar und zeigt, dass das schnelle Gerät Reck anfälliger für Urteilsverzerrungen dieser Art ist.

Abstract

In many sports, so in artistic gymnastics, referees decide over victory or defeat in competition. This capacity requires complex cognitive abilities, which at times are clearly overstressed (O´Brien, 1991; Plessner, 1997, 2004; Salmela, 1978). Social psychological literature shows that referees also underlie the principles of social judgement. Already in the 50s the biasing impact of the order effect was allocated impressively on judgement in different contexts (Asch, 1946; Anderson, 1959). The recency effect, as one type of order effect, the person is prone to evaluate later information more important then prior information. The opposite primacy effect can be found, when primarily presented information bias the judgement to a bigger extent.

Until today, however, no attention has been paid to order effects within an exercise of an aesthetic sport. The focus of interest here is in determining to what extent differing orders of fast movements within a routine bias the judge's rating thereof. So, does an athlete obtain different deductions for the same mistake dependant upon its sequence of occurrence, i.e. at the beginning or at the end of an exercise? Based on the research field of social information processing and using Hogarth and Einhorn's (1992) belief-adjustment model, the following prediction will be proved: sequential judgements cause referees who are evaluating a gymnastic routine to be prone to an order effect.

Data was collected for empirical analysis by asking judges of men's gymnastics to rate different videotaped exercises of gymnasts. They had to evaluate videos which were shown on a laptop. In doing so, they had to deliver a judgement regarding the execution of the routine and respecting the current code of points. One half of the test persons was permitted to take notes of all deductions for execution while watching the exercise. The other half was not permitted to make any notation during the evaluation and had to make an assessment after each routine. The exercises were shown on two different apparatuses and their presentation differed according to one fundamental aspect: the point in time of appearance of a major error in the exercise. The distinction between 'fast' and 'slow' apparatuses was used (Plessner, 1999) in order to determine whether the order effect demonstrates apparatus-specific distinctions.

Two experimental studies provide detailed insight into the deliberate judgement of referees, test the belief-adjustment model and develop propositions for the improvement of judgement training in aesthetic sports.

Within the data evaluation, a univariate analysis of variance tested, in order to determine if the assumptions of this doctoral thesis are valid within the special

field of research. The studies show that there is a difference in the ratings of judges which is not more frequent than random concerning the presented order of fault.

It can be supposed, that the predicted recency effect of Hogarth and Einhorn (1992) for complex tasks is nullified through a primacy effect and thereby invisible. The primacy effect is frequent in the field of sport (Greenless, Dicks, Thelwell & Holder, 2007) and can be traced back to the belief of the judge that ability is a stable characteristic and that, therefore, the first impression offers a relatively solid basis for judgment. For this reason, later information is included to a lesser extent in the judgement of performance or, rather, is interpreted in the according direction. Beyond that, other factors explaining the non-appearance of the order effect are assumed, i.e. specifics in the studies, or statistical restrictions of the study field.

The judgment type seems on the contrary not to be unimportant for artistic gymnastics — especially for less experienced judges. The findings let one assume that this factor brings differences in the ratings which are more frequent than random and should therefore be picked out as a central theme in educating judges. The apparatus effect is observable and shows that the fast apparatus high bar is prone to judging biases of this manner.

1 Forschungsproblem

Die olympischen Spiele 2004 in Athen führten bei den Turnentscheidungen zu einer Protestflut. Die russische Delegation reichte beim Internationalen Olympischen Komitee (IOC) einen offiziellen Protest gegen die Bewertungen der Kampfrichter im Gerätefinale ein. Der Russe Nemov hatte im Reck-Finale für seine Übung 9,725 Punkte erhalten. Die Zuschauer waren mit der geringen Wertung für diese spektakuläre Übung nicht einverstanden und machten ihrem Unmut minutenlang durch Buhrufe Luft. Der Verantwortliche des Technischen Komitees und der Oberkampfrichter korrigierten daraufhin die Ergebnisse des malaysischen und des kanadischen Kampfrichters leicht nach oben. Nemov erhielt dadurch eine Wertung von 9,762 und konnte letztlich aber doch keinen Medaillenplatz erreichen. Die Bewertungen der Kampfrichter gelten als Tatsachenentscheidungen. Sie dürfen aufgrund von Zuschauerrufen also gewöhnlich nicht korrigiert werden (Neumaier, 1988). Die Reaktion der Verantwortlichen ist demzufolge fragwürdig. Der Athlet selbst sowie die russische Delegation äußerten den Verdacht von vorherigen Absprachen und Vorurteilen gegenüber russischen Athleten und glaubten fortan nicht an eine faire Bewertung seitens der Kampfrichter.

Der Internationale Turnverband (FIG) musste weitere umstrittene Kampfrichterurteile zur Kenntnis nehmen. So gab es auch an den Ringen einen Protestfall der bulgarischen Delegation gegen die Bewertung des griechischen Olympiasiegers Tampakos. Hier trennten 0,012 Punkte die ersten beiden Ränge. Dieser Protest wurde abgewiesen, da zum einen eine Tatsachenentscheidung vorlag und zum anderen auf der Tatsache, dass man nur gegen die Bewertung des eigenen Athleten, nicht aber gegen die eines anderen Athleten Protest einlegen darf.

Auch die südkoreanische Delegation protestierte gegen eine falsche Bewertung. Die Barren-Übung des Koreaners erhielt einen Schwierigkeitswert von 9,9 statt von 10,0 Punkten. Dieses Zehntel hätte ihm die Goldmedaille im Mehrkampf eingebracht. Der Amerikaner Hamm erhielt aber die Goldmedaille, obwohl sie dem Koreaner Yang zustand. Die FIG gestand den Fehler der Kampfrichter ein, suspendierte diese, lehnte eine Korrektur der Wertung aber ab. So wurde dem Koreaner noch nach seinem Landsmann Kim, der die Silbermedaille erhielt, lediglich die Bronzemedaille überreicht. Auch die Amerikaner beriefen sich zu ihrer Entlastung auf die Unanfechtbarkeit von Tatsachenentscheidungen und wollten ‚ihre' Medaille nicht abgeben. Da die koreanische

Delegation nicht während des Wettkampfes, sondern zwei Tage später den Protest einlegte, wurde diesem aus formalen Gründen nicht stattgegeben und Hamm behielt seine Goldmedaille.

Die Beispiele zeigen, dass die subjektive Urteilsbildung und auftretende Fehlurteile bedeutend sein können, vor allem, wenn nur Zehntel eines Punktes über Sieg und Niederlage entscheiden (O´Brien, 1991). Suboptimale Bedingungen in der Urteilssituation führen oftmals dazu, dass für die Beurteilung von Personen und deren Leistungen nicht alle relevanten Informationen beachtet werden können (Bless & Keller, 2006). Irrelevante Informationen, wie Zuschauerzurufe, die nationale Herkunft des Athleten, Sympathie und Antipathie oder Gründe der Reputation sind als Ursachen eines Fehlurteils denkbar und dem Beurteiler oftmals nicht bewusst.

Das Handlungsfeld Sport stellt ein bestens geeignetes Feld für Untersuchungen der menschlichen Urteilsfindung und für mögliche systematische Beeinflussungen dar (Plessner & Raab, 1999). Die Sportart Gerätturnen bietet sich aus mehreren Gründen an. Einerseits, da verschiedene Personengruppen unterschiedliche Meinungen darüber haben, wer warum die höchste Bewertung erhalten sollte. Manchmal kommt es vor, dass ein Kampfrichter verhältnismäßig hohe oder auch niedrige Bewertungen vergibt. Nicht immer muss das direkt mit der erbrachten Leistung zusammenhängen, sondern kann durch die Berücksichtigung irrelevanter Informationen zustande kommen. Die meisten Untersuchungen von verzerrenden Einflüssen auf Bewertungsurteile im Sport wurden im Bereich Gerätturnen durchgeführt. Hier sind die Urteilssituation und die Urteilsaufgabe soweit strukturiert, dass die wichtigsten Einflussfaktoren experimentell gut kontrolliert werden können. Andererseits sollte in experimentellen Arbeiten allgemein und im speziellen Feld der Urteils- und Entscheidungsforschung ein hoher Anspruch bestehen die soziale Realität abzubilden. Im sportlichen Kontext kann eine ‚natürliche' Entscheidungssituation genutzt werden. Menschen werden nicht künstlich in eine für sie komplett ungewohnte Situation versetzt. Die gewählte Entscheidungssituation kommt so in der Realität vor, erhöht daher die externe Validität von experimentellen Untersuchungen und ist zudem selbst gewählt. Der Sport ist dadurch prädestiniert für Untersuchungen der Urteils- und Entscheidungsbildung und den damit zusammenhängenden Theorien.

Unterschiedliche Urteilsfehler wurden somit im Laufe der Jahre im sportlichen Kontext erforscht. Alle bisherigen Untersuchungen zum Positions-Effekt im Sport haben die Startreihenfolge und damit die

Reihenfolge einzelner Athleten betrachtet. Die Studien belegen, dass die Urteile der Kampfrichter, durch die Reihenfolge in der die Athleten einer Mannschaft den Wettkampf antreten, beeinflusst werden (Ansorge et al., 1978; Plessner, 1997, 1999; Scheer, 1973; Scheer & Ansorge, 1975, 1979).

Bis heute wurde der *Reihenfolge von schnellen Bewegungen* und somit innerhalb einer Übung keine Beachtung geschenkt[2]. Denkbar ist, dass auch in diesem Zusammenhang die Reihenfolge eine beeinflussende Wirkung auf das menschliche Urteil hat. Praktisch relevant ist diese Frage, wenn Turner und Trainer ab und an vor dem Problem stehen, dass sie bei der Komposition von Übungen unsicher sind, in welcher Reihenfolge sie ihre Übungselemente anordnen sollen. Beispielsweise in der Planungsphase einer Bodenübung ist nicht klar, ob die Höchstschwierigkeit am Anfang oder am Ende geturnt werden soll. Zumeist wird diese Frage aus pragmatischem Blickwinkel beantwortet, nämlich inwieweit die Kraft des Athleten ausreicht, um ein schwieriges und meist kraftraubendes Element am Ende noch turnen zu können. Was ist aber, wenn die Reihenfolge der Elemente dem Athleten dazu verhelfen könnte, eine möglichst gute Wertung im Wettkampf zu erzielen?

Ziel der Arbeit ist die experimentelle Untersuchung des Einflusses des Übungsaufbaus bzw. der Reihenfolge von Übungselementen auf Kampfrichterurteile im Gerätturnen der Männer. Die genaue Fragestellung lautet: *Führt ein Fehler abhängig davon, ob er zu Beginn oder am Ende der Übung gezeigt wird, zu unterschiedlichen Bewertungen?* Das männliche Gerätturnen bietet sich bei der vorliegenden Fragestellung besonders an, da die Mehrzahl an relevanten Untersuchungen in diesem Feld durchgeführt wurden (Ansorge et al., 1978; Plessner, 1997, 1999; Scheer, 1973; Scheer & Ansorge, 1975, 1979). Überdies lassen sich die geplanten experimentellen Untersuchungen nur an den Männergeräten realisieren (5.3.2).

Das ‚belief-adjustment'-Modell von Hogarth und Einhorn (1992) ist das einzige theoretische Modell, das detaillierte Angaben zu den Bedingungen von Reihenfolge-Effekten macht und diese voraussagt. Dabei geht es vor allem um die Frage, ob der erste Eindruck (Primacy-Effekt) oder der letzte Eindruck (Recency-Effekt) einen bedeutenderen Einfluss auf das Gesamturteil hat. Durch die Anwendung dieses Modells auf die

[2] Um begrifflich klar zum Ausdruck zu bringen, dass in der vorliegenden Arbeit der Positions-Effekt im Zusammenhang mit schnellen (Einzel-) Bewegungen gemeint ist, wird weiterhin dieser untersuchte Effekt als Reihenfolge-Effekt bezeichnet.

geschilderte Urteilssituation im sportlichen Kontext soll dessen externe Validität überprüft werden. Interessant erscheint, inwieweit die vorhergesagten Bedingungen wirklich auf Urteilssituationen des Sportbereichs übertragbar sind. Außerdem soll durch die Erkenntnisse der Untersuchung ein Beitrag geleistet werden, die Urteilssituation im Gerätturnen besser verstehen und erklären zu können.

In den folgenden Kapiteln geht es um den Einfluss der Reihenfolge auf soziale Urteile im sportlichen Kontext. Dazu werden zunächst Leistungsurteile im Sport in den Fokus gestellt (2). Zu Beginn soll die Güte der Beurteilung thematisiert werden (2.1). Daraufhin werden die Unparteiischen und ihre Entscheidungen beschrieben (2.2), bevor auf die Wertungsvorschriften (2.3) und die komplexe Aufgabe der Kampfrichterei (2.4) eingegangen wird. Das 3. Kapitel stellt die Urteilsfehler in den Vordergrund und soll die Entstehung dieser klären (3.1). Welche möglichen Einflüsse den Kampfrichter davon abhalten, objektiv ein Urteil zu fällen, stellt einen weiteren Gesichtspunkt der Arbeit dar (3.2). Der Reihenfolge-Effekt mit den unterschiedlichen Bedingungen, die entweder die ersten oder die letzten Informationen als bedeutender herausstellen, werden beleuchtet (3.3). Der theoretische Hintergrund, der mit Hilfe der Vorhersagen des ‚belief-adjustment'-Modells gebildet wird, schließt das Kapitel ab (3.4).

Aus dem aktuellen Forschungsstand und weiteren Überlegungen ableitend, beschäftigt sich das 4. Kapitel mit den Forschungsfragen der Studie. Die weiteren Kapitel widmen sich der für die Klärung der Forschungsfragen eingesetzten Methode und den Materialien (5). Dieses Kapitel beinhaltet das Design der geplanten und durchgeführten Untersuchungen (5.1), die Untersuchungsvariablen (5.2), das Videomaterial und die Präsentation der Untersuchung (5.3), sowie die eingesetzten Untersuchungsbögen (5.4). Kapitel 6 widmet sich der Beschreibung der Untersuchungsteilnehmer (6.1) und der Durchführung der Untersuchungen (6.2) und beleuchtet die im Zweitexperiment vorgenommenen Änderungen im Vergleich zum Erstexperiment (6.3). Weiterhin werden die abgeleiteten empirischen Hypothesen (7.1) dargestellt, die Planung der Untersuchungen (7.2), Datenbereinigung und -aufbereitung (7.3) sowie die statischen Auswertungsverfahren, die dieser Arbeit zugrunde liegen, beleuchtet (7.4). Die Beschreibung der ermittelten Ergebnisse der beiden durchgeführten Untersuchungen bildet einen weiteren Teil der Arbeit (8.1 & 8.2). Eine allgemeine Diskussion und Interpretation über die zentralen Befunde sowie deren praktische

Relevanz bilden den Kern des 9. Kapitels. Die Arbeit schließt mit einem Ausblick auf weitere mögliche Forschung in diesem Kontext (10).

2 Leistungsurteile im Sport

Für den Begriff *sportliche Leistung* gibt es eine Vielzahl unterschiedlicher Auslegungen (siehe Prohl, 2003a). Der Prozess sowie das Ergebnis einer Handlung oder die Anforderungen, die an jemanden gestellt werden, können als Leistung bezeichnet werden. Der Leistungsbegriff ist im Sport ein relevanter Begriff, auch wenn er in den verschiedenartigen Feldern des Sports, wie Leistungs-, Breiten-, Freizeit-, Schul- oder Gesundheitssport, in Bedeutung und Ausprägung recht unterschiedlich ist. Güldenpfennig (1996) beschreibt verschiedene Kennzeichen sportlicher Leistung und betont den Selbstzweck, die Freiwilligkeit und die Unabhängigkeit von äußeren Notwendigkeiten. Nach Kurz (1983) gibt es drei entscheidende Faktoren, die eine sportliche Leistung definieren: Das Gütekriterium, die festgelegten Wertungsvorschriften der Sportart und der Vergleichsmaßstab[3].

Die Leistung im Gerätturnen ist beispielsweise dadurch gekennzeichnet, dass derjenige Sportler gewinnt, der die definierten Bewegungsabläufe am akkuratesten zeigen kann. Das Gütekriterium der Gestaltoptimierung wird verwendet, so dass im Mittelpunkt der Betrachtung die Schwierigkeit und kunstvolle Gestaltung der ausgeführten Bewegung steht. „Wer am Sport teilnimmt, gibt das Versprechen ab, dass er die Mitgliedschaftsregeln einhalten wird“ (Güldenpfennig, 1996, S. 187). Das sportartspezifische Regelwerk gibt die vereinbarten Verhaltensweisen aller beteiligten Akteure wieder und definiert den zugrunde gelegten Leistungsbegriff und dessen Ermittlung. Es bildet die Grundlage jedes Urteils und soll Vergleiche über mehrere Wettkämpfe und Leistungsebenen hinweg ermöglichen (2.3). Die kriteriumsbezogene Bezugsnorm stellt somit den einheitlichen Maßstab dar, wenn es um die Ermittlung der Wertung geht. Auf dieser Basis wird die Leistungsrangfolge erstellt, indem die Leistungen der Konkurrenten vergleichend hinzu kommen und der Sieger ermittelt wird. Im Wettkampf zeichnet sich der Leistungsbegriff durch die Produktorientierung aus. Lediglich das Ergebnis wird zur Ermittlung des Urteils verwendet.

[3] Rheinberg (2001) unterscheidet die individuelle, die sozial orientierte und die sachbezogene bzw. kriteriumsbezogene Bezugsnorm.

Wenn Menschen die Leistung oder auch Eignung anderer Menschen beurteilen, spricht man von *sozialen Urteilen*. Diese sind in verschiedenen Kontexten zu finden, in alltäglichen Situationen sowie im Kontext Sport. Fiedler und Bless (2002, S. 128) halten fest, „dass soziale Urteile nur teilweise durch die Reize in einer gegebenen Situation festgelegt sind […]. Urteile werden auch stark vom Vorwissen abhängen, das wir in diese Situation einbringen“. Weiterhin betonen die Autoren, dass die Interpretation einer Situation abhängig vom Vorwissen ist und dadurch die aufkommenden Fragen oder die erinnerten Aspekte abgeleitet werden. In sportlichen Wettkämpfen, aber auch in anderen Situationen müssen soziale Urteile in meist kurzen Zeitspannen erfolgen. Zusätzlich werden sie fast immer einmalig präsentiert. Durch die begrenzt zur Verfügung stehende Zeit, in der meist zahlreiche und komplexe Informationen verarbeitet werden müssen, sinkt die Möglichkeit, notwendige und wichtige Informationen für eine objektive Urteilsbildung zu berücksichtigen (ebenda). Unser Denken wird stark durch die Begrenztheit der Verarbeitungskapazität beeinflusst.

Soziale Urteile unter Zeitdruck sind charakteristisch für *Leistungsurteile im Sport*. Diese stellen einen elementaren Bestandteil dar und sind nicht wegzudenken. Unvermeidbar sind im Sportsystem daher Schieds- und Kampfrichter, die unterschiedliche Aufgaben bewältigen und dazu meist sehr wenig Zeit zur Verfügung haben. Der Schiedsrichter hat nicht die Möglichkeit eine Videoaufnahme eines Fußballspiels in aller Ruhe zu sichten und dann seine Entscheidung zu fällen. Er muss direkt vor Ort und vor allem unverzüglich im Spielverlauf entscheiden, ob der Ball im Tor war oder nicht – wie etwa beim berühmten Wembley-Tor im Weltmeisterschaftsfinale Deutschland gegen England 1966. Um Entscheidungen unter Zeitdruck möglichst objektiv und damit richtig treffen zu können, bedarf es einer guten Ausbildung der Schieds- sowie Kampfrichter. Welche Entscheidung die richtige ist, wird im Regelwerk bzw. in den Wertungsvorschriften der jeweiligen Sportart festgehalten und in regelmäßigen Abständen den aktuellen Gegebenheiten angepasst. In allen Wettkampfsportarten sind menschliche Urteile involviert. In einigen wird die Leistungsbewertung der Athleten ausschließlich durch ein solches Urteil getroffen. Gerade in diesen Sportarten wird oftmals die Objektivität des vom Unparteiischen getroffenen Urteils angezweifelt.

Wie hoch die Anforderungen an den Beobachter sind, zeigen Fehlurteile im Wettkampf (Neumaier, 1988). Auch bei Experten auf diesem Gebiet, kann die begrenzt zur Verfügung stehende Verarbeitungskapazität

des Gedächtnisses ein Grund hierfür sein. Deshalb stellt gerade bei dem Personenkreis der Schieds- und Kampfrichter der Prozess der sozialen Informationsverarbeitung bis hin zur Urteilsbildung und den daraus resultierenden Urteilen und Entscheidungen eine interessante Thematik dar (3).

2.1 Die sportliche Leistungsbeurteilung und ihre Gütekriterien

Um eine sportliche Bewegung beurteilen zu können, muss der Beurteiler sensorische und kognitive Voraussetzungen erfüllen. Zudem erfolgt der Urteilsprozess stufenweise und beinhaltet die Wahrnehmung, die Kategorisierung, Gedächtnisprozesse und Urteils- und Entscheidungsprozesse (Fiedler & Bless, 2002) (3.1). Auf der ersten Stufe, der Wahrnehmung, muss die Fähigkeit des Bewegungssehens vorhanden sein. Das Bewegungssehen stellt die visuelle Wahrnehmung von Bewegungen dar und kann unbewusst und ohne spezifisches Ziel geschehen. Die Bewegungsbeobachtung stellt die Voraussetzung für eine Leistungsbeurteilung dar und kann als „das absichtliche, aufmerksamselektive visuelle Wahrnehmen von fremden Bewegungsabläufen mit dem Ziel, Ausführungsmerkmale in ihrer Ausprägung zu erfassen“ (Neumaier, 1988, S. 29), definiert werden. Daraufhin wird die wahrgenommene Beobachtung kategorisiert. Die Ausprägung einzelner Ausführungsmerkmale oder des gesamten Bewegungsablaufs werden mit repräsentierten Bewegungsnormen und definierten Kriterien verglichen und schließlich verbalen oder numerischen Kategorien, meist Punktwertungen, zugeordnet. Der Vergleichsprozess setzt voraus, dass es einheitliche Normen gibt, die bekannt sind. Außerdem sollte eine genaue Vorstellung der normgerecht optimal ausgeführten Bewegung vorliegen. Die Kriterien definieren die besondere Beachtung von Teilen der Bewegung und geben die optimale Bewegungsausführung vor (2.3). Das Resultat des Vergleichsprozesses wird in Form von Punkten ausgedrückt und ergibt die Bewertungsziffer. Das Ziel dieses Beurteilungsprozesses ist die objektive Bewertung anhand einer Bewertungsziffer (Thomas, 1978) und kann somit als Bewegungsbeurteilung bezeichnet werden. Die Bewegungsbeurteilung wird im Lernprozess und im Techniktraining durch eine differenzierte Rückmeldung zur Verbesserung der Bewegungsqualität eingesetzt und fungiert im sportlichen Wettkampf durch die interindividuelle Differenzierung als Mittel der Leistungsermittlung (Neumaier, 1988, S. 204).

Die Leistungsbeurteilung kann als Überbegriff verstanden werden, der sich in die beiden Stufen Leistungsfeststellung und Leistungsbewertung

unterteilen lässt (Jürgens, 2005). Beide Stufen sind notwendig, um die Leistung einzuschätzen. Die erste Stufe verhilft dem Beurteiler dazu, festzustellen, ob die geforderte Leistung beherrscht wird oder nicht, während die zweite Stufe diese Leistung in ein verständliches System einstuft und damit vergleichbar mit anderen Leistungen macht.

In ähnlicher Weise werden die Begriffe des Wertens und der Bewertung verwendet. Umgangssprachlich werden sie oftmals synonym benützt. Nach Lutter (1982, S. 97) stellt das Werten eine andere Art der quantitativen Leistungsfeststellung dar. Im Gerätturnen[4] werden unter anderem die Schwierigkeit und die Haltung gewertet. Die Aufgabe des Kampfrichters besteht darin, anhand von Wertungspunkten die Bewegungsfolge zu quantifizieren. Diese Punktevergabe stellt die Wertung dar. Das Bewerten hingegen drückt die Zuschreibung des ermittelten Wertes aus. Setzt man die beobachtete Leistung ins Verhältnis zu Altersgenossen oder einem anderen Vergleichsmaßstab, wird den ermittelten Daten ein Bedeutungsgehalt beigemessen (ebenda). Dieser Vergleich der erzielten Leistung relativiert diese und ermöglicht dessen Bewertung als gute oder eher schlechte Leistung.

Im meist pädagogischen Kontext verwendet man anstelle des Wertungs-Begriffs den der Leistungsmessung (Jürgens, 2005, S. 135). Dabei wird nicht die Messung im physikalischen Sinne[5] verstanden, sondern lediglich die systematische Sammlung von Informationen, die den aktuellen Leistungsstand beschreiben.

Der Begriff der Messung wird auch in unterschiedlichen Definitionen des Begriffs ‚sportliche Leistungsbewertung' verwendet. Im Zusammenhang damit beruht die Unterscheidung auf Sportartengruppen. Auf Basis der Ermittlung der sportlichen Leistungsbewertung charakterisiert Prohl (2003b) vier Sportartengruppen. Die erste Gruppe stellt die sogenannten c-g-s-Sportarten dar. Diese Sportarten, wie Leichtathletik oder Gewichtheben, können durch direkte *Messung* unmittelbar quantifiziert werden. Damit werden den Wettkampf überdauernde Leistungsvergleiche möglich, wie etwa Rekorde. Die zweite Gruppe, die Sportspiele, führen mittels Zählung von Toren bzw. Punkten zur Ermittlung des

[4] Der im Volksmund als ‚Kunstturnen' (Übersetzung der engl. Bezeichnung „artistic gymnastics") bezeichnete olympische Spitzensport wird inzwischen vom Deutschen Turner-Bund als ‚Gerätturnen' benannt. ‚Gerätturnen' ist eine Sammelbezeichnung für das Turnen an Geräten als Leistungs-, Breiten- und Freizeitsport (Deutscher Turnerbund, 2009).

[5] Eine Größe zu messen heißt im physikalischen Verständnis, das zu messende Objekt mit der Maßeinheit, unter Verwendung des MKS-Systems, zu vergleichen.

Siegers. Technokompositorische oder auch technisch-kompositorische Sportarten, wie das Gerätturnen oder der Eiskunstlauf, werden „durch die Zuweisung von Punktzahlen zur graduellen Abstufung der Qualität von Bewegungsausführungen“ entschieden und bilden die dritte Sportartengruppe (Prohl, 2003b, S. 337). Die vierte und letzte Gruppe beschreibt er als Sportarten, die durch eine Kombination der genannten Verfahren festgelegt werden. Beispielsweise ist das Skispringen eine Kombination aus Messen und Punkten, während bei Kampfsportarten vor allem die Erzwingung einer definierten körperlichen Position des Gegners, wie der ‚knock out' im Boxsport oder der Schultersieg im Ringen, für die Leistungsbewertung ausschlaggebend ist (ebenda).

Eine andere Strukturierung nimmt Göhner (1992) vor, der sich eingehend mit dem Thema der sportspezifischen Bewegungsaufgabe beschäftigt. Er unterscheidet fünf Komponenten, die eine Sportart beschreiben. Das Bewegungsziel, die -regeln, das -objekt, das -subjekt und der Bewegungsraum werden von ihm herausgestellt und dienen der detaillierten Beschreibung und Klassifizierung von Sportarten. Die Bewegungsziele werden weiterhin unterteilt in Vergleichs- und Erreichungsziele. Diese werden einerseits als Movendumziele mit wettbewerblicher Orientierung und andererseits als Bewegerziele, die nicht die äußerlich beobachtbare Rangordnung im Vordergrund haben, sondern die motorische Belastbarkeit, die psychische Befindlichkeit oder pädagogische Ziele des Sporttreibenden (Göhner, 1992, S. 49). Unter den Movendumzielen ordnet er den Vergleichszielen sechs unterschiedliche Ausdifferenzierungen unter, wie beispielsweise die Zeitminimierung (Schwimmen), die Trefferoptimierung (Fechten) oder die Fehlerminimierung (Gerätturnen) und benennt dazu jeweils Sportarten. Vergleichbar werden die vier Unterklassen der Erreichungsziele strukturiert, indem unter anderem Erhaltungsziele (Surfen) oder Formziele (Jazztanz) eingeordnet werden.

Bergholz (2003) kategorisiert die Sportarten in ähnlicher Weise wie Prohl und benennt jeweils eine Bewertungskategorie, die entweder quantitativ, qualitativ oder als eine Art Mischform vorkommt. Als quantitativ bezeichnet er die Gruppe, die die c-g-s-Sportarten (nach Prohl, 2003b) beinhaltet, wie beispielsweise Volleyball oder das Schießen. Seine zweite Bewertungskategorie entspricht der dritten Kategorie von Prohl (2003b), vereint technisch-kompositorische Sportarten und Sportarten mit technischen Wertungen wie Judo und Ringen. Weiterhin folgen zwei quantitativ-qualitative Kategorien.

> „Wettkampfleistungen sind bei der überwiegenden Mehrzahl der bekannteren Sportarten durch eine direkte Messung (unter Verwendung des MKS-Systems[6]) oder durch Trefferauszählung unmittelbar quantifizierbar. … Bei verschiedenen Sportarten kann nicht mehr von Messung im engeren Sinn, d.h. im physikalischen Verständnis gesprochen werden. Dort wird stattdessen die Ausführungsqualität der gezeigten Bewegungen von Kampfrichtern bewertet“ Bergholz (2003, S. 63).

Damit illustriert der Autor die Möglichkeit quantitativer und damit objektiver Messung der Sportarten der ersten Kategorie. Unbeachtet bleibt, dass auch in dieser Sportartengruppe Untersuchungen zu Fehlentscheidungen vorliegen (3) und somit ein Indiz für eine subjektive Leistungsermittlung auch hier gegeben ist. Laut Thomas (1978, S. 263) sind diese Probleme der Leistungsbeurteilung allerdings vielmehr messtechnischer Natur und entstehen allenfalls auf simplen Beobachtungsfehlern. Weiterhin führt er aus, dass die Beurteilungsleistung des Menschen hierbei eine eher untergeordnete Rolle spielt. Eine klare Gegenmeinung nehmen die Autoren ein, die, meist durch Untersuchungen im Fußball, belegen, dass vielfältige systematische Fehler auftreten können (u.a. Mascarenhas, O´Hare & Plessner, 2006). Somit soll im Rahmen dieser Arbeit kein bewertender Vergleich der Sportarten oder der Anforderungen der Beurteilungssituation erfolgen. Der Fokus dieser Arbeit liegt auf den technisch-kompositorischen Sportarten, die ausschließlich durch menschliche Urteile bewertet werden.

Gerade in diesem Zusammenhang stellt sich die Frage nach den Gütekriterien der getroffenen Urteile. Die Bewertungsverfahren der unterschiedlichen Sportarten werden fortwährend weiterentwickelt und verfolgen das Ziel, Fehlerquellen zu minimieren und die Objektivität zu erhöhen. Die Objektivität ist nichts Absolutes, sondern muss in Relation zu den Urteilen anderer Beurteiler gesehen werden (Thomas, 1978). Wenn aber alle Beurteiler gleichen Fehlern unterliegen, gehen diese, im absoluten Sinn, unidentifiziert als solche in das Urteil ein. Wie objektiv, im Sinne von frei von subjektiven Einflüssen, sind die Urteile der Unparteiischen?

Deutlich formuliert Kurz bereits 1983 (S. 64) den subjektiven Charakter sportlicher Leistungsbewertung: „Leistungen im Sport sind weder objektiv noch absolut!“. Um sportliche Leistungen, die keine direkte Auseinandersetzung der Sportler beinhalten, ermitteln und vergleichen zu

---

[6] System, das auf der Messung von Weite/Höhe (Meter), Gewicht (Kilogramm) und Zeit (Sekunde) beruht.

können, ist eine Quantifizierung der Leistungsergebnisse vonnöten. Quantifizierung bedeutet eine Messung im weitesten Sinne und somit die Zuordnung von Zahlen zu Merkmalen, Eigenschaften oder Ähnlichem des Beobachtungsinteresses. Die erbrachte Bewegungsleistung muss möglichst objektiv in Zahlen ausgedrückt werden, um eine Leistungsrangfolge zu bilden, die die Ermittlung des Siegers im Vergleich zu den Konkurrenten zur Folge hat (Thomas, 1978, S. 261).

> „So wird im Bereich des Kunstturnens versucht, durch exakte Vorschriften (die allerdings in ihrer Komplexität nur dem Experten verständlich sind) ein Höchstmaß an Objektivität in der Benotung zu erreichen: Trotzdem lässt sich eine subjektive Bewertung nicht ausschließen“ (Lutter, 1982, S. 99).

Diese Ambivalenz von Objektivität und Subjektivität zeigt sich immer wieder, wenn es darum geht, Leistungen im Sport zu erfassen und zu bewerten. Oftmals entstehen unterschiedliche Meinungen darüber, welches Urteil das richtige sei. Wie unterschiedlich ein und dieselbe sportliche Leistung bewertet wird, wird deutlich, wenn man sich die *unterschiedlichen anwesenden Bewertungsgruppen* vor Augen hält.

Nicht selten entstehen hitzige Diskussionen darüber, wer warum die höchste Wertung in einem Wettkampf bekommen hat. Unparteiische, Trainer, Athleten, Zuschauer und Funktionäre vertreten unterschiedliche Meinungen und präsentieren den durch das getroffene Urteil der Schieds- und Kampfrichter entstanden Unmut auf unterschiedliche Art und Weise. Jede Beurteilungsgruppe für sich meint, unter *rationalen Gesichtspunkten* – entsprechend den definierten Normen der Sportart – die Entscheidung über den Sieger für sich getroffen zu haben. Auf dieser Basis kann ermittelt werden, was korrekt ist und was nicht. Doch die psychologische Entscheidungsforschung geht davon aus, dass derartige Urteile nicht wirklich rational gefällt werden, sondern fast ausschließlich unter suboptimalen Bedingungen stattfinden und in irgendeiner Art und Weise verzerrt werden (Fischer, Greitemeyer & Frey, 2006).

Je komplexer das zu fällende Urteil, desto größer ist die Wahrscheinlichkeit für das Zustandekommen eines nicht korrekten Urteils. Betrachtet man also die Aufgabe des Schieds- bzw. Kampfrichters, lässt sich eine Variation innerhalb der unterschiedlichen Sportarten von relativ einfachen Urteilen bis hin zu *komplexen Urteilen* (2.4), beispielsweise in technisch-kompositorischen Sportarten, bei der Identifikation von Fehlern im Gesamtverlauf der Bewegung (Neumaier, 1988, S. 390), erkennen. „Man muss sich also immer dann des Messinstruments

Mensch bedienen, wenn es um die Erfassung verhältnismäßig komplexer sportlicher Leistungen geht“ (Haase, 1972, S. 346). Weiterhin muss die „Funktionsfähigkeit des Menschen als Messinstrument“ ausreichen, um fehlerfrei alle relevanten Leistungsmerkmale zu erfassen. Fehlurteile in sportlichen Wettkämpfen verdeutlichen die hohen Anforderungen an den Beobachter bei der Beurteilung von Bewegungen. Es ist anzunehmen, dass nicht motivationale Gründe dafür verantwortlich sind, „sondern [Fehlurteile] häufig aus einer Überforderung des Beobachters resultieren“ (Neumaier, 1988, S. 225). Hinzu kommt die Einmaligkeit und Nicht-Wiederholbarkeit des Ereignisses (Thomas, 1978). Derartige Entscheidungen werden als Tatsachenentscheidungen bezeichnet und wurden in der Vergangenheit nicht selten durch Aufzeichnungen in Frage gestellt. Auf der Grundlage von Aufzeichnungen soll, generell laut Wettkampfregeln, dem Unparteiischen Recht gegeben werden. Begründet wird diese Haltung dadurch, dass „der Aufnahmewinkel des elektronischen Mittels in der Regel anders ist als der Blickwinkel des Schiedsrichters, der darüber hinaus erheblich näher am Handlungsort ist als die Kamera“ (Thieß & Tschiene, 1999, S. 43). Das menschliche Auge und Gehirn sind dermaßen komplex und leistungsfähig, dass man nicht zwangsläufig sagen muss, es sei schlechter als die technischen Hilfsmittel. Dennoch gibt es keine Garantie dafür, dass die Dinge, die in der Außenwelt vor sich gehen, auch von zwei oder mehr Individuen in identischer Weise wahrgenommen werden (3). Nach Neumaier (1988, S. 211) ist

> „eine objektive Beurteilung … nur dann möglich, wenn die sinnesphysiologisch und psychologisch bedingten Anforderungen der einzelnen Elemente sowie deren gleichzeitige Beachtung und Verknüpfung im Gesamturteil keine Überforderung der Wahrnehmungsfähigkeit des Beurteilers darstellen“.

Durch die Zielsetzung, ein „Höchstmaß an Objektivität“ (Lutter, 1982, S. 99) bei der Beurteilung sportlicherer Leistung zu erbringen, ergibt sich ein gewisser Qualitätsanspruch. Selbst erfahrene Kampfrichter zeigen eine hohe Varianz der Noten (Wilson, 1976a & b). In der Wettkampfsituation entstehen Streuungen der Wertungen durch situative und subjektive Einflüsse (Bard, Fleury, Carrière & Hallé, 1980). Ein situativer Einfluss stellt beispielsweise die Wettkampfatmosphäre dar, während unter einem subjektiven Einfluss die menschlichen Wahrnehmungsgrenzen bei hoher Bewegungsgeschwindigkeit verstanden werden (Taha, Osman & Ehlerz, 1991). Es besteht ein Unterschied zwischen erfahrenen und unerfahrenen Personen (Bard et al., 1980; Plessner & Schallies, 2005; Ste-Marie, 1999, 2000), systematische Urteilsfehler

lassen sich aber auch bei Experten finden (Oudejans, Verheijen, Bakker, Gerrits, Steinbrückner & Beek, 2000; Plessner & Schallies, 2005; Scheer, 1973; Ste-Marie & Lee, 1991; Wilson, 1976a & b).

Die *Gütekriterien* der klassischen Testtheorie beschäftigen sich mit der Qualitätsbeurteilung von Messungen und der Abschätzung von Messfehlern (Bortz & Döring, 2003). Übertragen auf die Thematik dieser Arbeit lassen sich im Folgenden die Gütekriterien als Grundlage zur Einschätzung der Beurteilungssituation im sportlichen Wettkampf der technisch-kompositorischen Sportart Gerätturnen erläutern. Die Leistungsmessung soll sich an den Haupt-Gütekriterien der Objektivität, Reliabilität und der Validität orientieren (Jürgens, 2005).

Unter *Objektivität* ist die Unabhängigkeit vom Beurteiler gemeint (Bortz & Döring, 2003). Im konkreten Fall bedeutet das, das die Wertung einer gezeigten Übung in einem Wettkampf immer dieselbe sein müsste, unabhängig davon, welcher Kampfrichter diese wertet (Wilson, 1976b). Eine wichtige Grundlage für die objektive Beurteilung ist die präzise Definition der zu erhebenden Variablen, also der Kriterien, die beachtet werden.

Die Durchführungsobjektivität erfasst die Vereinheitlichung und Reglementierung der Bedingungen, unter denen die sportliche Leistung ermittelt wird. In den Wertungsvorschriften sind diese Bedingungen, wie beispielsweise die Zeitdauer einer Übung, genau festgelegt und die ‚Aufgabenstellung' damit genau definiert.

Die Auswertungsobjektivität ist gegeben, wenn die Vergabe von Testpunkten vom Auswerter unbeeinflusst ist (ebenda). Die Punktevergabe bzw. Vergabe von Abzügen sollte immer identisch durchgeführt werden und damit klar sein, wie viele Punkte bzw. Abzüge für welches Element bzw. Fehler zu vergeben sind. Auch das ist in den Wertungsvorschriften vorgegeben. Die Verinnerlichung – das deklarative Wissen – und die sichere Handhabung – das prozedurale Wissen – dieser Vorgaben, soll durch eine gute Ausbildung sichergestellt werden, birgt aber Fehlerquellen. Das deklarative Wissen stellt das reine Wissen über den Inhalt der Wertungsvorschriften und wird in der Kampfrichterausbildung erlernt (Plessner, 2001a). Die Kampfrichter sollten diese ‚blind' beherrschen, da Wissenslücken zu Fehlurteilen führen. Das prozedurale Wissen birgt ebenso eine Fehlerquelle und schmälert die Objektivität des Urteils. Die korrekte Anwendung des Wissens in der realen Situation des Wertens benötigt viel Übung, da das Erkennen von Elementen automatisch und möglichst schnell ablaufen sollte (ebenda). Sind weiterhin die Kriterien zu allgemein, zu wenig eindeutig oder unterschiedlich

gewichtet, führt das automatisch dazu, dass der subjektive Einfluss des Kampfrichters in das Urteil eingeht und dadurch die Wertungen stärker streuen. So stellte zum Beispiel die Möglichkeit der Vergabe von 0,05 Punkte (P), anstelle der üblichen 0,1P, um eine bessere Differenzierung zwischen den Übungen zeigen zu können, einen unnötigen Interpretationsspielraum des Kampfrichters und führte damit zum größeren Einfluss subjektiver Eindrücke (ebenda). Diese Möglichkeit besteht seit dem CdP des Jahres 2006 nicht mehr und ist ein Beispiel für die ständige positive Anpassung der Wertungsvorschriften.

Die Interpretationsobjektivität, als letzte Unterform der Objektivität, beinhaltet die abschließende Bewertung. Anhand vorhandener Normen, im speziellen Fall die Alters- oder Geschlechtsgenossen und nicht Rekorde wie in anderen Sportarten, werden die Wertungen verglichen und der Gewinner ermittelt. Somit gibt es keinen absoluten, sondern nur einen relativen Vergleich, indem bei der endgültigen Beurteilung der Leistung die soziale Bezugsnorm zum Einsatz kommt. Ein Athlet kann mit einer 14 Punkte-Übung einen Wettkampf gewinnen, da alle anderen weniger Punkte haben und im nächsten Wettkampf mit dieser Wertung nicht einmal ins Finale eines Geräts kommen.

Im Zusammenhang mit der Erhöhung der Objektivität gibt es unterschiedliche Vorschläge, welche Maßnahmen ergriffen werden müssen, je nachdem, welche Fehlerquelle betrachtet wird. Gefordert wird allgemein die Ausschaltung subjektiver Einflüsse auf Seiten der Beurteiler, da diese zu Beurteilungsfehlern führen können (Jürgens, 2005). Eine generelle Möglichkeit, die Objektivität der Kampfrichterurteile zu erhöhen, besteht darin, eine wiederholte Sichtung und eine zeitliche Dehnung der Urteilssituation mit Hilfe von Videoaufnahmen zu schaffen (Bard et al., 1980; Johnson, 1971; Neumaier, 1988; Plessner & Haar, 2006). Die Videomethode verhilft den Kampfrichtern dazu, mehr Fehler in den Übungen zu erkennen, soll allerdings nicht generell als das Mittel der Evaluation eingesetzt werden, so O´Brien (1991). Die situativen und subjektiven Einflüsse der Realsituation können anhand von Videoaufnahmen reduziert werden und führen sowohl zwischen den Kampfrichtern als auch zwischen zwei Messzeitpunkten zu homogeneren Urteilen (Taha et al., 1991). Der notwendige Zeitaufwand, die unterschiedlichen Blickperspektiven von Videoaufnahme und Kampfrichter sowie die Tatsachenentscheidung werden als Gründe genannt, die gegen einen Videoeinsatz sprechen (Puhl, 1980; Taha et al., 1991). Der Videoeinsatz ist in den Wertungsvorschriften als Kontrollmöglichkeit

festgeschrieben worden. Eingesetzt werden die Aufnahmen nur in strittigen Fällen (2.3).

Eine weitere Möglichkeit zur Reduzierung der Überforderung der Informationsverarbeitung besteht in einer Partialisierung von Kampf- und Schiedsrichteraufgaben (Plessner & Haar, 2006). Die Wertungsvorschriften sind dieser Forderung nachgegangen (2.3). Ob allerdings die Neuerung der Aufteilung der Kampfrichter in zwei Kampfgerichte ausreicht, um die Überforderungssituation zu minimieren oder gar zu verhindern, muss kritisch untersucht werden.

Objektivität bedeutet im Kontext der aufgegriffenen Thematik die Übereinstimmung der Beurteilung durch verschiedene Personen (Landers, 1970; Scheer, 1973; Wilson, 1976b, 1977). Da allerdings alle Beurteiler gleichen Urteilsfehlern unterliegen können, „erlaubt der Grad an Übereinstimmung keine Aussagen über die wahre und einzig richtige Zuordnung von beobachteter Leistung und Punktzahl in einem Wettkampf“ (Neumaier, 1988, S. 226). Begründet kann diese Meinung dadurch werden, dass systematische Fehler zu hohen Reliabilitätswerten führen können. Lediglich unsystematische Fehler lassen sich in derartigen Ergebnissen nachweisen.

Die Notwendigkeit von Reliabilitäts- und aufwendigen Validitätsprüfungen ist im Rahmen der Beobachtung sportlicher Leistung eher unklar. Die *Reliabilität* beschreibt die Genauigkeit der Messung und wird oftmals unter dem Begriff der Objektivität oder der Übereinstimmung von Kampfrichtern berichtet (Wilson, 1976a). Unter gleichen Rahmenbedingungen sollte bei einem reliablen Test eine exakte Reproduzierbarkeit des Ergebnisses möglich sein. Mögliche Methoden stellen der Retest, der Paralleltest, die Testhalbierung und die interne Konsistenz dar (Bortz & Döring, 2003). Bei der Retest-Methode, als einzige anwendbare Methode in diesem Zusammenhang, wird derselbe Test derselben Stichprobe zweimal vorgelegt (Wilson, 1976a & b).

In Untersuchungen zur Reliabilität von Kampfrichterurteilen im Turnen werden immer wieder Unterschiede bezüglich der bewerteten Geräte berichtet. Insgesamt zeigen Untersuchungen moderate bis hohe Reliabilität an allen Geräten, vor allem an Boden, Barren und Ringen bei den Männern und Stufenbarren bei den Frauen, nicht aber am Sprung (Wilson, 1976a & b).

Die *Validität* bezieht sich auf die Frage, ob das, was man messen wollte, auch gemessen wurde. Die Unterpunkte sind die Inhalts-, die Kriteriums-, die Konstruktvalidität und die Testfairness. Die Inhaltsvalidität,

auch als augenscheinliche Validität bezeichnet, wird nur durch logisches Denken, nicht durch Techniken feststellbar (Bortz & Döring, 2003). Sie kann als gegeben angesehen werden, wenn der Inhalt der Test-Items das Zielmerkmal hinreichend genau definiert. Im konkreten Fall: Lässt sich mithilfe der Wertungsvorschriften die sportliche Leistung im Gerätturnen erfassen? Dieser Frage kann im Rahmen der Arbeit nicht nachgegangen werden.

Unterpunkte der Kriteriumsvalidität sind die Übereinstimmungs- und die Vorhersagevalidität. Die Übereinstimmung eines Messinstrumentes mit anderen relevanten Merkmalen, sogenannten Außenkriterien, spiegelt die Übereinstimmungsvalidität wider (ebenda). Im Sport soll die sportliche Leistung beurteilt werden. Das Training stellt zwar ebenfalls eine Situation dar, in der die Leistung überprüft werden könnte, entspricht aber der Sinngebung des ‚Wettkampf'-Sports nicht. Um die Validität zu überprüfen hat Wilson (1976a & b) die Real-Wertungen eines Wettkampfs mit entsprechenden Video-Wertungen verglichen. Vermutet werden kann, dass mit dieser Methode die situativen Einflüsse des Wettkampfs minimiert, die subjektiven systematischen Fehlerquellen aber nicht ausgeschlossen werden können.

Die Vorhersagevalidität ist gegeben, wenn die Ergebnisse Prognosen für zukünftige Ergebnisse ableiten lassen. Die individuelle Leistungsentwicklung der Sportler ist durch Faktoren wie Verletzungen, Trainingsintensität und –umfang sowie durch Motivation und viele weitere Faktoren bestimmt und führt daher eher zu ungenauen Aussagen.

Die Konstruktvalidität ist gegeben, wenn die gemessenen Eigenschaften mit einem theoretischen Modell übereinstimmen. Im sportpsychologischen Bereich werden allerdings meist Konstrukte gemessen, die nicht direkt erfassbar sind. Es liegen meist mehrere Theorien vor, die anhand unterschiedlicher Determinanten versuchen, das entsprechende Phänomen zu beschreiben und zu erklären. Die sportliche Leistung müsste detailliert erklärt und mit den Richtlinien der Sportart verglichen werden. Dieses Vorgehen ist noch nicht hinreichend untersucht und auch nicht Thema dieser Arbeit.

Als letzten Unterpunkt der Validität ist die Testfairness zu nennen, die besagt, dass die Leistungsbeurteilung nicht durch die Zugehörigkeit zu einer bestimmten Gruppe beeinflusst wird, wie beispielsweise das Geschlecht oder die soziale Schicht.

Zusammenfassend kann gesagt werden, dass die Güte der Beurteilung sportlicher Leistungen stark durch den Faktor Mensch und die entsprechenden Rahmenbedingungen beeinflusst ist. Eine objektive, reliable und valide Bestimmung, im absoluten Sinne, ist in diesem Zusammenhang eher unrealistisch. Umso mehr scheinen die urteilenden Personen des Handlungsfelds Sport interessant für weitere Forschung zu sein.

2.2 Kampf- und Schiedsrichter und ihre Entscheidungen

Im sportlichen Kontext, mit seinen verschiedenartigen Sportarten, lassen sich auch unterschiedliche Bezeichnungen für den Unparteiischen erkennen. Am häufigsten findet man aber die beiden Begrifflichkeiten ‚Schiedsrichter' und ‚Kampf- oder Wertungsrichter'. Je nach Sportart und Funktion werden weitere Bezeichnungen verwendet, wie der Punkt- oder Preisrichter, aber auch Starter, Wende- oder Linienrichter. Für Bezeichnungen, die eine Funktion beschreiben, wie beispielsweise beim Starter, kommt die Frage nach dem ‚warum' nicht schnell auf. Aber bei der Definition und Unterscheidung des Schieds- und des Kampfrichters gibt es entsprechende Auslegungen.

Die offiziellen internationalen Wertungsvorschriften im Gerätturnen benennen die Unparteiischen in der deutschen Version als Kampfrichter (2.3). Somit wird in der vorliegenden Arbeit derselbe Titel für das Gerätturnen und verwandte technisch-kompositorische Sportarten verwendet.

Teipel (2003b, S. 460) beschränkt sich bei der Beschreibung des *Schiedsrichters* auf die Sportartengruppe der Sportspiele. Der Schiedsrichter verantwortet die Einhaltung der Spielregeln.

> „Die Rechte und Pflichten eines Schiedsrichters … sind in der jeweiligen Spielordnung niedergelegt. Für die Tätigkeit … werden eine differenzierte Kenntnis der Spielregeln, eine gute Schnelligkeit und hohe Genauigkeit in der Wahrnehmung, eine ausgeprägte Entscheidungsfreudigkeit, Unvoreingenommenheit und Souveränität im Auftreten als wesentliche Voraussetzung angesehen. … Bei den Entscheidungen … handelt es sich um Tatsachenentscheidungen, die nach bestem Wissen und Gewissen gefällt werden. Nur gravierende Fehlentscheidungen … können evtl. durch die sportspielspezifische Sportgerichtsbarkeit verändert werden".

Wenn er hingegen *Kampfrichter* beschreibt, bezieht er sich auf technisch-kompositorische Sportarten, Kampfsportarten, aber auch auf die Sportart Leichtathletik und stellt die kennzeichnende Funktion als Unterscheidungsmerkmal heraus. Der Kampfrichter, als der Leiter von spezifischen Wettkämpfen, hat

> „die Aufgabe, als Unparteiischer die Einhaltung der sportartspezifischen Wettkampfregeln zu überwachen, Strafen zu verhängen sowie in Zweifelsfällen Entscheidungen über Gelingen oder Misslingen von Aktionen zu treffen. Er übernimmt in Abhängigkeit von der jeweiligen Sportart spezifische Funktionen, das Bewerten von Leistungen (z.B. im Turnen, in der Rhythmischen Sportgymnastik), die Vergabe von Punkten (z.B. im Judo, im Boxen) und das Messen von Leistungen (z.B. in der Leichtathletik)“ Teipel (2003a, S. 287).

Der Unparteiische, ob Schieds- oder Kampfrichter, hat die Überwachung und Einhaltung der Regeln bei einer Sportveranstaltung sicherzustellen. Bezüglich der verwendeten Bezeichnung für die Unparteiischen schlagen Brand und Ness (2004) vor, zwischen Urteilssituationen in Sportspielen und anderen Sportarten zu unterscheiden. Ähnlich wie bereits bei Teipel (2003a & b) zu lesen, sollte nach Brand und Ness (2004, S. 130) „in (zukünftigen) Untersuchungen begrifflich wie inhaltlich konsequent zwischen Schiedsrichtern in Sportspielen und Punkt- oder Kampfrichtern in anderen Sportarten unterschieden werden“. Die inhaltliche Begründung liegt dabei in der Aufgabe der Unparteiischen. Schiedsrichter beurteilen Situationen, und damit sportliche Leistungen, die „in Sportspielen prozesshaft und in direkter Interaktion zwischen wettstreitenden Parteien“ (ebenda) stattfinden. „Schiedsrichter greifen fortwährend bzw. mehrfach in ein sich entwickelndes Gesamtspielgeschehen ein“ (ebenda), während Punkt- oder Kampfrichter nicht in das sportliche Geschehen eingreifen. In von Punkt- oder Kampfrichtern zu beurteilenden Situationen herrscht keine direkte Interaktion zwischen Konkurrenten vor, sondern es kommt zu einer Darstellung der sportlichen Leistung. Somit haben Punkt- oder Kampfrichter nach Brand und Ness (2004) keine prozesssteuernde Funktion. Unberücksichtigt bleibt allerdings, in der von den Autoren vorgeschlagenen inhaltlichen Zuordnung der Begrifflichkeiten, die direkte Konkurrenteninteraktion bei Kampfsportarten, wie beispielsweise Taekwondo oder Judo. Sie widerspricht dem Kriterium, ein Kampf- oder Punktrichter hätte keine prozesssteuernde Funktion in einem sportlichen Wettkampf.

Kampf- und Schiedsrichter müssen *immer eine Entscheidung* treffen. Unabhängig davon, ob sich diese als eine einfache, klare oder aber als eine unklare oder schwierige Aufgabe herausstellt. Auch wenn ein Bewegungsablauf mit hoher Ereignisdichte und Geschwindigkeit nicht detailliert genug erfasst werden kann, muss ein Urteil resultieren. Die Schwierigkeit liegt darin, nach einmaliger Sichtung, schnell das richtige Urteil zu fällen, auch wenn zu wenige Informationen vorliegen (Messner & Schmid, 2007) (3.1). Gute Leistungen oder fehlerfreie

Entscheidungen belassen den Urteilenden unauffällig im Hintergrund. Souveräne Leistungen werden somit nicht positiv hervorgehoben oder gelobt. Vermeidliche oder tatsächliche Fehlentscheidungen hingegen lassen seine generelle Entscheidungs- und Handlungsfähigkeit schnell in die Kritik geraten.

Entsprechend dem Lehrbuch für Schiedsrichter (Ebersberger, Malka & Pohler, 1996, S. 96), sollen die Unparteiischen nicht zu empfindlich auf Kritik reagieren, konsequent sein und sich zu guter Letzt nicht beeinflussen lassen. Nicht selten stehen Fehlentscheidungen von Schieds- oder Kampfrichtern im Mittelpunkt von Diskussionen. Besonders im Gedächtnis bleiben Skandale, wie beispielsweise der des Fußball-Schiedsrichters Robert Hoyzer[7] oder etwa der Gerätturn-Wertungs-Skandal der Olympischen Spiele in Athen 2004, der zu Beginn der Arbeit dargestellt wurde.

Aber auch bei zweifelsfrei richtigen Entscheidungen des Kampfrichters kommt es zu Unmut auf Seiten von Athleten und Trainern, wie der aktuellste Fall der Olympischen Spiele 2008 in Peking demonstriert. Ein „Ausraster beim Taekwondo“, so ein Bericht der Deutschen Presseagentur, beschreibt ein solches Szenario. Der sich im Halbfinale in Führung befindende Athlet benötigt verletzungsbedingt eine Pause, die er überzieht und daraufhin der Kampfrichter den Kampf als verloren erklärt. Der Vorfall, der zu einer lebenslangen Sperre führt, entwickelt sich aus aufkeimenden Tumulten über die Entscheidung des Kampfrichters, woraufhin der Athlet gegen den Kopf des Hauptkampfrichters tritt.

Falsche Urteile sind nicht gleich falsche Urteile. Kampf- und Schiedsrichtern wird oftmals vorgeworfen, vorsätzlich für oder gegen jemanden zu entscheiden. Wissenschaftliche Studien belegen, dass es Unparteiischen schwer fällt, vorhandene Parteilichkeiten nicht in ihre Urteile und Entscheidungen einfließen zu lassen (u.a. Ansorge & Scheer, 1988; Mohr & Larsen, 1998). Kritiker der Entscheidungen von Schieds- und Kampfrichtern bedenken meist nicht, dass es sich bei Fehlurteilen oftmals nicht um motivational gesteuerte Fehlentscheidungen, sondern um *funktionale Gründe für Verzerrungen* von Urteilen handelt (3). Im Sport ist es in den meisten Fällen möglich, durch Videoanalysen zu

[7] In der Saison 2004/2005 manipulierte der damals 25-jährige Fußball-Schiedsrichter Robert Hoyzer nachweislich Regionalliga- und DFB-Spiele, indem er Mannschaften zum Sieg verhalf, auf deren Sieg er und Eingeweihte bei einer Wettlotterie horrende Summen gesetzt hatte.

ermitteln, welche die objektiv richtige Entscheidung ist. Allerdings erlaubt in erster Linie der dafür notwendige Zeitaufwand oftmals eine derartige Überprüfung nicht.

Diese mehr oder weniger folgenschweren systematischen Urteilsverzerrungen entstehen aufgrund unterschiedlichster Ursachen, wie beispielsweise dem Zeitdruck, der Reihenfolge der Informationspräsentation oder aufgrund anderer Faktoren (dazu mehr in 3). Absichtliche Fehlurteile dürften in der Regel die Ausnahme darstellen. Die vorliegende Arbeit stellt unbewusste Fehler der Unparteiischen in den Mittelpunkt, die entstehen, obwohl oder gerade weil sie motiviert sind ‚richtig' zu entscheiden.

Doch was stellt die richtige Entscheidung dar? Diese Frage kann klar mit den *Regeln der Sportart* beantwortet werden. Jede Sportart hat festgeschriebene Richtlinien, die vorgeben, wie sich der Unparteiische in bestimmten Situationen entscheiden soll. Darin wird genau festgehalten, wie Urteile zu fällen sind, um möglichst objektiv den tatsächlichen Sieger zu ermitteln (FIG, 2006, Art.1). Die eingesetzte Beurteilungsmethode soll einen möglichst hohen Grad an Reliabilität und Validität aufweisen, da das Ergebnis häufig bedeutende Folgen nach sich zieht.

2.3 Wertungsvorschriften und Kampfrichterlizenzen im Gerätturnen

Als *Code de Pointage* (CdP) werden die internationalen Wertungsvorschriften im Gerätturnen bezeichnet. Der CdP wird vom internationalen Turnerbund (FIG) für das weibliche und das männliche Gerätturnen getrennt formuliert. Alle vier Jahre werden sie den aktuellen Gegebenheiten in der Weltspitze angepasst. Der erste Satz der deutschen Fassung des CdP (FIG, 2006, Art.1) unter dem Titel „Ziel und Zweck der Wertungsvorschriften" lautet:

> „Der Hauptzweck der Wertungsvorschriften besteht darin, für alle Ebenen regionaler, nationaler und internationaler Wettkämpfe ein objektives Mittel zur Bewertung von Übungen für das Kunstturnen bereitzustellen".

Darüber hinaus sollen dem Trainer und dem Turner Unterstützung bei der Zusammenstellung der Übungen geleistet werden (ebenda, Art.1). Neben allgemeinen Verhaltensweisen für Turner, Trainer und Kampfrichter und den Grundlagen für die Übungsbewertung (ebenda, K.4-6) enthalten die Wertungsvorschriften einen Überblick über die einzelnen Elemente und ihre Schwierigkeit (ebenda, K.7-13).

Zu Beginn des Jahres 2007 traten die neuen Wertungsvorschriften im Gerätturnen der Männer[8] in Kraft, welche die Abschaffung der Höchstnote 10 (Punkte) mit sich brachten. Ziel dieser Neuerung ist, die Wertungsskala nach oben hin zu öffnen, um Leistungsunterschiede im oberen Leistungsbereich deutlicher herausstellen zu können.

Um eine Überforderung der Kampfrichter zu verhindern oder zumindest deren Entstehung zu reduzieren (2.1), setzt sich die Gesamtnote aus einer A-Note und einer B-Note zusammen, wobei die A-Note die ‚Schwierigkeit' und die B-Note die ‚Ausführung' ausdrückt. Die Noten werden von zwei unterschiedlichen Kampfgerichten, dem A- und dem B-Kampfgericht, getrennt ermittelt.

Zur Ermittlung der *A-Note* werden die neun schwierigsten Elemente der Übung und das letzte Element (der ‚Abgang') zusammengefasst. Dazu sind die verschiedenen Elemente in Schwierigkeitsgrade von A bis F unterteilt. Ein A-Element bringt einen Bonus von 0,1 Punkten, ein F-Element einen Bonus von 0,6 Punkten. Hinzu können Bonuspunkte für die Verbindung von Elementen (0,2 Punkte) kommen. Die Addition der Faktoren Schwierigkeit der Elemente, Erfüllung der Elementgruppen und des Abgangs sowie mögliche Bonuspunkte ergeben die A-Note (ebenda, K.4, 5 & 7-12).

Die neuen Vorschriften fördern eine stärkere *Gewichtung der Ausführung* (B-Note). Um diesen Trend noch zu verstärken, wurden die Abzüge für Verstöße gegen technische, haltungsmäßige und kompositorische Anforderungen stark erhöht. Die Höchstnote 10 (Punkte) bleibt insofern erhalten, dass in der B-Note von 10,0 Punkten beginnend abgezogen wird. Der kleine Fehler bleibt bei einem Abzug von 0,1 Punkten, der mittlere wird schon mit 0,3 Punkten bestraft und der grobe Fehler zieht einen Abzug von 0,5 Punkten nach sich. Ein Sturz mit 0,8 Punkten Bestrafung (äquivalenter Wert von zwei D-Teilen) stellt eine sehr starke Sanktion dar. Die Aufgaben des Kampfgerichts während des Wettkampfes sind genau definiert und sehen für das B-Kampfgericht vor, dass jeder Kampfrichter die präsentierte Übung bewertet und die Summe der Technik-, Kompositions- und Haltungs-Fehler (Abzüge) selbständig, ohne Kontakt mit den anderen Kampfrichtern bestimmt.

> „Die Ausdifferenzierung der Bewegungsnorm in möglichst viele Beobachtungseinheiten, die gleichzeitig zu beachten sind, erhöht die Genauigkeit der Beurteilung keineswegs automatisch. Es ist im

[8] Das männliche Gerätturnen umfasst die Geräte Boden, Pauschenpferd, Ringe, Sprung, Barren und Reck, die in dieser olympischen Reihenfolge geturnt werden.

> Gegenteil mit dem Entstehen einer Überforderungssituation und damit einer Verschlechterung der Beurteilungssituation zu rechnen“ (Neumaier, 1988, S. 223).

Zeitliche Bestimmungen bestehen für jeden Kampfrichter des B-Kampfgerichts, da die Abzüge innerhalb von zehn Sekunden nach Beendigung der Übung berechnet und übermittelt bzw. angezeigt werden müssen (FIG, 2006, K.4 & 6-12). Darüber hinaus wird, wie in anderen technisch-kompositorischen Sportarten üblich, die Regel verfolgt, dass die höchste und die niedrigste Note des B-Kampfgerichts gestrichen[9] werden. Die sogenannten Streichnoten werden nicht in die Berechung der B-Endnote berücksichtigt (ebenda, K.4, Art.11).

Die verbleibenden, mittleren Noten werden addiert und durch die Anzahl der Kampfrichter[10] dividiert, dadurch ergibt sich die Note des B-Kampfgerichts. Für die mittleren Summen der Abzüge (Noten) besteht eine Regel über die zugelassenen Unterschiede zwischen den (vier bzw. zwei) Noten (ebenda, K.4 & 6). Diese festgelegten maximalen Abweichungen stehen im Verhältnis zum errechneten Mittelwert. Bei einem größeren Unterschied kann der Supervisor (Kontrollkampfrichter) des Gerätes, nach Prüfung seiner Note, einschreiten und die Note ändern lassen. Falls derartige Abweichungen auffällig werden, können Sanktionen für die betroffenen Kampfrichter verhängt werden (ebenda, K.2, Art.5).

Die festgelegte Note des A-Kampfgerichts und der gemittelte Wert des B-Kampfgerichts werden addiert und ergeben die Endnote für die gezeigte Übung.

Für die Kampfrichter sind in den Wertungsvorschriften ebenfalls Regelungen beinhaltet, die beispielsweise die Platzordnung und die Arbeitsweise vorschreiben. Mitinbegriffen ist dabei, dass die Kampfrichter bei offiziellen FIG-Wettkämpfen schriftliche Aufzeichnungen ihrer persönlichen Wertungen anfertigen müssen, die auf Wertungsblätter fixiert werden (ebenda, K.3, Art.9). Die Verpflichtung, während des Übungsablaufs Aufzeichnungen zu machen, führt zu einer zusätzlichen

---

[9] Das gestutzte Mittel wird angewendet, um Daten, die zu hoch bzw. zu niedrig im Vergleich zu den restlichen Daten sind, weniger Gewicht zu verleihen. Man sortiert die Werte nach aufsteigender Größe, schneidet eine gleiche Anzahl von Werten am Anfang und Ende ab und bildet das arithmetische Mittel der verbleibenden Werte (Stahel, 2007).

[10] Bei offiziellen FIG-Wettkämpfen bilden sechs (vier) Kampfrichter das B-Kampfgericht. Vier (zwei) Noten bilden die Berechnungsgrundlage für die B-Note.

Erschwerung der Wahrnehmungssituation, da immer wieder eine Aufmerksamkeitsteilung oder der völlige Abzug von der Übungsdurchführung notwendig ist (Neumaier, 1988, S. 410). Anhand der Aufzeichnungen ist es möglich, bei Unklarheiten gegebenenfalls noch einmal den Übungsverlauf und die entsprechenden Mängel in der Übung nachzuvollziehen. Allerdings ist diese Methode der nachträglichen Prüfung nicht frei von unbewussten Fehlern im Urteilsprozess, da diese auf kognitiver Ebene stattfinden und im Moment der Niederschrift bereits vorherrschend sind. Im Gegensatz dazu ist für die nachträgliche Prüfung der Noten mit den neuen Wertungsvorschriften eingeführt worden, dass der Einsatz von Videoaufnahmen (2.1) erlaubt ist. Die Kampfrichter können somit seit 2006 nach Einsicht der Videoaufzeichnung ihre Entscheidung nachträglich ändern. Die Kriterien zur Einsicht der B-Note sind im Vergleich zur Einsicht der A-Note deutlich strenger formuliert. Begründet wird dies damit, dass man als Kampfrichter bezüglich der Ausführung der Übung immer verschiedene Sichtweisen haben kann und dies zu unüberschaubaren Protesten[11] führen würde.

Die generelle Verteilung der Aufgaben auf zwei unabhängige Kampfgerichte führt zur Entlastung der Kampfrichter und erhöht die Reliabilität der Urteile (Landers, 1970; Neumaier, 1988), stößt aber dennoch an die Grenzen der menschlichen Informationsverarbeitung oder überschreitet diese (O´Brien, 1991; Plessner, 1997, 2004; Salmela, 1978). Die beschriebenen Unterschiede in der Bildung der beiden Noten werden durch die Organisation in zwei Kampfgerichte unterstrichen. Das A-Kampfgericht setzt sich aus zwei Kampfrichtern zusammen, während das B-Kampfgericht aus mehreren (bis zu sechs) Personen besteht, die aus unterschiedlichen Blickwinkeln die Übungen bewerten. Da die Wertung des B-Kampfgerichts zu den meisten Diskussionen und Kontroversen führt (Salmela, 1978), wird durch die größere Anzahl an Kampfrichtern versucht, die Reliabilität zu erhöhen (Landers, 1970).

Die unterschiedlichen Blickwinkel der Kampfrichter sollen die Reliabilität der Urteile erhöhen (Wilson, 1976a & b). Die B-Note soll objektiviert werden, da man davon ausgeht, dass sich die subjektiven Anteile (3.2), wie beispielsweise die Präferenz für die eigene Nation, durch die höhere Anzahl an Kampfrichtern herausmitteln (O´Brien, 1991).

[11] Proteste, die zur Kontrolle der Videoaufzeichnungen führen, dürfen lauf Reglement bei der Schwierigkeitsnote (A-Note), wenn sie „vermutlich nicht korrekt ist“ und bei der Ausführungsnote (B-Note), wenn „eine erhebliche Abweichung innerhalb der vier zählenden B-Noten auftritt“ (FIG, 2006, K.2, Art.5), erhoben werden.

Verzerrungen dieser Art versucht man explizit im CdP (FIG, 2006, K.2, Art.7.1) in Form des Kampfrichtereids zu unterbinden:

> „Ich erkläre bei meiner Ehre, dass ich mich, in meiner Funktion als Kampfrichter einzig durch den Geist der sportlichen Loyalität und Würde leiten lassen werde, und ich verpflichte mich, die gezeigten Übungen gewissenhaft und ohne Ansehen der Person oder Nation zu bewerten".

Trotz den durchgeführten Verbesserungen an den Wertungsvorschriften kann man davon ausgehen, dass diese empfindlich gegenüber dem in dieser Arbeit untersuchten Reihenfolge-Effekt sind. Die Athleten, Trainer und Kampfrichter sind sich der größeren Bedeutung der B-Note bewusst und achten vermehrt auf die Einhaltung der geforderten Ausführungskriterien. Dadurch ergibt sich eine relativ ‚saubere' Übungsausführung aus der Perspektive der Kampfrichter. Die entstehenden mittleren oder groben Fehler treten stärker hervor. Somit sind die Übungen, aus Sicht der B-Kampfrichter, einfacher zu bewerten (O´Brien, 1991), da sich nicht so viele ‚winzige' Fehler anhäufen und sich auch weniger überlagern. Die Reduzierung von Fehleranhäufung macht es dem gut ausgebildeten B-Kampfrichter somit auch überschaubarer, Athleten mit hohem Leistungsniveau zu bewerten, im Gegensatz zu Athleten mit niedrigerem Leistungsniveau. Wenn sich die einzelnen Fehler klarer abzeichnen, kann das eine günstigere Voraussetzung für einen eventuell vorherrschenden Reihenfolge-Effekt darstellen. Somit kann beispielsweise der mittlere Fehler, der zu Beginn der Übung geturnt wird, auch als solcher wahrgenommen und definiert werden. Er geht nicht unter in der ‚Flut mannigfacher Fehler', die die gesamte Übung durchziehen.

Im Gegensatz dazu bringt eine Übung auf hohem Leistungsniveau für weniger gut ausgebildete Kampfrichter vermutlich Hindernisse mit sich. Alleine die gezeigten Schwierigkeiten und die Verbindungen solcher Übungen bereiten dem Kampfrichter mit geringerer Expertise möglicherweise durch die ungewohnte Situation Probleme, da er eher weniger Erfahrungen auf diesem Gebiet sammeln konnte. Auch wenn dieser lediglich die Ausführung der Elemente einschätzen soll, könnten hochklassige und damit ungewohnte Übungen durch eine mögliche Begrenzung der Aufmerksamkeit die Informationsverarbeitung einschränken und somit Fehler schüren. Hochqualifizierte Wertungsrichter verfügen über ein umfangreiches und sicheres Wissen bezüglich der verbindlichen Wertungsvorschriften (Salmela, 1978; Ste-Marie, 1999). Überdurchschnittliche Kenntnisse der zu beurteilenden Bewegungsabläufe

sparen die kognitiven Ressourcen des Kampfrichters ein und umgehen Einschränkungen der menschlichen Informationsverarbeitungskapazität (Salmela, 1978; Ste-Marie, 2000).

Eine Möglichkeit, den Qualifikationsgrad der Kampfrichter zu bestimmen, stellt die Betrachtung der erworbenen *Wertungslizenz* dar. Um als Kampfrichter bei Wettkämpfen eingesetzt zu werden, wird neben einem gesunden Urteilsvermögen und Fairness eine Kampfrichterlizenz gefordert. Ohne diesen ‚Schein', der die Qualität der eigenen Ausbildung und die Verantwortung gegenüber den Aktiven ausdrückt, wird kein Kampfrichter zu einem Wettkampf zugelassen. Welche Lizenzstufe notwendig ist, um einen Wettkampf werten zu dürfen, hängt von der Wettkampfausschreibung ab. Die Einteilung der Kampfrichterlizenzen reicht von der sogenannten D-Lizenzstufe bis hin zu der internationalen Lizenz. Die nationale Kampfrichterlizenz D berechtigt das Werten von Wettkampfübungen auf Gau- und Kreisliganiveau. Mit der C-Lizenz ist auch der Einsatz bei Landeswettkämpfen möglich. Die Kampfrichterlizenzstufe B ermöglicht einen vollständigen Einsatz bei Landeswettkämpfen und teilweise den Einsatz bei nationalen Wettkämpfen. Das Werten auf höchsten nationalen Wettkämpfen wird mit der A-Kampfrichterlizenz möglich. Die internationale Kampfrichterlizenz ermöglicht den Einsatz bei allen nationalen und je nach Kategorie bei internationalen Wettkämpfen. Man kann vier Kategorien der internationalen Lizenz unterscheiden, bis hin zur ersten Kategorie, mit der man berechtigt ist, alle internationalen Wettkämpfe wie beispielsweise Europameisterschaften, Weltmeisterschaften oder Olympische Spiele zu werten. Prinzipiell gilt, dass der Einstieg in die Ausbildung mit der Kampfrichterlizenzstufe D gewählt und danach in den höheren Lizenzstufen immer mehr Professionalität erreicht wird, z.B. durch Lernen der Symbolschrift.

Grundsätzlich sind die Kampfrichterausbildungen der benachbarten deutschsprachigen (und anderen) Ländern sehr ähnlich, da der CdP weltweit Gültigkeit findet. Die Aus- und Fortbildung sowie die Prüfung der internationalen Lizenz sind in allen Ländern identisch und werden weltweit von Kampfrichterbeauftragten der FIG durchgeführt. Auf nationaler wie auch auf internationaler Ebene sind die Wertungsvorschriften verbindliche Grundlage der Prüfungsinhalte.

2.4 Die komplexe Aufgabe der Kampfrichter

Die Aufgabe der Kampfrichter stellt sich als eine *sehr komplexe* dar (O´Brien, 1991; Salmela, 1978). Die Komplexität der Aufgabe wird in

vielerlei Hinsicht sichtbar. Grundsätzlich ist die Endgültigkeit des Urteils zu nennen, dass aufgrund der Tatsachenentscheidung im Sport nicht widerrufen werden kann. Zudem stellt die Einmaligkeit des Geschehens, die im Normalfall keine Wiederholung zulässt und die Schnelligkeit, in der die Bewegungen stattfinden, eine Besonderheit der Urteilsaufgabe dar. Wenn eine Bewegung bzw. Handlung nicht gesehen wurde, muss der Kampfrichter trotzdem sein Urteil abgeben. Er muss die Übung als Ganzes betrachten und dennoch die einzelnen Details erfassen. Wie detailliert er dabei vorgeht, hängt von den Wertungsvorschriften (2.3) ab (Thomas, 1978).

Neumaier (1988, S. 385 - 400) unterscheidet drei Einflussgrößen, die sich auf die Bewertung der komplexen Wettkampfleistung auswirken. Erstens erläutert er die *sehobjektabhängigen* Einflussgrößen und zählt den Sehwinkel, die Winkelgeschwindigkeit, die Darbietungszeit, den Beobachterstandort, die Beleuchtungsbedingungen sowie die Darbietungsform auf. Auch Plessner (1997, S. 54) ist der Meinung, dass „der Hauptfaktor, der eine reliable und objektive Bewertung verhindert, … die Schnelligkeit der Übungsfolge zu sein [scheint]“.

Zweitens nennt er die *aufgabenabhängigen* Einflussgrößen und beschreibt vier Bereiche, die Beurteilungsaufgabe, den -umfang, die Beobachtungseinheiten und die Informationsdichte. Der Beurteilungsgegenstand des Kampfrichters, also die vielen Einzelbewegungen mit den enthaltenen zahlreichen Details, die sehr hohe Informationsdichte bei sehr schnell ablaufenden Bewegungen (O´Brien, 1991) und bei geringen Zeitabständen zwischen den einzelnen Übungsteilen, macht eine optimale Bewertung beinahe unmöglich. Nicht nur, dass die Kampfrichter die Übungen anschauen müssen, sie sollen die gezeigten Elemente mit den definierten Bewegungsnormen gedanklich vergleichen (Salmela, 1978). Abweichungen von dieser Norm gelten als Fehler und werden anhand von Winkeln und anderen festgeschriebenen Kriterien (wie beispielsweise die Beugung von Armen und Beinen) in Abzüge umgerechnet (Bard et al., 1980). Der Kampfrichter soll Unmögliches leisten, indem er beispielsweise unterscheiden soll, ob ein Kreuzhang eine Arm-Rumpf-Winkelstellung von 30 oder 31 Grad hat (Plessner, 2001a). Dabei erfolgt mitunter eine Einschätzung des Ausmaßes des Fehlers (FIG, 2006, K.6, Art.19; O´Brien, 1991). Die extrem hohe Anzahl an Regeln und Elementen, die der Kampfrichter beherrschen muss, führt zu einer hohen kognitiven Beanspruchung (O´Brien, 1991). Die Bewertungsaufgabe verlangt, dass der Kampfrichter ein Übungsteil möglichst vollständig und lange Zeit im Kurzzeitspeicher verfügbar hält, um diese

Simultan- und Sukzessivbewertungen vornehmen zu können. Die Dauer einer Gesamtübung übersteigt hingegen die Verweildauer von Informationen im Kurzzeitspeicher deutlich. Die Wahrnehmungen werden in den Langzeitspeicher übertragen, falls sie nicht sofort protokolliert werden können. Die fortgeführte Beobachtungs- und Beurteilungsaktivität verhindert aber ein ausreichendes Memorieren zur Sicherung der Information. Es kommt zu Zerfallserscheinungen (Neumaier, 1988, S. 409).

> „Die wettkampfgerechte Beurteilung einer Bodenturnübung mit dem Versuch, die vorliegenden Bewertungsrichtlinien vorschriftsmäßig zu befolgen, [stellt] in Anbetracht des Beurteilungsumfangs und der Informationsdichte sowie der verlangten Mehrfachaufgabe eine völlige Überforderung des Kampfrichters dar“ (ebenda, S. 410).

Als dritten Punkt, summiert der Autor unter den *beurteilerabhängigen* Einflussgrößen die visuelle Leistungsfähigkeit, aber auch Vorinformation, Sehstrategie, Kapazität des Kurzzeitgedächtnisses, psychologisch bedingte Beurteilungsfehler, die Motivation, Emotionen, die Konzentrationsfähigkeit und kognitive Verarbeitungsstile.

Nicht zu vergessen ist der ständig vorherrschende Lärmpegel oder auch Zuschauerreaktionen, die die Informationsverarbeitungskapazität weiter einschränken können (O´Brien, 1991). Die Konzentrationsfähigkeit der Kampfrichter muss stark ausgebildet sein, wenn man bedenkt, welchen Ablenkungen sie in einer derartigen Veranstaltung widerstehen müssen (ebenda). Auch die physischen Anforderungen stellen eine Besonderheit dar, denkt man an Aspekte wie die Bequemlichkeit der Sitzmöglichkeiten oder die Hallentemperatur. Hinzu kommt die Ermüdung durch stundenlang andauernde Wettkämpfe (O´Brien, 1991; Pflughoeft, 1984). Diese finden teilweise unter schlechten Bedingungen statt, wie beispielsweise als Kampfrichter am Reck, der ständig nach oben – und damit in die Scheinwerfer – schauen muss oder die manchmal nicht optimale Versorgung der Kampfrichter.

Plessner (1997) beschreibt die erschwerte Aufgabe durch die Anwesenheit *anderer Kampfrichter*. Üblicherweise werten Kampfrichter in Kampfgerichten (2.3). Dabei kann es zu zwei möglichen Folgen dieser Art der Benotung kommen. Zum einen passt sich eine Person in ihrem Urteil dem der Mehrheit an (4.e in 3.2), da sie von der Richtigkeit der anderen überzeugt ist oder weil sie nicht von ihnen abweichen möchte (Scheer, Ansorge & Howard, 1983). Zum anderen kann diese Art der Bewertung Leistungseinbußen bewirken. Die Sozialpsychologie

betrachtet Konformitätsurteile differenziert und unterscheidet Faktoren, die die Tendenz eines Beurteilers zu konformem Verhalten beeinflussen. So steigert die Verantwortlichkeit – also die Verpflichtung zur Rechenschaft – des getroffenen Urteils, die Konformität. Wenn allerdings das Bedürfnis des Urteilers nach Korrektheit hinzu kommt, wird der Konformitätsdruck ignoriert und das Urteil unabhängig von anderen Personen gefällt (vgl. Aronson, Wilson & Akert, 2004, S. 314). Man geht davon aus, dass die Bearbeitung einer einfachen und gut geübten Aufgabe unter Beisein anderer besser gelingt, während es bei einer komplexen, schwierigen Informationsverarbeitungsaufgabe zu einer eher schlechten Leistung führt. Besonders wenn negative Konsequenzen eher als positive zu erwarten sind, wie die im CdP festgeschriebene Maßnahme der Sanktion bei starkem Abweichen der urteilenden Kampfrichter. Die Anstrengung ist dabei zwar größer, der Einfluss aufgabenirrelevanter Informationen erhöht sich allerdings. Dieser Effekt ist bei alleiniger Wertung geringer (6.2).

Auch das *Mitschreiben* der wahrgenommenen Elemente (bei der A-Note) bzw. der beobachteten Fehler (B-Note) stellt für O´Brien (1991) einen Hinweis dar, dass es sich um eine sehr komplexe Aufgabe handelt. Somit ist es üblich, dass die Kampfrichter während der Übungspräsentation Notizen machen. Somit stellt ein ‚normaler' Wertungsprozess eine schrittweise (step by step - SbS) kognitive Bewertung dar. Diese Methode des Mitschreibens erhöht die Anforderungen und ist eher schlecht für die Beurteilungsqualität (Neumaier, 1988, S. 413). Allerdings lässt ein Mitschrieb kaum noch spätere heuristische Verzerrungen aufgrund beispielsweise des Gesamteindrucks zu. Somit ist es sinnvoll, die Übungen manuell anhand der Symbolschrift oder Fehlerabzügen aufzuzeichnen. Dies sollte aber frühzeitig erlernt und häufig angewendet werden (Plessner, 2001a), um nicht zur Überforderung der Informationsverarbeitung zu führen. Es gibt allerdings Kampfrichter, die vor allem bei der Bewertung der Ausführung (B-Note), keine Notizen machen (5.2.2).

Plessner (1997, 1999) nimmt weiterhin an, dass die Komplexität und damit Schwierigkeit ein Urteil zu fällen, je nach *Gerät,* unterschiedlich einzustufen ist. Er unterscheidet dabei die schnellen Geräte Pauschenpferd, Sprung und Reck von den langsamen Geräten Boden, Ringe und Barren (5.2.1). Dabei stellen die schnellen Geräte höhere Anforderungen an den Beurteiler im Vergleich zu den langsamen.

Mit den beschriebenen Faktoren wird deutlich, dass der Kampfrichter eine große Menge an Informationen verarbeiten muss. Die zu

verarbeitende Informationsmenge ist bedeutend, wenn es darum geht ein Urteil zu bilden. So kann man sich vorstellen, dass viele Informationen durchaus hilfreich sein können. Kommt es aber dazu, dass zu viele Informationen vorherrschen, kann sich eine Art Verwässerung einstellen, und die Wahrnehmung und Bewertung verändert sich. Durch das Hinzunehmen irrelevanter Informationen wird zudem der Einfluss der für das Urteil relevanten Informationen getrübt.

Die Bewertung im Gerätturnen entsteht nach Salmela (1978) auf Grundlage der wahrgenommenen, erinnerten, verarbeiteten und zum Teil mitgeschriebenen Informationen. Diese Art der Noten-Erschließung bedarf einer guten Ausbildung und langjährigen Erfahrung – als Turner, Trainer und Kampfrichter (auch Thomas, 1978, S. 266). Auch Weigelt (2006, S. 117) betont, dass komplexes Handeln anderer Akteure nur dann richtig eingeschätzt und bewertet werden können, wenn man diese Handlungen auch selbst beherrscht (5.4). Somit müsste man meinen, dass eine vielfältige und gute Ausbildung ausreicht, um Fehler im Urteilsprozess zu vermeiden und dazu verhilft, immer die ‚richtige' Entscheidung zu treffen. Die Anzahl der Jahre, die ein Kampfrichter vorweisen kann, sagt nichts über die Qualität der Urteile aus (Wilson, 1976a).

Immer wieder kommt es zu Fehlentscheidungen, die mehr oder weniger folgenschwer sind. Je schwieriger die Urteilsaufgabe ist bzw. je mehr Unsicherheit beim Urteilen vorherrscht, desto stärker müssten sich subjektive Einflüsse auf das Urteil auswirken. Oftmals merken die Beurteiler nicht, dass sie irrelevante Informationen in ihr Urteil aufgenommen und damit ein verzerrtes Urteil gebildet haben.

Was ist, wenn mehrere oder gar alle Kampfrichter derselben unbewussten systematischen Einflussnahme unterliegen und damit gar nicht in der Lage sind das ‚richtige' Urteil zu fällen (3)? Im folgenden Kapitel wird genau dieser Frage nachgegangen.

3 Urteilsfehler

Jedes menschliche Urteil hat einen subjektiven Charakter und schließt die Entstehung von Fehlern im Urteilsprozess nicht aus. „Urteile in sozialen Situationen werden häufig unter suboptimalen Bedingungen gefällt (Bless & Keller, 2006, S. 294). Als Ursache können *funktionale* bzw. kognitive aber auch *motivationale Faktoren* angesehen werden, die zu einem falschen Urteil führen (Fischer et al., 2006).

Eine nicht leistungsgerechte Beurteilung kann aufgrund von strategischen Überlegungen, leistungsfremde Gesichtspunkte in das Urteil integrieren und führt dadurch zur Vergabe bestimmter Noten. Diese absichtliche, motivational gesteuerte Beeinflussung von Urteilen ist für die Beantwortung der Forschungsfrage nicht von Interesse und daher zu vernachlässigen.

Vielmehr werden Urteilsfehler bzw. -verzerrungen thematisiert, die unbewusst ablaufen und vom Urteilenden, im speziellen Fall dem Kampfrichter, normalerweise nicht wahrgenommen werden. Da in manchen Fällen nicht eindeutig zu klären ist, welches Urteil objektiv richtig und welches falsch ist, wird in der Sozialpsychologie vorsichtiger von *Verzerrungen* oder *Täuschungen* statt von *Fehlern* in der Urteilsbildung gesprochen (Kanning, 1999; Plessner & Raab, 1999). Im sportlichen Kontext ist es meist möglich festzustellen, ob es zu Abweichungen von den definierten Vorschriften der Sportart kommt, indem beispielsweise technische Hilfsmittel wie die Videomethode eingesetzt werden. Daher könnte man klar von Urteilsfehlern sprechen. Da es manchmal eben nicht eindeutig ist, ob ein Fehler begangen wurde oder ob beispielsweise die Blickperspektive unterschiedliche ‚richtige' Einschätzungen zulässt, werden in dieser Arbeit die Begriffe ‚Urteilsfehler' und ‚Urteilsverzerrung' synonym verwendet.

3.1 Urteilsbildung und die Entstehung von Urteilsfehlern

Menschen sind *‚soziale Wesen'* und machen sich daher Gedanken darüber, wer derjenige ist, mit dem sie es zu tun haben. Die Einschätzung anderer gehört zum alltäglichen Leben. Da der Mensch nicht unfehlbar ist, unterlaufen ihm auch auf diesem Gebiet mehr oder weniger folgenschwere Fehler. Im Alltag nimmt man Fehlurteile leichter in Kauf, aber im beruflichen Kontext sollte die eigene Urteilsfähigkeit bewusst und im Idealfall fehlerfrei ausgebildet sein, da die Folgen von Fehlurteilen durchaus schwerwiegend sein können.

Im Forschungsfeld der sozialen Kognition, haben sich unterschiedliche Ansichten entwickelt, wie der Mensch ein Urteil bildet. Sie beschäftigen sich mit den geistigen Prozessen, die zwischen der Reizwahrnehmung und dem Verhalten vermitteln. Soziale Urteile werden nicht nur durch den äußeren Reiz bestimmt, sondern auch davon, wie der Mensch diesen wahrnimmt und versteht. Die nachstehenden beiden Auffassungen sollen einen kurzen Einblick in dieses Forschungsgebiet ermöglichen.

Unter dem Begriff ‚kognitive Algebra' ist der mathematische Ansatz bekannt. Das *‚Informations-Integrations'-Modell* von Anderson (1974) beschreibt und modelliert, wie eine Person Informationen unterschiedlicher Quellen integriert, um sich ein Urteil zu bilden. Das auf algebraischen Richtlinien basierende Modell geht davon aus, dass ein Stimulus zunächst in einen Zahlenwert übersetzt und damit bewertet werden muss. Dann wird eine Rechenoperation, wie beispielsweise durch Addition oder Multiplikation, eine Gewichtung vorgenommen, um das neue Urteil bilden zu können. Diese Betrachtungsweise entspricht der Grundannahme, dass man den Prozess der Informationsintegration bei der Eindrucksbildung in mathematische Modelle fassen kann. Dieser Ansatz liefert zwei unterschiedliche Modelle der Eindrucksbildung – das Summenmodell und das Durchschnittsmodell. Beim Summenmodell wird der Gesamteindruck als Summe aller Einzeleindrücke, seien sie noch so marginal, gebildet. Im Gegensatz dazu vertritt das Durchschnittsmodell die Auffassung, dass der Gesamteindruck dem arithmetischen Mittel der Einzeleindrücke entspricht. Beide Modelle bringen unterschiedliche Ergebnisse hervor. Später wandelte Anderson sein Modell dadurch ab, indem er nicht alle Merkmale gleich gewichtete, um die unterschiedlichen Einflüsse abzubilden. Das *gewichtete Durchschnittsmodell* entstand.

Lassen sich Prozesse der Eindrucksbildung des alltäglichen Lebens tatsächlich auf einfache mathematische Formeln reduzieren? Einerseits müssten die Sympathiewerte von Persönlichkeitsmerkmalen dauerhaft und unveränderlich sein und andererseits sollte die Eindrucksbildung einen einfachen, rationalen, kognitiven Prozess darstellen. Die zweite Annahme kann verworfen werden, was einen tatsächlichen Gebrauch dieses Modells bereits unmöglich macht.

Asch (1946) geht davon aus, dass Menschen dazu neigen, die Eindrucksbildung als *ganzheitlichen Prozess* anzusehen. Dabei geben sie bestimmten zentralen Merkmalen einen unverhältnismäßig großen Einfluss. Andere periphere Merkmale haben weitaus geringeren Einfluss auf den Eindruck, den man sich über andere Personen bildet. Asch

belegt seine Aussagen eindrucksvoll, indem seine Probanden eine Liste mit Adjektiven erhalten, die eine Zielperson beschreiben. Die Aufgabe der Probanden ist die Wiedergabe des Gesamteindrucks über die Zielperson anhand dieser Persönlichkeitsmerkmale. Die Liste der Adjektive wird einmal in einem zentralen Merkmal verändert und einmal in einem peripheren Merkmal. Unabhängig von der präsentierten Reihenfolge der Adjektive beeinflussen bestimmte Informationen, weil sie besonders stark gewichtet werden und weil sie die Bedeutung anderer Informationen modifizieren. In Aschs Untersuchungen war dies die Information „warmherzig“ bzw. „kühl“, die besonders stark gewichtet wurde und infolgedessen die Eigenschaft „intelligent“ völlig unterschiedlich interpretiert wurde. Somit scheinen zentrale Merkmale einer Person, im Vergleich zu peripheren Merkmalen, einen bedeutenden Einfluss auf die Beurteilung auszuüben. Welche Determinanten dazu führen, dass ein Merkmal als zentral bzw. als peripher angesehen wird, scheint in diesem Zusammenhang eine wichtige Frage zu sein.

In vergleichbar aufgebauten Untersuchungen zur Urteilsbildung entdeckte Asch, dass Einzelinformationen aufgrund ihrer Position bei der Präsentation unterschiedlich stark in den Gesamteindruck der Probanden eingehen (ebenda). Informationen, die am Anfang einer Liste standen, hatten dabei einen größeren Einfluss auf die Urteilsbildung als diejenigen Informationen, die in der Mitte dieser Liste präsentiert wurden. Der sogenannte *Primacy-Effekt* zeigt sich beispielsweise in einem alltäglichen Beispiel, indem man sich einen ersten Eindruck zu Beginn des Kennenlernens macht und dieser die Bewertung der folgenden Informationen über die andere Person beeinflusst. Der *Recency-Effekt* beschreibt hingegen die besondere Gewichtung der letztgenannten Information in der Beurteilung anderer Personen. Unter bestimmten Bedingungen kann man diesen Effekt beobachten (3.3).

Die Reihenfolge der präsentierten Informationen stellt einen verzerrenden Einfluss auf die Urteilsbildung dar. Je nachdem in welcher Reihenfolge die Informationen dargeboten werden, entsteht ein unterschiedliches Urteil. Doch auch andere Faktoren können das Urteil in unterschiedlicher Art und Weise beeinflussen (3.2).

Eine eher molekulare Herangehensweise, wie sie etwa Tversky und Kahneman (1973,1974; Kahneman & Tversky, 1972) bevorzugen, stellt die verzerrenden Einflussfaktoren im Urteilsfindungsprozess dar. In zahlreichen einflussreichen Artikeln haben die Autoren die Bedeutung dreier Phänomene herausgestellt. Diese sind in der Literatur als *Urteilsheuristiken* bekannt. Weniger statistisches Kalkül, sondern

vereinfachende Entscheidungsregeln verhelfen dem Urteilenden eine Entscheidung zu treffen. Bei einer Überforderung von Kampfrichtern im Gerätturnen werden demnach vorliegende Informationen gar nicht genutzt, sondern verkürzende Urteilsstrategien (Plessner, 1997). Eine derartige Strategie besteht darin, Wertungen aufgrund des Gesamteindrucks zu bilden oder auf Basis der Schwierigkeit des Abgangs einer Übung aus diesem Abgang Rückschlüsse auf die gesamte Übung zu ziehen (Plessner, 2001b). Urteilsheuristiken vereinfachen die menschliche Informationsverarbeitung und erfordern nur ein geringes Ausmaß an kognitiver Kapazität (Bless & Keller, 2006). Allerdings werden dadurch entscheidungsrelevante Informationen nicht oder nur teilweise eingesetzt. Heuristiken sind kognitiv basiert und stellen sich je nach Urteilssituation unterschiedlich dar. So verschieden und zahlreich die Urteilssituationen sind, so mannigfach stellen sich auch die Urteilsverzerrungen dar. Besonders durch die Arbeiten von Amos Tversky und Daniel Kahneman wurde allgemeinen Heuristiken, die also nicht auf spezifischen Inhalten beruhen, vermehrt Beachtung geschenkt. Drei der wichtigsten allgemeinen Heuristiken sind die Verfügbarkeitsheuristik, die Repräsentivitätsheuristik und die Mechanismen der Verankerung und Anpassung (ebenda). Ob die Leichtigkeit der Abrufbarkeit, die Wahrscheinlichkeit der Zugehörigkeit zu einer Kategorie oder die Orientierung an einen festgelegten Ausgangswert die gewählte Urteilsbildungs-Strategie darstellt, sie führt zu einem systematischen Fehlurteil.

Urteile in sozialen Situationen werden häufig unter suboptimalen Bedingungen gefällt (Bless & Keller, 2006, S. 294). Suboptimale Bedingungen können sich auf vielerlei Art und Weise ergeben und führen meist zu verzerrten Urteilen. In Kapitel 2.2 wurde beschrieben, dass die Unparteiischen des Sports – die Kampf- und Schiedsrichter – immer eine Entscheidung treffen müssen. Sie sehen die Situation einmal, ohne Wiederholung, und müssen direkt eine Entscheidung fällen. Dabei ist unberücksichtigt, ob der zu beurteilende Bewegungsablauf sehr komplex ist und durch die hohe Geschwindigkeit gar nicht so detailliert erfasst werden kann. Aber auch das eingeschränkte Sehvermögen (Neumaier, 1988) oder der ungünstige Blickwinkel führen zu einer nicht realen Wahrnehmung oder überfordern die kognitive Kapazität und produzieren daher Urteilsverzerrungen (3.2). Erschwerend kommt im Sport hinzu, dass der Schiedsrichter oder der Kampfrichter einen gewissen Ermessensspielraum zwischen unterschiedlich harten Maßnahmen hat. So kann der Fußball-Schiedsrichter in manchen Spielsituationen zwischen einer mündlichen Verwarnung oder einer gelben Karte wählen (Messner & Schmid, 2007). Oftmals geht das einher mit uneindeutigen

Situationen, in denen selbst durch nachträgliche Videoanalysen nicht eindeutig festzustellen war, ob diese Situation mit einem Elfmeter bestraft werden sollte (ebenda).

Auch bei zu wenig vorhandenen relevanten Informationen, unabhängig davon, ob diese nicht vorhanden oder nicht wahrgenommen wurden, sollte es zu einem objektiv richtigen Urteil kommen. Der *Stichprobenansatz* von Fiedler (2000) vertritt diese Auffassung, um zu erklären, wie Urteilsfehler entstehen. Dieser besagt, dass ein Urteilsfehler nicht unbedingt durch kognitive Defizite, sondern durch den Prozess des Sammelns von Informationen entsteht. Sämtliche Urteile basieren auf einer Stichprobe an Beobachtungen, die aus der Grundgesamtheit aller prinzipiell existierenden Informationen in der Umwelt gezogen wurde. Die Stichproben, die den Alltagsurteilen zugrunde liegen sind selten repräsentativ, da beispielsweise die selektive Wahrnehmung oder bestehende Erwartungen manche Informationen eher verfügbar machen als andere. Die resultierende Informationsstichprobe kann somit aus vielerlei Gründen unzulänglich oder verzerrt sein. Interessant ist, dass sich Menschen dieser Unzulänglichkeiten und Verzerrungen normalerweise überhaupt nicht bewusst sind.

Grieve und Hogg (1999) gehen davon aus, dass ein Urteil, das unter einem Grad an *Unsicherheit* gefällt wird, die Gefahr mit sich bringt, irrelevante Informationen zu berücksichtigen. Dieses unter Unsicherheit gefällte Urteil wirkt sich negativ auf die Urteilsqualität aus. Stehen zu wenig relevante Informationen zur Urteilsfindung zur Verfügung, werden häufig irrelevante Informationen berücksichtigt, was zu systematisch verzerrten Urteilen führt (Messner & Schmid, 2007). Irrelevante Informationen sind beispielsweise Rufe des Publikums, die kulturelle Ähnlichkeit oder aber auch die Trikotfarbe der Athleten (3.2).

Zusammenfassend hat die Überforderung des Beurteilers, aus welchen Gründen sie auch verursacht wird, eine Abweichung vom objektiven Urteil zur Folge (Neumaier, 1988).

Die Betrachtung der *kognitiven Stufen der sozialen Urteilsbildung* stellt nach Plessner und Haar (2006) eine sinnvolle Möglichkeit dar, den unterschiedlichen Ursachen von Urteilsfehlern auf den Grund zu gehen. Bis ein Urteil gefällt werden kann, müssen unterschiedliche Prozesse im Gedächtnis durchlaufen werden. Von der Aufnahme von Informationen über die Verarbeitung bis hin zum Ergebnis spielt nicht nur das Reizereignis eine bedeutende Rolle, sondern auch das individuelle Vorwissen. Dabei bauen die kognitiven Stufen der sozialen

Informationsverarbeitung aufeinander auf und beeinflussen sich gegenseitig (Fiedler & Bless, 2002).

Wie Abbildung 1 zeigt, werden erst unterschiedliche kognitive Stufen durchlaufen, bevor das Gedächtnis in der Lage ist, eine Schlussfolgerung bzw. ein Urteil zu fällen. Nachdem ein Reiz wahrgenommen wurde, kommt es zur Enkodierung und Interpretation. Der Enkodierungszustand ist dabei abhängig vom Vorwissen. In der Stufe der Enkodierung wird der äußere Reiz in eine internale Repräsentation umgesetzt. Die enkodierte Wahrnehmung kann nun im Gedächtnis abgespeichert werden. Dieses neu erworbene Wissen und das im Gedächtnis vorhandene Vorwissen bilden die Grundlage jeglicher Schlussfolgerungen und Urteile. In manchen Fällen folgt daraufhin ein beobachtbares Verhalten.

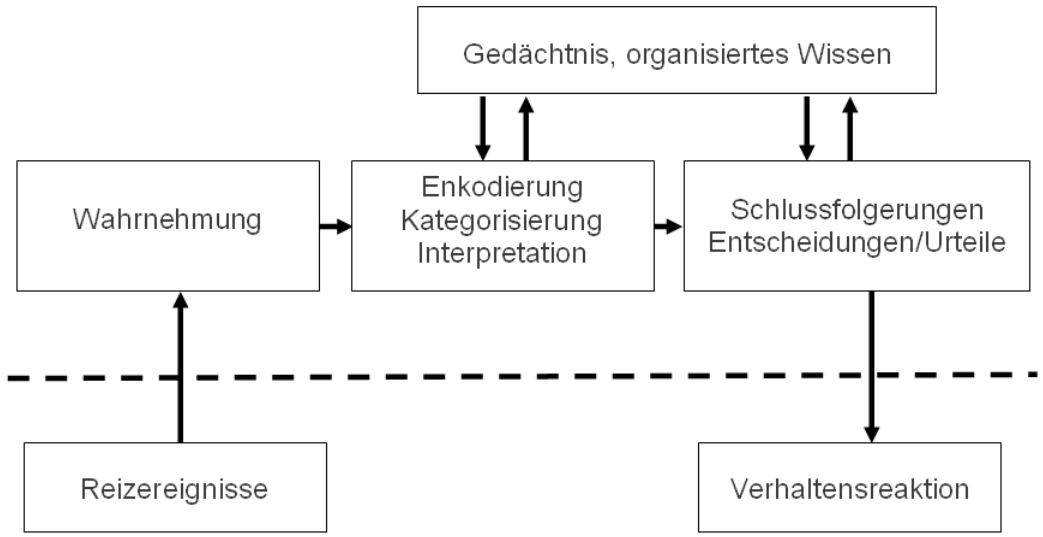


Abbildung 1: Die Stufen der sozialen Informationsverarbeitung (Fiedler & Bless, 2002, S. 133)

Wenn diese kognitiven Schritte durchlaufen werden müssen, um ein Urteil zu bilden, kann ein Fehler, so klein er auch sein mag, einen großen Einfluss auf das Urteil haben. Die Kette der im Urteilsprozess involvierten kognitiven Schritte kann mit dem bekannten Kinderspiel *‚Stille Post'* verglichen werden. Ein Kind flüstert einem anderen ein Wort zu und dieses flüstert es einem weiteren Kind zu und so weiter. Wie sich dieses Wort im Laufe der Kette verändern kann, so kann es auch im Prozess der Urteilsbildung zu vergleichbaren Modifikationen kommen. Die Originalinformationen werden durch zahlreiche Entwicklungen in der Urteilsbildungskette der kognitiven Stufen verändert.

An einem Handballbeispiel hat Plessner (2004) die Stufen der Informationsverarbeitung veranschaulicht und damit auf mögliche Fehlerquellen aufmerksam gemacht. Die von ihm mit Beispielen versehenen Stufen sind, wie bereits nach Fiedler und Bless (2002), die Wahrnehmung, die Kategorisierung, die Gedächtnisprozesse und die Urteils- und Entscheidungsprozesse. Die Wahrnehmung, als erste Stufe, stellt beispielsweise das Sehen eines klammernden Abwehrspielers dar. Diese Handlung wird vom Schiedsrichter als Foulspiel bewertet und durchläuft auf diese Weise die zweite Stufe, die der Kategorisierung. Die Gedächtnisprozesse, im Beispielfall die Erinnerung, dass derselbe Spieler bereits verwarnt wurde, bilden die dritte Stufe. Die vierte und letzte Stufe, Urteils- und Entscheidungsprozesse, führt alle Informationen zu der Entscheidung zusammen, dass in Form eines Freiwurfes und einer Zweiminutenstrafe ausgesprochen wird. Eine Fehlentscheidung, also keine Vergabe des Freiwurfs, könnte somit unterschiedliche Ursachen haben. Das Übersehen des Klammerns, das als weniger roh bewertete Foul oder das Vergessen der bereits ausgesprochenen Verwarnung können Möglichkeiten sein, wie ein Fehlurteil entstehen konnte. All diesen, grob aufgeführten Stufen der Informationsverarbeitung können wissenschaftliche Befunde über nachgewiesene Urteilsverzerrungen zugeordnet werden (Überblick siehe Plessner & Raab, 1999). Diese unterschiedlichen Quellen verursachen *systematische Fehlurteile.* Die meisten Menschen würden demzufolge denselben Fehler in diesen Situationen machen.

Eine zusätzliche Einteilung in lokale und globale Urteile nehmen Plessner und Haar (2006) vor. Dabei stellt ein lokales Urteil eines dar, das zeitlich und räumlich limitiert ist und nur die aktuelle Situation fokussiert, wie beispielsweise eine Gerätturnübung. Globale Urteile werden dadurch charakterisiert, dass sie über einen größeren Zeitraum hinweg, also über die momentane Situation hinaus, beeinflusst werden.

Die Überforderung des Beurteilers und die daraus ableitbaren Fehlurteile stellen ein interessantes Forschungsgebiet dar. Verschiedene Ursachen führen zu diesen Fehleinschätzungen und können auf den Stufen der sozialen Informationsverarbeitung eingeordnet werden. Dadurch ergibt sich eine hervorragende Möglichkeit einen spezifischen Einfluss, im vorliegenden Fall den Reihenfolge-Effekt, besser zu verstehen.

3.2 Verzerrende Einflüsse auf Kampfrichterurteile

Es gibt eine riesige Anzahl systematischer Fehler im Prozess des Urteilens, die auf unterschiedlichen Gebieten wie der Rechtssprechung oder der Wirtschaft oder auch der Kommunikationswissenschaft untersucht werden. Eine Kategorisierung dieser vielfältigen Verzerrungen erweist sich als schwierig, ist oftmals abhängig von der speziellen Urteilsaufgabe und sollte nicht verallgemeinernd, sondern im speziellen Kontext betrachtet werden.

Bei der Beurteilung anderer Personen kann davon ausgegangen werden, dass man vereinfachende Strategien nutzt und die erhaltene Information so zu kategorisieren versucht, dass sie unseren Erfahrungen, unserem Wissen über Personentypen und den Normen und Erwartungen unserer Kultur entsprechen. Im Forschungsfeld der Eindrucksbildung wurden unterschiedliche Ansätze entwickelt, die versuchen, das komplexe Thema detaillierter abzubilden.

Systematische *Urteilsverzerrungen im Sport* sind hingegen seltener erforscht worden. Vorwiegend psychische Ursachen in der Person des Beurteilers führen zu sehr unterschiedlichen Arten von Fehlurteilen. Sie sind dadurch charakterisiert, dass der Beurteiler eine leistungsgerechte Beurteilung nicht durchführen kann. Der Beurteiler beabsichtigt diese Fehler nicht und sie sind ihm oftmals auch nicht bewusst. Vor allem unter Zeitdruck und in Situationen, die aus verschiedenen Blickwinkeln zu unterschiedlichen Urteilen führen, neigen Menschen dazu, mehr Fehler und auch systematische Fehlurteile zu produzieren (Messner & Schmid, 2007). Es gibt eine Reihe von systematischen Fehlern beim Beurteilen sportlicher Leistungen, die wissenschaftlich mehr oder minder gut dokumentiert worden sind. Technisch-kompositorischen Sportarten, wie dem Gerätturnen, wurde vermehrt Aufmerksamkeit geschenkt. Begründet wird dies durch die Natur der Beurteilung dieser Sportartengruppe, die teilweise subjektiven Kriterien unterliegt (2). Im Gerätturnen entscheiden Kampfrichter über den Ausgang des sportlichen Wettstreits und obwohl sich die Erfahrenen unter ihnen von den weniger Erfahrenen unterscheiden (Bard et al., 1980; Plessner & Schallies, 2005; Ste-Marie, 1999, 2000), lassen sich systematische Urteilsfehler nicht nur unter Laien, sondern auch unter Experten finden (Plessner & Schallies, 2005; Scheer, 1973; Ste-Marie & Lee, 1991; Wilson, 1976a & b).

Abbildung 2 zeigt einen Überblick über Urteilsverzerrungen bei Kampfrichtern, eingeordnet nach den Stufen der sozialen Informationsverarbeitung (3.1) – Wahrnehmung, Kategorisierung, Gedächtnis- und

Urteils- und Entscheidungsprozesse. Es gibt darüber hinaus weitere Effekte, die im sportlichen Kontext aufzufinden sind (Plessner & Haar, 2006) und in dieser Arbeit nicht aufgeführt werden, da sie bezüglich Kampfrichterurteile eine eher geringe oder keine Bedeutung haben. Die dargestellten Verzerrungen stellen keine umfassende Aufarbeitung aller Urteilsfehler dar, sondern sollen denkbare Einflüsse auf Kampfrichterurteile aufzeigen. Sie können sich inhaltlich überschneiden, wodurch eine Trennung voneinander nicht immer klar ist.

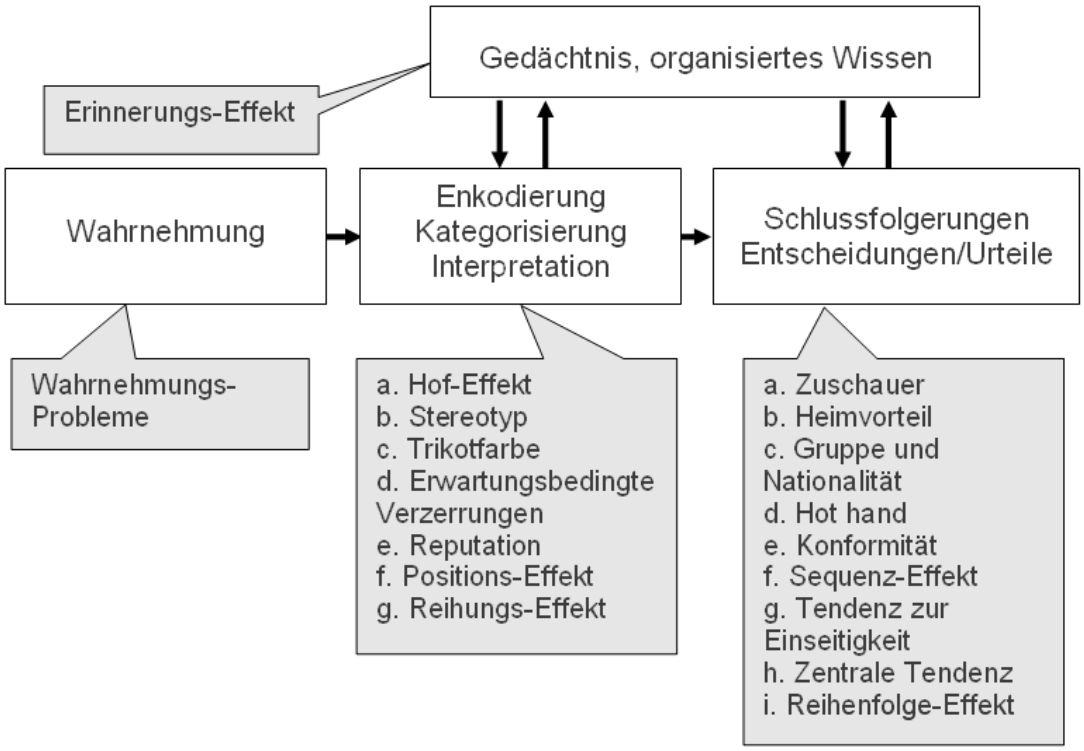


Abbildung 2: Übersicht möglicher verzerrender Einflüsse auf Kampfrichterurteile im Gerätturnen zugeordnet zu den Stufen der sozialen Informationsverarbeitung (eigene Darstellung, angelehnt an Fiedler & Bless, 2002, S. 133)

1. Wahrnehmung

Beobachter leiten ihre Urteile aus dem ab, was sie sehen. Sie hinterfragen meist nicht das Gesehene, sondern nehmen es als wahr an. Somit stellt bereits diese erste Stufe der Informationsverarbeitung eine wichtige Fehlerquelle dar, die im Folgenden genauer beschrieben wird.

Wahrnehmungsprobleme

Unterschiedliche *Blickwinkel* können uneindeutige Situationen schaffen. Eine Untersuchung zur Abseits-Entscheidung im Fußball hat gezeigt,

dass die Position des Schiedsrichters auf dem Spielfeld eine entscheidende Rolle spielt. Der Blickwinkel auf das Spielgeschehen und die damit für den Schiedsrichter zur Verfügung stehenden Informationen können zu objektiv gesehenen Fehlurteilen führen, obwohl aus Sicht des Unparteiischen eine richtige Entscheidung getroffen wurde (Oudejans et al., 2000). Ford, Gallagher, Lacy, Bridwell und Goodwin (1997) weisen im Baseball den Einfluss der Beobachtungsperspektive auf die Genauigkeit von ‚ball-strike'-Entscheidungen nach und belegen den Vorteil einer Position, die weiter weg und höher ist als die übliche.

Dasselbe Phänomen konnte in ähnlicher Weise auch im Gerätturnen aufgedeckt werden. Plessner und Schallies (2005) untersuchten unter anderem den Einfluss des Blickwinkels der Kampfrichter beim Bewerten von Übungen an den Ringen. Die Probanden bewerteten Photographien eines Halteelements, die aus unterschiedlichen Blickwinkeln aufgenommen wurden. Sie hatten die Aufgabe, Abweichungen der Arme bezogen auf die Horizontale anhand von Winkelgradangaben anzugeben. Eines der Ergebnisse besagt, dass sich Experten signifikant vom Blickwinkel beeinflussen lassen. Mit zunehmender Abweichung des Blickwinkels von der frontalen Ansicht nimmt auch die Fehleranfälligkeit zu. Je schräger somit die Aufnahme des Athleten ist, desto stärker weichen die angegebenen Winkelabgaben der Kampfrichter von den tatsächlichen ab. Das subjektive Urteil des Kampf- oder Schiedsrichters kann demnach in dessen Augen ein objektives darstellen. Die beiden Beispiele können auf die unterschiedlichen Perspektiven der am Wettkampf beteiligten Beobachtergruppen ausgedehnt werden. Es verwundert somit nicht, dass es schnell zu Diskrepanzen zwischen Zuschauern, Trainern, Athleten und den Unparteiischen kommt, nehmen die einzelnen Parteien unterschiedliche Positionen ein.

Aber nicht nur der Blickwinkel, sondern auch Faktoren, die durch technische Hilfsmittel ermöglicht werden, kommen als ein weiteres Beispiel in Frage, wenn Meinungsverschiedenheiten entstehen. Der Fernsehzuschauer sieht *Zeitlupenaufnahmen*, die der Schiedsrichter im Entscheidungsmoment auf dem Spielfeld nicht zur Verfügung hat. Technische Hilfsmittel sollen im Zweifelsfall die Einschätzung darüber erleichtern, welches das richtige Urteil ist, und können oftmals Abhilfe schaffen. Aber es gibt auch Situationen, die auch mithilfe von ausführlichen Videoanalysen uneindeutig bleiben. Studien in den Sportarten Tennis (Jendrusch, 2002), Fußball (Nevill, Balmer & Williams, 2002) und Handball (Jendrusch, Schmidt, Wilke & de Marées, 1993) haben gezeigt, dass es trotz technischer Hilfen oft zu Fehlurteilen kommt.

Begründet werden können diese Fehlurteile unterschiedlich. Jendrusch (2002) hat mit seiner Studie im Tennis gezeigt, dass es dem menschlichen Auge nicht möglich ist zu sehen, ob der Ball auf der Linie oder knapp daneben ist. Die im Spiel generierten Geschwindigkeiten des Balls überfordern die menschliche Wahrnehmungsfähigkeit (2.1).

Vermutet werden kann, dass dieser Faktor auch im Gerätturnen von Bedeutung ist. Gerätturnen ist eine Sportart, die durch sehr hohe Geschwindigkeiten des Körpers bzw. einzelner Körperteile der Athleten charakterisiert werden kann. So ist vorstellbar, dass das menschliche Auge auch hier überfordert ist. Nicht immer ist beispielsweise klar, ob am Boden die Begrenzungslinie übertreten wurde. Dabei führt dieser Fehler des Athleten zu einem Abzug und kann somit wichtige Zehntel eines Punktes bedeuten.

Die besser ausgeprägte visuelle Leistungsfähigkeit von erfahrenen Beurteilern im Vergleich zu weniger erfahrenen konnte in Untersuchungen im Handball (Jendrusch et al., 1993) sowie im Gerätturnen (Bard et al., 1980) belegt werden.

2. Kategorisierung

Die Kategorisierung stellt nach Fiedler und Bless (2002) die zweite Stufe des sozialen Informationsverarbeitungsprozesses dar. Die Zuordnung zu einem Bedeutungssystem, wie beispielsweise die Klassifizierung der Schwierigkeit oder von Fehlern, birgt für Kampfrichterurteile eine Reihe wissenschaftlich belegter Fehlerquellen.

a. Der Hof-Effekt

Menschen haben die Tendenz, Verhaltens- oder Charaktereigenschaften eher im Sinne einer globalen und nicht im Sinne einer differenzierten Eindrucksbildung wahrzunehmen (Asch, 1946). So auch beim Hof-Effekt der erstmals von Frederic Wells (1907) beobachtet und 1920 von Edward Thorndike benannt wurde. Er drückt die Tendenz aus, aufgrund eines dominierenden Einzelmerkmals den Gesamteindruck über eine Person zu bilden. Somit kommt es zur Übertragung eines Urteils von einem auf das andere Merkmal und führt zu einer gleichgerichteten Verfälschung der Beurteilung (Thorndike, 1920). Der Beurteiler hat dabei die Tendenz, frühere Urteile oder Überzeugungen eher zu bestätigen als zu widerlegen und somit widerspruchsfrei zu bleiben. Faktisch unabhängige Personeneigenschaften werden fälschlicherweise als zusammenhängend wahrgenommen. Der Effekt tritt meist auf, wenn sich der zu Beurteilende durch besonders hervorstechende, ausgeprägte Eigenschaften oder Verhaltensweisen auszeichnet. Besonders starken

Einfluss hat der Effekt, wenn der Beurteiler speziell auf diese Eigenschaften Wert legt. Weiterhin neigt der Beurteiler zu diesem Effekt, wenn das eigentlich zu beurteilende Merkmal nur schwer beobachtbar oder ungewöhnlich ist, nicht präzise genug definiert ist (Bortz & Döring, 2003) oder wenig Zeit zur Beurteilung eingeräumt wird. Folglich entsteht eine Über- oder Unterbewertung der beurteilenden Person.

Ein Beispiel für den Hof-Effekt ist, wenn ein Lehrer annimmt, dass ein freundlicher und gut aussehender Schüler auch gute Leistungen erbringt. Auch die Einschätzung von Übergewichtigen als gutmütig und Brillenträgern als klug sind Folgen des Effekts. Zu interessanten Hof-Effekten kommt es, wenn man vom äußeren Erscheinungsbild einer Person auf die Persönlichkeitsmerkmale schließt. Einfache periphere Merkmale werden herangezogen, wie die Attraktivität einer Person, um komplexe Sachverhalte zu beurteilen (Bless & Keller, 2006, S. 294).

Man kann sich vorstellen, dass auch Beurteilungen im Sport dem Einfluss der Attraktivität unterliegen. Dem Faktor experimentell nachgegangen ist Wilson (1976a), der zwar Tendenzen nachweist, aber aufgrund des ‚Vorstufen'-Charakters der Studie kaum Aussagen zu den Ergebnissen macht. Das äußere Erscheinungsbild, wie das tragen langer Haare als Mann oder eines Barts, ein schludriger Gang (Landers, 1970), eine überheblich wirkende Verhaltensweise oder eine Tätowierung könnten den Kampfrichter daran hindern, die höchst mögliche Wertung für den entsprechenden Athleten zu vergeben. Aber auch eine Überbewertung aufgrund eines sehr selbstsicheren Auftretens des Athleten vor dem Kampfgericht ist denkbar. Boen, Vanden Auweele, Claes, Feys und De Cuyper (2006) haben für das Gerätturnen die Übertragung bzw. die Verallgemeinerung der Leistung einer Dimension auf andere Dimensionen beschrieben. Dem Hof-Effekt experimentell nachgegangen sind Bormann (1975) und Moormann (1994). So könnte sich die auffallende Schnelligkeit der Bewegungen oder die Eleganz eines Turners auf die Annahme des Beurteilers übertragen, vergleichbar gute Leistungen bezüglich der Genauigkeit zu erwarten. Der Reputations-Effekt ist diesem recht ähnlich und wird im weiteren Verlauf erläutert.

b. Der Stereotyp

Ein Stereotyp ist die kognitive Komponente einer voreingenommenen Einstellung und ist definiert als eine Verallgemeinerung über eine Gruppe, wobei nahezu allen Mitgliedern identische Merkmale zugeordnet werden, ohne Rücksicht auf bestehende Variationen unter den Mitgliedern (Aronson, Wilson & Akert, 2004, S. 485). So besteht beispielsweise ein Stereotyp im nationalen Kontext „Die Deutschen sind

ein fleißiges Volk“. Diese Form der Eindrucksbildung lernen wir automatisch, denn sie erleichtert alltägliche soziale Wahrnehmungsprozesse und entlastet von kontinuierlichen Entscheidungs- und Prüfprozeduren. Stereotype sind somit Gruppenmeinungen, die durch den Gebrauch eine Zeitersparnis bei der Beurteilung anderer bewirken. Nachteilig ist, dass auf der Grundlage früherer Erfahrungen des Beobachters nicht klar ist, ob diese selbst gebildete Theorie im Einzelfall zutreffend ist. Anlass zu Stereotypisierungen geben gewöhnlich auffällige Merkmale wie die Hautfarbe, der Akzent oder das Geschlecht.

So belegen Untersuchungen zum Geschlechtsstereotyp im Handball (Souchon, Coulomb-Cabagno, Traclet & Rascle, 2004) und Fußball (Coulomb-Cabagno, Rascle, Souchon, 2005), dass die Urteilsbildung von Schiedsrichtern durch das Geschlecht beeinflusst wird. Aggressives Verhalten wird bei Spielerinnen stärker geahndet als bei Spielern.

Vorstellbar ist im Kontext Gerätturnen, dass sich entsprechende stereotype Sichtweisen, die dem Kampfrichter nicht bewusst sein müssen, das Urteil beeinflussen. Ein Beispiel stellt die mögliche Meinung von Kampfrichtern dar, dass ein gewisser Körperbau besonders gut geeignet bzw. nicht geeignet ist, an einem Gerät sehr gute Leistungen zu erbringen. Durch die Größe oder die Muskulatur des Athleten definiert, bestehen eventuell auf Erfahrungen zugrunde gelegte Vorstellungen darüber, wie denn ein ‚typischer Reckturner’ aussieht. Unbedacht könnte dabei bleiben, dass ein eher untypisch aussehender Turner ebenfalls sehr gute Leistungen am Reck erbringen kann. Allerdings liegen keine bekannten wissenschaftlichen Studien mit diesem Augenmerk vor.

c. *Die Trikotfarbe*

Frank und Gilovich (1988) thematisieren die Bedeutung der Farbe schwarz im sportlichen Wettbewerb in den Sportarten Fußball und Eishockey. Da schwarz in nahezu allen Kulturen als Farbe des Bösen und des Todes angesehen wird, untersuchen die Autoren, ob Mannschaften, die diese Farbe tragen, aggressiver sind als diejenigen die kein schwarz tragen. Sie weisen nach, dass die schwarz tragenden Mannschaften mehr Foulbeurteilungen und damit Strafen im Spielverlauf verhängt bekommen. Anhand zweier Laborexperimente zeigen sie, dass dieses Ergebnis auf zwei Ursachen zurückzuführen ist: die verzerrten Urteile der Schiedsrichter und die erhöhte Aggressivität der Spieler. Eine ähnliche Begründung, allerdings bezogen auf die Farbe rot, haben Hill und Barton (2005) als Grundlage ihrer Untersuchung im Kampfsport. Vermutet wird, dass die Farbe rot mit den Eigenschaften

Dominanz und Aggression gleichgesetzt wird. Die Ergebnisse der Olympischen Spiele 2004 in Athen wurden auf Grundlage der Trikot- bzw. Brustschutz-Farben rot und blau untersucht. Sie finden in den vier nachgeprüften Kampfsportarten Boxen, Taekwondo sowie Ringen in den Stilarten Freistil und griechisch-römisch signifikant mehr ‚rote' als ‚blaue Gewinner'. Auch im Fußball können die Autoren die überzufällige Wirkung der Farbe rot zeigen. Empirisch belegen Hagemann, Strauß und Leißing (2008) die Bedeutung der Trikotfarbe im sportlichen Wettkampf als wichtigen Einflussfaktor auf die Leistungsbewertung. Erfahrenen Taekwondo-Kampfrichtern wurden manipulierte Videos mit unterschiedlichen Kampfsequenzen zum Werten vorgelegt. Mittels Videobearbeitung wurden die Brust- sowie Kopfschutzfarben rot und blau der Kämpfer entsprechend vertauscht und damit variiert. Die identischen Szenen der beiden Videoversionen unterschieden sich somit lediglich durch die Farbe des Brust- und Kopfschutzes. Obwohl die Kampfrichter den identischen Kampf sahen, bewerteten sie die Leistungen der Sportler je nach getragener Farbe anders. Die Leistung der Taekwondo-Kämpfer mit roter Schutzausrüstung wurde besser bewertet als die der Athleten mit blauem Schutz, gerade wenn die Leistungen der beiden identisch waren. Im Durchschnitt erhielten die ‚roten Kämpfer' 13 Prozent mehr Punkte als die ‚blauen Kämpfer'.

Dem Einfluss der Trikotfarbe sehr ähnlich kann die *Kleidung* generell als beeinflussender Faktor herausgestellt werden. Aber auch die damit eng verbundene *Körpersprache* hat zu Untersuchungen geführt, die eine entsprechend verzerrende Wirkung aufzeigen konnten (Greenlees, Buscombe, Thelwell, Holder & Rimmer, 2005).

Es ist denkbar, dass sich die Farbe der Bekleidung auch in den technisch-kompositorischen Sportarten auf das Urteil der Kampfrichter niederschlägt. So kann vermutet werden, dass eine besonders elegante Turnerin durch ihren Turnanzug diese Eigenschaft betont oder eine explosiv turnende Athletin durch einen Anzug in auffälliger Farbe noch temperamentvoller wirkt. Im Gegensatz dazu, kann ein ungepflegter Anzug einem Athleten einen negatives Gesamtbild und vielleicht auch ein geringeres Urteil verschaffen. Gerade in den ästhetischen Sportarten könnte sich dieser Faktor beeinflussend auswirken, wurde aber bis dato empirisch noch nicht untersucht.

d. Erwartungsbedingte Verzerrungen

Bevor der Kampfrichter einen Wettkampf wertet, erklärt er sich bereit, dies zu tun. Bereits in dieser Phase versucht er sich auf den bevorstehenden Einsatz einzustimmen. Dabei kommt es zu einer Art

Orientierungsleistung, die Erwartungen in ihm hervorruft. Dabei werden automatisch alle Erfahrungen, die er in diesem Zusammenhang gemacht hat, einbezogen. Er versetzt sich, mehr oder weniger bewusst, in die entsprechend zu erwartende Situation. Diese individuelle ‚Färbung' kann allerdings zu Fehleinschätzungen führen. So ergibt sich für ihn eine besondere Situation, wenn zum Beispiel sehr leistungsfähige Sportler an den Start gehen. Er wird in diesem Fall besondere Leistungen erwarten, die er nicht vermutet, wenn keine Favoriten beteiligt sind (Thomas, 1978, S. 265). Aufgrund dieser antizipatorischen Prozesse haben es nicht favorisierte Athleten schwerer, für eine ausgezeichnete Leistung die entsprechende Benotung zu bekommen.

Der Einfluss von Erwartungen zeigt sich auch bezüglich der Startposition oder des Ansehens eines Athleten. Der sogenannte Positions-Effekt kommt aufgrund der Erwartung zustande, dass der beste Turner nicht zu Beginn, sondern am Ende eines Mannschafts-Wettkampfes eingesetzt wird. Dieser Einfluss wird gesondert unter f. erläutert. Die Reputation eines Sportlers kann ebenfalls zu Urteilsfehlern führen und wird im Folgenden beschrieben.

e. Die Reputation

Die genannten Erwartungen der Kampfrichter über die Leistungen der Turner lassen sich mit der Tendenz, die Bewertung eines Sportlers aufgrund seines Ansehens zu fällen (Boen et al., 2006), vergleichen.

Der Einfluss der Reputation aggressiver Mannschaften auf Schiedsrichterurteile im Fußball (Jones, Paull & Erskine, 2002) sowie der verfälschende Einfluss von ‚Star-Spielern' auf Urteile im Basketball (Lehman & Reifman, 1987) konnten empirisch nachgewiesen werden.

Im Gerätturnen spricht sich Landers (1970) dafür aus, dem Einfluss der Reputation wissenschaftlich nachzugehen. Die Kampfrichter leben nicht in einem Vakuum, dadurch gewöhnen sie sich an Athleten und die Tendenz steigt, eine Wertung bzw. einen Wertungsbereich mit dem Athleten zu verbinden. Unterstützend kommt hinzu, dass die Neigung besteht, ein einmal getroffenes Urteil beizubehalten. Speziell im Wettkampf wird der Einfluss der Reputation geschürt durch die spätere Startposition besserer Athleten (vgl. f) und die Lautsprecheransagen vor der Übung, die laut verkünden, welche Leistungen bzw. errungene Titel der Athlet bereits gewonnen hat. Nach Chaplan (1990) kann der bekannte Turner, der Turner einer bekannten Mannschaft oder der international angesehene Trainer davon profitieren. Dennoch ist es für einen unbekannten Turner möglich, einen bekannten Athleten zu

besiegen, dafür muss dieser allerdings bedeutend besser als der Bekannte sein. Der Unbekannte muss perfekt turnen, während der Bekannte ‚verturnen' muss. Turner einer bekannten Mannschaft bekommen höhere Wertungen als Turner einer weniger bekannten Mannschaft oder Turner ohne Mannschaft, da sie den Kampfrichtern besser in Erinnerung bleiben. Findlay und Ste-Marie (2004) untersuchten im Eiskunstlauf den Reputations-Effekt und belegen, dass unbekannte Sportler durchschnittlich 0,17 Punkte weniger in ihrer technischen Wertung bekommen als bekannte.

Für Calkin (1979) ist kein Effekt so offensichtlich wie dieser von ihm als Bekanntheits- der Persönlichkeits-Effekt genannte Einfluss. In vielen Fällen verturnt ein Star eine Übung und wird dennoch mit einer guten Wertung bewertet. Bei Mannschaften lässt sich dieser Effekt auch feststellen.

f. *Der Positions-Effekt*

Die auf der Startposition beruhenden Effekte (Ansorge et al., 1978; Bruine de Bruin, 2005, 2006; Calkin, 1979; Moormann, 1994; Plessner, 1997, 1999; Scheer, 1973; Scheer & Ansorge, 1975, 1979, 1980; Wilson, 1977), wurden in der Literatur am häufigsten genannt und stellen einen weiteren Einflussfaktor auf Kampfrichterurteile dar. Im Gerätturnen entwickelte sich im Laufe der Zeit ein ungeschriebenes Gesetz der Startreihenfolge bei Mannschaftswettkämpfen. Die Trainer stellen dabei meist die Turner entsprechend ihrer Leistungsfähigkeit am jeweiligen Gerät auf. Der schwächste Turner beginnt den Wettkampf, während der stärkste Turner ihn beendet. Dadurch entsteht bei den Wertungsrichtern eine bestimmte Erwartung an die Leistungen der Turner entsprechend dieser Reihenfolge (Boen et al., 2006). Nachgewiesen werden konnte dieser Effekt von Studien, die zeigen, dass dieselbe Übung an erster Stelle schlechter bewertet wird, als wenn sie an letzter Startposition im Wettkampf geturnt wird. Empirische Belege für den Positions-Effekt wurden für das männliche (Ansorge et al., 1978; Plessner, 1997, 1999; Scheer, 1973; Scheer & Ansorge, 1975) sowie für das weibliche Gerätturnen (Ansorge et al., 1978) und für das Synchronschwimmen (Wilson, 1977) ermittelt. So werden Punkteabweichungen berichtet, die sich zumeist zwischen einem und zwei Zehntel eines Punktes bewegen und durchaus praktische Relevanz haben. Wilson (1977) berichtet von signifikanten Unterschieden zwischen der ersten und der zweiten oder dritten Gruppe im Synchronschwimmen, allerdings nur bei den WM-Schwimmerinnen und nicht bei den Amateurinnen. Eine Untersuchung im männlichen Gerätturnen berichtet von signifikanten Unterschieden in

den Wertungen, allerdings nur an bestimmten Geräten (Scheer, 1973). Um zu überprüfen, ob es abhängig vom gewählten Gerät zum Positions-Effekt kommt, wirft Plessner (1997, 1999) erstmalig die Unterscheidung langsamer und schneller Geräte auf (5.2.1). Dabei zeigt er den Positions-Effekt bei schnellen, nicht aber bei langsamen Geräten. Es lässt sich ein Positions-Effekt an den schnellen Geräten nachweisen, der für den fünften Turner im Vergleich zum ersten, durchschnittlich eine um zwei Zehntel Punkte höhere Wertung zur Folge hat. Damit zeigt sich ein kleiner bis mittlerer Effekt (d = 0,40, Bortz & Döring, 2003). Calkin (1979) benutzt den Begriff ‚Treppen-Effekt' und merkt an, dass der Effekt nur entsteht, wenn es sich um knappe Unterschiede handelt, nicht aber wenn Übungen verturnt werden.

g. Der Reihungs-Effekt

Der von Calkin (1979) als Mehrfach-Sitzungs-Effekt bezeichnete Einfluss ist kein separater Effekt, sondern eine Version des Positions-Effekts. Wenn man sich die Erklärung der Entstehung des Positions-Effekts genauer ansieht fällt auf, dass das ungeschriebene Gesetz der ‚Startreihenfolge' seitens der Trainer Grund für die verzerrte Beurteilung durch die Kampfrichter ist. Der vorliegende Effekt könnte einen entgegengesetzten Erklärungsansatz liefern. Bei großen Veranstaltungen gibt es mehrere Durchgänge und oft kann man erkennen, wie die Kampfrichter im Verlauf der Veranstaltung von Durchgang zu Durchgang ‚milder' werden (Pflughoeft, 1984). Die Wertungen im letzten Durchgang sind signifikant höher, als die des ersten Durchgangs (Bruine de Bruin, 2005). Die ersten Beurteilungen fallen damit strenger aus als die letzten. Somit werden zu Beginn gezeigte gute und sehr gute Leistungen unterbewertet. Da später turnende Athleten besser sein könnten und man so die Überlegenheit ausdrücken können muss, wird ‚nach oben Luft gelassen' (Chaplan, 1990). Damit handeln die Trainer auf die ansteigende Beurteilung der Kampfrichter, indem sie ihre besten Athleten später turnen lassen und somit davon profitieren. Das hier angesprochene ‚Henne-Ei-Problem' kann im Rahmen dieser Arbeit nicht geklärt werden.

3. Gedächtnisprozesse

Die Art und Weise, wie Informationen gespeichert und organisiert werden stellt eine weitere Fehlerquelle dar. Das erworbene Vorwissen beeinflusst und schränkt die Speicherung neuer Informationen ein. Systematische Erinnerungsvorteile einzelner Informationen können infolgedessen das Urteil verändern.

Der Erinnerungs-Effekt

Eine Reihe von Untersuchungen hat nachgewiesen, dass das Wissen um die Leistung eines Turners zur Beeinflussung der Wertung führt (Boen et al., 2006). Vor dem Wettkampf gewonnene Informationen über die Leistungen der Athleten, beispielsweise während des Einturnens, formen die spätere Beurteilung der Übungen. Verschiedene Studien haben gezeigt, dass eine vorher in einer Lernphase präsentierte Leistung die Testphase beeinflusst (Ste-Marie, 2003; Ste-Marie & Lee, 1991; Ste-Marie & Valiquette, 1996; Ste-Marie, Valiquette & Taylor, 2001). In beiden Phasen besteht die Aufgabe der Kampfrichter darin, die Ausführung der gezeigten Bewegungen als perfekt oder fehlerhaft einzustufen. In der Testphase sind entweder dieselben oder aber andere Ausführungen der Bewegungen zu sehen. Die größten Urteilsfehler sind dann zu finden, wenn sich die Ausführung der Testphase von der Lernphase unterscheidet. Wenn die Ausführung einer Bewegung in der Lernphase fehlerhaft ist, wird sie in der Testphase auch als fehlerhaft eingestuft, obwohl sie perfekt geturnt wird. Dieser Effekt ist auch nach bis zu einer Woche Pause zwischen Lern- und Testphase nachweisbar (Ste-Marie & Valiquette, 1996). Aus diesen Erkenntnissen kann beispielsweise abgeleitet werden, dass die Beobachtung des Trainings, auch einige Tage vor dem Wettkampf, zu ähnlichen Urteilsfehlern führen kann. Die Fehler tauchten ebenso auf, wenn die Ausführung der Lernphase nur beobachtet, nicht aber bewertet wird (Ste-Marie et al., 2001). Auch nach einer Veränderung von äußeren irrelevanten Merkmalen, wie etwa der Kleidung, konnte der Effekt gezeigt werden (Ste-Marie, 2003). Überraschend erwies sich, dass das Wissen des Sachverhalts und die zusätzliche Aufforderung sich nicht beeinflussen zu lassen, diesen Effekt nicht beseitigen konnten (Ste-Marie & Lee, 1991).

4. Urteils- und Entscheidungsprozesse

Auf der Grundlage aller Informationen, die wahrgenommen, kategorisiert und aus dem Gedächtnis abgerufen werden, wird abschließend ein Urteil gebildet. So können vielfältige irrelevante Einflussfaktoren, wie die Urteile anderer, die Urteilsfähigkeit des Kampfrichters beeinflussen. Außerdem kann auch eine vereinfachende Strategie, wie der Gesamteindruck, herangezogen werden (3.1) um das Urteil zu bilden.

a. Die Zuschauer

Ein erster Einflussfaktor, der die Urteilsbildung der Kampfrichter auf dieser Stufe der Informationsverarbeitung verändert, ist das Publikum. Zu Beginn der Arbeit wird die Situation beschrieben, in der das

Publikum eine Übung anders als die wertenden Kampfrichter beurteilt und dies auch laut äußert. Die Zuschauer sind Teil des sportlichen Wettkampfs. Der Einfluss der Zuschauer wird als direkt bezeichnet, wenn diese ihre Athleten durch lautes Anfeuern unterstützen (Alfermann & Würth, 2009). Sie können dadurch Einfluss auf die Urteile der Kampfrichter nehmen, wenn sie sich durch lautes Anfeuern, Applaus oder auch Auspfeifen bemerkbar machen. Die reine Anwesenheit hingegen stellt einen indirekten Einfluss auf das sportliche Geschehen dar (ebenda). Im Basketball (Lehman & Reifman, 1987) sowie im Fußball (Courneya & Carron, 1992; Nevill et al., 2002) durchgeführte Studien belegen, dass Zuschauer Schiedsrichterentscheidungen beeinflussen.

Im Gerätturnen ist dieser Einflussfaktor, über das anfängliche Beispiel dieser Arbeit hinaus, in anderer Weise denkbar (Chaplan, 1990; O´Brien, 1991). Viele Turnveranstaltungen haben, im Vergleich zu einem Sportspiel wie dem Fußball, relativ wenige Zuschauer. Aber auch die geringere Anzahl an Beobachtern kann in einer Sportart, die während der Präsentation einer Übung im Normalfall unter absoluter Ruhe stattfindet, einen großen Einfluss haben. Der wertende Kampfrichter kann in einer solchen Situation konzentriert arbeiten. Wenn allerdings mehrere Turner gleichzeitig an unterschiedlichen Geräten turnen, findet der Applaus nach einer Übung statt, während an einem anderen Gerät noch gewertet wird. Bei riskanten Stürzen während einer Übung neigen die Zuschauer ebenfalls dazu, laut ihr Mitgefühl zu äußern. Diese Zuschauerreaktionen können sich störend und damit negativ auf den Urteilsbildungsprozess des Kampfrichters auswirken und machen das Auftreten von Urteilsfehlern wahrscheinlicher als in einer ruhigen Umgebung.

Nicht nur die Anwesenheit des Publikums, sondern auch die Erwartungen der Zuschauer, die Anzahl an Beobachtern und weitere Merkmale der Zuschauer werden wissenschaftlich hinterfragt (Alfermann & Würth, 2009) und kontrovers diskutiert. Vor allem im Zusammenhang mit dem Heimvorteil, der im nächsten Abschnitt beschrieben wird, sind diese Aspekte von großem Interesse.

b. Der Heimvorteil

Sehr eng verbunden mit dem Einfluss der Zuschauer ist das Phänomen des Heimvorteils. Es besagt, dass die Mannschaft bzw. die Athleten, die vor heimischer Kulisse einen Wettbewerb bestreiten, bevorzugt behandelt werden. Empirische Befunde gibt es vermehrt in den Sportspielen. Auf der Grundlage der Nachspielzeit in der Fußballbundesliga untersuchten Sutter und Kocher (2004) dieses Phänomen und fanden

heraus, dass länger nachgespielt wird, wenn die Gastmannschaft statt der Heimmannschaft mit einem Tor in Führung liegt. Eine Übersicht stellen Courneya und Carron (1992), die in Sportarten wie Baseball, Football, Eishockey, Basketball und Fußball die Einflussfaktoren beim Heimvorteil zeigen. Balmer, Nevill und Williams (2001) haben anhand von Auswertungen der Winter Olympiaden der Jahre 1908 bis 1998 überprüft, inwieweit ein Heimvorteil für die ausrichtende Nation besteht. Belege für einen Heimvorteil konnten sie in den Sportarten Eiskunstlauf, Ski Freestyle, Skispringen, Ski Alpine und Kurzbahn-Eisschnelllauf finden, während in den Sportarten Eishockey, nordische Kombination, Skilanglauf, Bobfahren, Rennrodeln, Biathlon und Eisschnelllauf ein geringer oder kein Heimvorteil gefunden wurde.

Vertreter der Gegenposition nehmen an, dass der Heimvorteil überschätzt wird bzw. nur vorherrscht, wenn man daran glaubt. Die Selbstbestätigung von Prognosen (Schlicht & Strauß, 2003) und die erhöhte Selbstwirksamkeitsüberzeugung der Athleten (Strauß, 1999) führt dazu, dass sie bessere Leistungen erbringen. Der hohe Erwartungsdruck, der von den Zuschauern auf die Spieler der Heimmannschaft ausgeübt wird, kann mit steigender Bedeutung des Spiels belastend auf die Athleten wirken und diese behindern. Der Heimvorteil kehrt sich in einen Heimnachteil um (Baumeister & Steinhilber, 1984). Diese Leistungsverschlechterung in Drucksituationen stellt ein Phänomen dar, dass unter dem Begriff des ‚choking under pressure' diskutiert wird (Schlicht & Strauß, 2003).

Im Gerätturnen gibt es keinen empirischen Nachweis. Chaplan (1990) nimmt an, dass Turner die vor lauten Fans turnen ebenfalls von einem Heimvorteil profitieren und somit denen gegenüber bevorzugt sind, die vor leiser Kulisse turnen. Calkin (1979) beschreibt das Szenario, dass der Kampfrichter eine nur mäßig gute Übung sieht. Das Publikum jubelt aber nach dem Abgang euphorisch. Der Kampfrichter wundert und fragt sich, ob er etwas nicht richtig gesehen hat. Daraufhin vergibt er eine gute Wertung und begründet diese für sich „im Zweifelsfall für den Angeklagten". Weiterhin vermutet der Autor, dass Kampfrichter bessere Wertungen für die Heimmannschaft vergeben und weniger dazu neigen die Schwierigkeit der Übungen zu reduzieren. Denkbar sind weiterhin Aspekte, die auf den Athleten Einfluss nehmen und somit indirekt zu einem Heimvorteil führen, da der Athlet tatsächlich bessere Leistung zeigt. Beispielsweise die für ihn gewohnte Umgebung oder bei internationalen Veranstaltungen die Sprache, die gesprochen wird, und so vorkommende Probleme am Wettkampfort schneller behoben werden.

Aber auch zusätzliche Betreuer, die hauptsächlich bei Heimwettkämpfen vor Ort sein können. Einen besonders bedeutenden Einfluss nehmen die Geräte ein, die im Fall eines Heimwettkampfs meist die sind, an denen Wochen zuvor bereits trainiert wurde. Heimathleten nutzen bei internationalen Wettbewerben die Möglichkeit, fabrikneuer Geräte bereits Tage vorher, bevor alle anderen Athleten anreisen, zu testen und sich somit besser auf den Wettkampf vorzubereiten.

c. Die Gruppe und die Nationalität

Der Nationalitäts-Effekt beschreibt die Tendenz, die Athleten aus der eigenen Nation bzw. aus der eigenen Kultur zu bevorzugen (Boen et al., 2006; Landers, 1970). In unterschiedlichen Sportarten konnte dieser Einfluss auf das Urteil des Unparteiischen empirisch gelegt werden. Die Australian Football League diente einer Untersuchung, die sich der Bevorteilung der eigenen Gruppe bei Freistoßentscheidungen widmet (Mohr & Larsen, 1998). Dass eine Bevorzugung durch den Schiedsrichter aufgrund der kulturellen Ähnlichkeit im Fußball resultiert, zeigen Messner und Schmid (2007) in ihrer Untersuchung, indem sie sich die kulturellen Besonderheiten der Schweiz zunutze machen. Unterscheiden sich die zwei Mannschaften kulturell dadurch, dass die eine einen französisch- und die andere einen deutschsprechenden Hintergrund hat, hat die Mannschaft einen Vorteil, die aus derselben Kultur wie der Schiedsrichter entstammt. 1991 untersuchten Seltzer und Glass im Rahmen der Olympischen Winterspiele zwischen 1986 und 1988 im Eiskunstlauf 417 Starter bzw. Paare. Sie fanden heraus, dass die Kampfrichter signifikant höhere Wertungen an Teilnehmer vergeben, die aus ihrer Nation kommen. Die Ergebnisse decken sich mit denen im Rahmen der olympischen Winterspiele 1980 in Lake Placid (USA) im Eiskunstlauf (Fenwick & Chatterjee, 1981).

Im Gerätturnen belegen Ansorge und Scheer (1988), sowie Ste-Marie (1996), dass es einen Nationalitäts-Effekt gibt. Ansorge und Scheer (1988) beleuchten dabei die Realsituation der Olympiade 1984 in Los Angeles (USA) und stellen zwei Arten von Verzerrungen fest. Zum einen herrschen Verzerrungen zugunsten des Athleten des eigenen Landes der Kampfrichter vor. Zum anderen gibt es auch welche, die zuungunsten der Athleten eines anderen Landes als derer der Kampfrichter wirken. Um den Nationalitäts-Effekt zu überprüfen, vergleicht man jede Kampfrichter-Wertung mit denen der anderen. Gezeigt wird, dass ein Kampfrichter, Athleten seines Landes höher als mit dem der Durchschnittswert der anderen bewertet. Die Länder, die im Wettkampf am dichtesten an der Wertung der eigenen Athleten sind, werden ferner

niedriger als der Durchschnittswert eingestuft. Die höhere Bewertung der eigenen Athleten ist Grundlage der Untersuchung von Ste-Marie (1996). Sie vermutet einen nicht beabsichtigten Vorgang (Thomas, 1978) und erklärt die Verzerrung durch die wiederholte Sichtung der Übung, die automatisch zu einer besseren Bewertung aufgrund der größeren Vertrautheit führt. Sie findet keine signifikanten Belege für ihre Annahme, führt dies selbst auf die experimentellen Schwierigkeiten des Sachverhalts zurück.

d. Der ‚hot-hand'-Effekt

Ebenfalls zur Überbewertung eines Athleten oder einer Mannschaft bzw. zu einem milderen Urteil führt das ‚Hot-hand-Phänomen'. Es kennzeichnet den Glauben der Zuschauer, Sportreporter und auch Trainer, dass beispielsweise ein Basketballspieler nach zwei oder drei Treffern eine bessere Chance hat, einen erneuten Korb zu erzielen, im Vergleich zur entsprechenden Anzahl an missglückten Versuchen (Gilovich, Vallone & Tversky, 1985).

Im Gerätturnen vermutet Calkin (1979) ebenfalls eine derartige Wirkung, weist aber keinen empirischen Beleg dafür auf. Seiner Meinung nach folgt einer sehr guten Übung ein gewisser Grad an Aufregung. Wenn allerdings einer Mannschaft eine Reihe solcher Übungen gelingt, bewirkt diese eine elektrisierende Atmosphäre und führt zur Überbewertung der einzelnen Übungen.

e. Der Konformitäts-Effekt

Eine weitere an sich irrelevante Informationsquelle, die Kampfrichter in ihre Bewertungen mit einbeziehen, sind die Urteile anderer Kampfrichter. Der Konformitäts-Effekt beschreibt die Tendenz, die eigene Bewertung an die der anderen Kampfrichter anzupassen (Boen et al., 2006).

Scheer, Ansorge und Howard (1983) untersuchten mit manipulierten Videoaufnahmen, ob es Konformitätsurteile im Gerätturnen gibt. Die Untersuchung zeigt, dass sich die Versuchspersonen (VPn) an die Wertungen der anderen Wertungsrichter im gezeigten Video anpassen und dadurch Unterschiede von mehreren Zehntel einer Note entstehen. Diese Anpassung an die Gruppennorm kommt zustande, da jeder Kampfrichter nach jedem Wertungsdurchgang Informationen zum Urteil der Kollegen erhält und sich dadurch Auswirkungen auf das eigene Urteil ergeben können. Dieses System der offenen Wertung fördert Konformität bei den Kampfrichtern und konnte in einer Untersuchung im Synchronschwimmen (Vanden Auweele, Boen, De Geest & Feys, 2004) sowie einer Replikation der Studie im Rope Skipping (Boen et al., 2006)

nachgewiesen werden. Die Streuung der Kampfrichterurteile ist zu Beginn eines Wettkampfs größer als im weiteren Verlauf (Landers, 1970). Eine Untersuchung im Eiskunstlauf zeigte, dass sich die Kampfrichter sogar im Laufe einer Saison in ihrer Einschätzung bezüglich eines Eislaufstils aneinander annähern (Wanderer, 1987).

Im Gerätturnen und in anderen technisch-kompositorischen Sportarten wird die Konformität der Kampfrichter geradezu von den internationalen Wertungsvorschriften, dem Code de Pointage (CdP) gefordert, da Abweichungen von den Wertungen anderer Kampfrichter bestraft werden (2.3). Dadurch kann nicht davon ausgegangen werden, dass dieser Einfluss ohne eine Separierung der Kampfrichter zukünftig verhindert werden kann.

f. Der Sequenz-Effekt

Der Sequenz-Effekt beschreibt den Einfluss, den einmal gefällte Entscheidungen auf darauffolgende Urteile haben. Die Entscheidung, im Fußball eine Elfmeter-Strafe für ein Team zu verhängen, ist davon abhängig, wie in einer bereits vorherigen Spielsituation entschieden wurde (Plessner & Betsch, 2001). Sportliche Leistungsbeurteilungen hängen somit auch vom speziellen Kontext ab, in dem sie stattfinden. In einer Studie im Basketball wurden Schiedsrichtern anhand von Videoaufnahmen unterschiedliche Foulszenen präsentiert (Brand, Schmidt & Schneeloch, 2006). Diese sollten die Szenen dahingehend einschätzen, ob ein Offensiv-, ein Defensivfoul oder kein Foul zu sehen war. Manipuliert waren die Szenen in der Hinsicht, dass sie einzelnen im originalen Spielverlauf oder in zufälliger Reihenfolge gezeigt wurden. Die Ergebnisse zeigen, dass Schiedsrichter, die eine zufällige Abfolge der Foulszenen sehen, rigoroser Strafen verhängen als diejenigen, die die Szenen im Originalverlauf zu sehen bekommen.

Ob sich die Wertung eines Beurteilers beim sequenziellen Urteilen durch die vorherige Wertung beeinflussen lässt, wurde im Gerätturnen untersucht. Überprüft wurde, ob die Bewertung einer Übung durch die zuvor gezeigte Darbietung eines anderen Athleten beeinflusst wird und, ob somit die Strategie des Vergleichens eingesetzt wird (Damisch, 2004; Damisch, Mussweiler & Plessner, 2006). Die Untersuchungen zeigen, dass die wahrgenommene Ähnlichkeit der Turner festlegt, in welche Richtung das Urteil des Kampfgerichts abweicht. Die Nationalität dient als Unterscheidungsmerkmal und zeigt bei einer wahrgenommen Ähnlichkeit zwischen Kontroll- und Versuchsperson einen Effekt der Annäherung der Wertungen – einen Assimilationseffekt. Wenn beispielsweise zwei Athleten gleicher Nationalität nacheinander turnen und

der erste Turner eine gute Übung zeigt, bekommt der zweite Turner eine vergleichsweise gute Wertung. Wenn hingegen keine Ähnlichkeit wahrgenommen wird und die Athleten unterschiedlicher Nationen sind, kommt es zu einem Kontrast in der Beurteilung – einem Kontrasteffekt. Wenn der erste Turner eine schlechte Übung zeigt, wird die Übung des zweiten Turners besser bewertet, als wenn der erste eine gute Übung geturnt hätte.

g. Die Tendenz zur Einseitigkeit

Die Tendenz zur Einseitigkeit ist dadurch geprägt, dass der Urteiler systematisch eine zu hohe oder zu geringe Wertung vergibt. Beim *Mildefehler* neigt der Beurteiler zu einer positiven Bewertung (genauere Informationen siehe Jürgens, 2005). Ein Erklärungsansatz besagt, dass besonders gegenüber gut bekannten und sympathischen Personen dieser Fehler auftritt[12]. Beim *Strengefehler* hingegen herrscht die persönliche Neigung vor, im Zweifelsfall lieber etwas anspruchsvollere Maßstäbe anzulegen als zu milde. Diese Tendenz kann als Folge der Ablehnung einer Person auftreten oder wenn Spezialisten Leistungen beurteilen. Ein weiterer Erklärungsansatz besagt, dass diese Tendenzen Persönlichkeitseigenschaften entsprechen und milde von strengen Urteilern unterschieden werden können.

So stellt die eigene Erfahrung mit bestimmten Geräten im Turnen eine mögliche Begründung dieser Tendenzen dar (Haase, 1972). Die unterschiedlichen Anforderungen der Kampfrichter an die Leistungen an den einzelnen Geräten könnten sich bezüglich der eigenen Erfahrung mit dem Gerät unterscheiden. Hatte der Kampfrichter als aktiver Turner selbst Schwierigkeiten mit dem Gerät, wertet er vermutlich milder als er das an einem anderen Gerät tut.

h. Die zentrale Tendenz

Eine weitere Tendenz der Urteilsbildung beschreibt nicht die zu gute oder zu schlechte Einschätzung einer sportlichen Leistung, sondern die Vorliebe für mittlere Urteile (genauere Informationen siehe Jürgens, 2005). Die Beurteilungsskala wird meist dann eingeengt, wenn eine gewisse Unsicherheit gegeben ist und der Urteiler befürchtet, sein Urteil rechtfertigen zu müssen. Wenn eine zu beurteilende Leistung nur teilweise oder überhaupt nicht gesehen wurde, eine ungerechte

[12] Gerade an diesem Beispiel lässt sich der enge Zusammenhang zur Reputation zeigen. Wechselwirkungen unterschiedlicher Einflüsse werden keine Beachtung geschenkt, könnten sich allerdings als spannendes Forschungsfeld erweisen.

Benotung vermieden werden soll und / oder der Beurteiler zu gutmütig oder feige ist, ein differenziertes Urteil abzugeben. Ziel der mittleren Bewertung ist es, eine möglichst geringe Abweichung zu den Urteilen der Mitbewerter zu erreichen. Diese Tendenz ist ein möglicher Erklärungsansatz für den Konformitäts-Effekt, der bereits beschrieben wurde.

i. Reihenfolge-Effekt

Der Reihenfolge-Effekt stellt den thematischen Mittelpunkt dieser Arbeit dar und wird anhand eigener Studien experimentell bewährt. Sequenzielle Informationen, die in schneller Folge präsentiert werden, führen dazu, dass der Beurteiler nicht alle Informationen als gleichbedeutend erachtet. Durch eine variierende Reihenfolge dieser Informationen erfolgt auch eine unterschiedliche Bewertung. Damit werden die ersten oder die letzten Informationen als wichtiger für das Urteil eingestuft. Möglich ist aber auch, dass die Reihenfolge schneller Bewegungen keinen Einfluss auf das Urteil hat. Diese Verzerrung ist dem Positions-Effekt recht ähnlich, unterscheidet sich aber dadurch, dass nicht aufgrund von Erwartungen bestimmter Leistungen ein Urteilsfehler resultiert, sondern aufgrund einer fehlerhaft ablaufenden Urteilsbildung.

3.3 Der Reihenfolge-Effekt

Der Reihenfolge-Effekt stellt ein Phänomen des täglichen Lebens dar und lässt sich in sehr unterschiedlichen Bereichen finden. Um hier nur einige zu nennen: In der Wirtschaft wie etwa beim Vorstellungsgespräch oder im Verkaufsgespräch, in Rundfunk und Fernsehen wie beispielsweise in den Fernsehnachrichten oder in der Werbung, im Bereich der Kommunikation und Persuasion sowie in der Schule, in der Rechtssprechung aber auch im Leistungssport begegnet man dem psychologischen Phänomen (Hogarth & Einhorn, 1992). Hogarth und Einhorn (1992, p. 3) definieren den Reihenfolge-Effekt wie folgt:

> “There are two pieces of evidence, A and B. Some subjects express an opinion after seeing the information in the order A - B; others receive the information in the order B - A. An order effect occurs when opinions after A - B differ from those after B – A”.

Damit beschreiben die Autoren, dass die sequenzielle Position und somit die Reihenfolge, in der Informationen präsentiert werden, einen bedeutenden Einfluss auf die Meinung hat, die man sich über einen bestimmten Sachverhalt bildet. Im Marketing der einzelnen Fernsehsender, macht man sich darüber Gedanken welche Werbespot-Reihenfolge dem Zuschauer präsentiert wird. Dabei wird ein besonderer

Schwerpunkt auf die Anzeigen zu Beginn und zum Schluss eines Werbeträgers gesetzt, da diese besonders wertvoll zu sein scheinen (Nieschlag, Dichtl & Hörschgen, 1988). Darin spiegelt sich das psychologische Phänomen in seinen zwei möglichen Ausprägungen wider. Zum einen der auch im Deutschen als Primacy-Effekt (Zimbardo, 1995) genannte Effekt und zum anderen der Recency-Effekt.

In der kognitionswissenschaftlichen Forschung wurde Ebbinghaus (1885) erstmals auf den seriellen Positions-Effekt aufmerksam, der ein Phänomen der Gedächtnisforschung beschreibt und durch die beiden erwähnten Effekte gebildet wird. Die Reproduktionsleistung, d.h. die Erinnerung von VPn, hängt von der Position des präsentierten Items ab. Die ersten Wörter werden sehr gut erinnert, die letzten ausgezeichnet und die mittleren werden eher schlecht wiedergegeben.

Im Folgenden werden die beiden Reihenfolge-Effekte, Primacy- und Recency-Effekt, erläutert. In diesem Zusammenhang erfolgt eine Beschreibung des Effekts und relevante Untersuchungen werden vorgestellt. Weiterhin werden mögliche Erklärungen zur Entstehung bzw. Bedingungen, unter denen der Effekt wahrscheinlicher ist, aufgeführt und die Relevanz des Effekts dargestellt.

Der Primacy-Effekt

Die ersten Informationen beeinflussen das Urteil über eine andere Person stärker als die nachfolgenden Informationen. Sie bestimmen somit die Richtung des Gesamteindrucks und führen zu einer verzerrten Interpretation der folgenden Informationen (Asch, 1946; Lund, 1925).

Eine der klassischen Untersuchungen zum Reihenfolge-Effekt stammt von Asch (1946). Der *Primacy-Effekt* wurde erstmals von ihm dargestellt. Wenn man sich ein Urteil über eine Person bildet, nimmt der Gesamteindruck eine bedeutende Position ein. In Untersuchungen entdeckte Asch, dass Einzelinformationen aufgrund ihrer zeitlichen Position bei der Präsentation unterschiedlich stark in den Gesamteindruck der Probanden eingehen. Er beschrieb bei der ersten Probandengruppe eine Person, beginnend mit positiven Eigenschaften, als intelligent (intelligent) – fleißig (Industrious) – impulsiv (Impulsive) – kritisch (critical) – eigensinnig (stubborn) – neidisch (envious), bei der anderen Hälfte als neidisch – eigensinnig – kritisch – impulsiv – fleißig – intelligent. Inhaltliche Unterschiede bestehen nicht, lediglich die Reihenfolge der Adjektive ist eine andere. Die Probanden der ersten Gruppe beurteilten die beschriebene Person durchweg positiver, als diejenigen in der

zweiten Gruppe. Jones, Rock, Shaver, Goethals und Ward (1968) untersuchten diesen seriellen Reihenfolge-Effekt unter realistischeren Bedingungen: Sie ließen Probanden einen Test zweier Person beobachten. Unter Bedingung eins konnte die getestete Person zu Beginn fast alle Fragen beantworten, ließ aber in der zweiten Hälfte stark nach. Unter Bedingung zwei hatte die getestete Person einen schlechten Start, konnte den zweiten Teil der Fragen sehr gut beantworten. In beiden Tests erreichte die Person jeweils 15 richtige Antworten. Die Beobachter hielten die Person unter Bedingung eins für intelligenter und räumten ihr auch bessere Chancen für weitere Tests ein (siehe auch McAndrew, 1981). Greenless, Dicks, Holder und Thelwell (2007) haben in ihrer Untersuchung im sportlichen Kontext einen Primacy-Effekt aufgedeckt.

Informationen, die am Anfang einer Liste stehen, haben einen größeren Einfluss auf die Urteilsbildung als diejenigen Informationen, die in der Mitte dieser Liste präsentiert werden (Anderson & Hubert, 1963). Doch wie kommt es zum starken Einfluss der ersten Informationen? In der Literatur sind unterschiedliche Erklärungsansätze und situationale Bedingungen für das Zustandekommen eines Primacy-Effekts zu finden:

1. Der *Aufmerksamkeitsverlust über die Serie:* Die Aufmerksamkeit des Beurteilers lässt im Laufe der Informations-Präsentation nach (Anderson, 1965; Hogarth & Einhorn, 1992; Kruglanski & Webster, 1996). Spätere Informationen erhalten weniger Aufmerksamkeit und haben einen geringeren Einfluss auf das Urteil (siehe auch Anderson & Hubert, 1963). Vielleicht aufgrund von Langeweile oder der Annahme, dass die wichtigen Informationen zu Beginn präsentiert werden, kommt es zu diesem Effekt (Richter & Kruglanski, 1998). Dies ist die Annahme klassischer Informations-Integrations-Modelle, nach denen das relative Gewicht der ersten Information über eine Person größer als die nachfolgenden Informationen ist.
2. Die *Abwertung inkonsistenter Informationen:* Wenn Beurteiler annehmen, dass nicht alle Informationen die gleiche Zuverlässigkeit besitzen, ergibt sich die Tendenz, darauffolgende Informationen, die mit den früheren inkompatibel sind, in ihrer Bedeutung herabzusetzen oder gar zu ignorieren (Nisbett & Ross, 1980).
3. Die *Assimilierung:* Anfängliche Informationen werden als eine Art Anker verwendet und beeinflussen spätere inkonsistente Informationen, indem sie im Licht des ersten Eindrucks interpretiert werden und einem Bedeutungswandel unterliegen. Diese Auffassung vertritt Asch (1946), der behauptet, dass die Bedeutung der späteren

Adjektive durch die zuerst präsentierten Adjektive geformt wird; d.h. eigensinnig wird positiv als eigenständig kodiert, weil die betreffende Person intelligent ist und ihren Mitmenschen zu Recht häufig widerspricht.

4. Die *Art der Verarbeitung:* Wenn ein Urteil in einem ‚end of sequence'-Prozess gebildet wird und dazu zunächst alle Informationen gesammelt werden, bevor es zu einem Urteil kommt und damit die Möglichkeit für ein Zwischenurteil nicht besteht, ist die Wahrscheinlichkeit für das Zustandekommen eines Primacy-Effekts recht groß (Hogarth & Einhorn, 1992).
5. Das *Bedürfnis nach kognitivem Abschluss:* Das unspezifische Bedürfnis, ein Urteil zu bilden, ist motivationaler Natur und kann anhand eines Kontinuums von dem starken Bedürfnis ein Urteil zu fällen, bis zu dem starken Wunsch, einen solchen Abschluss zu vermeiden, beschrieben werden (Richter & Kruglanski, 1998). Urteiler wollen zumeist ein Urteil schnell fällen und situationale Bedingungen wie Zeitdruck oder mentale Müdigkeit (Webster et al., 1996) schüren diese Tendenz.
6. Die *mentale Ermüdung:* Es kommt zu einem Konzentrationsabfall über die Zeit, der durch die psychische Ermüdung erklärt werden kann (Anderson, 1965; Anderson & Hubert, 1963; Hogarth & Einhorn, 1992; Kruglanski & Webster, 1996; Scheer, 1973; Webster, Richter & Kruglanski, 1996).
7. *Zeitdruck:* Es kommt eher zur Tendenz, die ersten Informationen stärker in das Urteil zu integrieren, wenn das Urteil unter Zeitdruck gefällt wird (Kruglanski & Webster, 1996). So ergibt sich auch bei der Beurteilung sportlicher Leistung die Situation, dass sehr viele Athleten in nur kurzer Zeit bewertet werden müssen.
8. Die *Stabilität des Gebildes Leistungsfähigkeit*: Dieser Faktor scheint in der Reihenfolge-Effekt-Thematik ebenfalls einen bedeutenden Stellenwert zu haben. Wird die Fähigkeit des zu Beurteilenden als stabil erachtet, führt das zu einem Primacy-Effekt (Jones et al., 1968; Jones & Welsh, 1971; McAndrew, 1981).

Zusammengefasst scheint ein Primacy-Effekt recht wahrscheinlich aufzutreten, wenn die ersten Items beim Urteiler einen anfänglichen Eindruck hervorrufen. Dieser Eindruck ist so fest, dass er dazu führt, darauffolgende Items, die dem Eindruck widersprechen, zu ignorieren oder subtile Veränderungen der Bedeutung vorzunehmen. Die Annahme

besteht, dass der Primacy-Effekt durch die frühe Bildung einer Hypothese entsteht (Nisbett & Ross, 1980).

Ein Beispiel aus dem alltäglichen Leben, das sich mit der Erklärung der Abwertung inkonsistenter Informationen (Punkt 2) deckt, ist die Situation des Bewerbungsgesprächs. Der Personalchef fragt sich, ob ihm die Person sympathisch ist, ob sie arrogant wirkt oder er sie gar als unfähig für die ausgeschriebene Stelle einschätzt. Derartige Urteile können aufgrund des ersten Eindrucks resultieren. Der erste Eindruck stellt eine Form des Primacy-Effekts dar, dem viel Bedeutung geschenkt wird. Der oft bleibende Eindruck beeinflusst die Bewertung der folgenden Informationen über die andere Person. „Jene Verhaltensweisen, die einen ersten Eindruck widerlegen, [werden] vorübergehenden Konstellationen oder äußeren Ursachen [zugeschrieben]“ (Zimbardo, 1995, S. 701). Somit wird inkonsistentes Verhalten auf situative Merkmale zurückgeführt, während mit vorherigen Beobachtungen übereinstimmendes Verhalten eher auf stabile Persönlichkeitsmerkmale attribuiert wird. Wenn nun zwei Kandidaten im Vorstellungsgespräch ohne Pause direkt hintereinander sprechen und die Entscheidung erst später am Bewerbertag getroffen wird, ergibt sich recht wahrscheinlich ein Primacy-Effekt (Aronson et al., 2004, S. 239).

Der Recency-Effekt

Dieser Effekt beschreibt, im Gegensatz zum Primacy-Effekt, die besondere Gewichtung der letztgenannten Informationen in der Beurteilung anderer Personen (Cromwell, 1950). Dabei stellt eine Information nicht zwingend eine Eigenschaft dar, sondern kann auch, in einem weiteren Sinn, durch ein Argument oder gar eine Person repräsentiert sein.

Auch der Recency-Effekt konnte in einigen Bereichen nachgewiesen werden, wie in der Eindrucksbildung (Ashton & Ashton, 1988; Jones & Berglas, 1976; Richter & Kruglanski, 1998; Wyer, 1973), der Einstellungsänderung (Miller & Campbell, 1959), im medizinischen Kontext (Bergus, Levin & Einstein, 2002), im Kontext strategische Videospiele (Jones & Welsh, 1971) oder im Gerichtswesen (Costabile & Klein, 2005).

Informationen, die am Ende einer Liste stehen, haben dabei einen größeren Einfluss auf die Urteilsbildung als diejenigen Informationen, die in der Mitte dieser Liste präsentiert werden (Anderson & Hubert, 1963). Doch wie kommt es zum starken Einfluss der letzten Informationen? Unterschiedliche Bedingungen sind für das Zustandekommen eines

Recency-Effekts verantwortlich. Im Folgenden werden mögliche Erklärungsansätze vorgestellt:

1. Die *Aufmerksamkeit:* Der im Primacy-Effekt beschriebene ‚Aufmerksamkeitsverlust über die Serie' kann behoben werden und resultiert in einem Recency-Effekt, wenn die Aufmerksamkeit eines wenig motivierten Beurteilers gesteigert wird, er dadurch konzentriert die Situation bis zum Schluss verfolgt und so ein Urteil trifft (Kruglanski & Webster, 1996). Durch eine anfängliche Aufforderung von Personen, möglichst alle Teile der Präsentation gleich zu berücksichtigen, bevor ein Gesamturteil gefällt wird, und sich kein vorschnelles Urteil zu bilden, wird der Primacy-Effekt reduziert (Luchins, 1957). Zu einer Umkehr in einen Recency-Effekt kommt es, wenn diese Aufforderung nicht zu Beginn, sondern zwischen zwei Informationsdarbietungen formuliert wird (ebenda). Die entstehende Verzögerung zwischen der Informationsaufnahme und der Fällung des Urteils bewirkt diesen Effekt.
2. *Vergessensunterschiede*: Die zuletzt präsentierten Eigenschaften sind besser aus dem Gedächtnis abrufbar als frühere Informationen. Je weiter die anfängliche Information vom Zeitpunkt der Urteilsbildung entfernt ist, desto stärker fallen Behaltensunterschiede ins Gewicht. Dieser Erklärungsansatz basiert auf den Untersuchungen von Ebbinghaus (1885), der das Gedächtnis daraufhin untersuchte, wie Menschen Informationen vergessen (Miller & Campbell, 1959).
3. Der *Kontrasteffekt*: Eine Person die im Einzelurteil eine mittlere Bewertung erhält wird relativ positiv beurteilt, wenn ihr eine negativ bewertete Person vorausgeht und relativ negativ, wenn sie auf eine positiv bewertete Person folgt (Miller & Campbell, 1959). Dieser Kontrasteffekt kann die abschließende Beurteilung verschlechtern oder eben verbessern (Jones & Goethals, 1972).
4. Die *Art der Verarbeitung:* Der ‚step by step'-Prozess, in dem das Urteil mit Hilfe von Zwischenurteilen getroffen wird, begünstigt die Überbewertung späterer Informationen (Hogarth & Einhorn, 1992; Nisbett & Ross, 1980). Zwischenurteile entstehen, wenn die Urteilsaufgabe dem Urteiler das Anfertigen von Notizen während der Präsentation erlaubt, wie beispielsweise im juristischen Bereich oder bei Kampfrichtern in technisch-kompositorischen Sportarten. Durch diesen Prozess wird die Aufmerksamkeit nicht nur auf die ersten, sondern auf alle Informationen verteilt, wobei die letzten

Informationen am stärksten davon betroffen sind und damit stärker das Urteil beeinflussen.

5. Die *zusätzliche kognitive Aufgabe*: Weiterhin kommt es zu einem Recency-Effekt, wenn man während der Beurteilung zusätzlich eine kognitive Leistung zu erbringen hat oder eine Ablenkung stattfindet (Jones & Goethals, 1972; Luchins, 1957). In sportlichen Wettkämpfen haben die Beurteiler meist sehr viele Personen zu beurteilen und oftmals müssen von ihnen zusätzliche Aufgaben bewältigt werden. Unter diesen Bedingungen kommt es eher zu einem Recency-Effekt (Greenless et al., 2007).
6. Die *Logik der Entwicklung*: Wenn ein Lern- oder Entwicklungsprozess zu Grunde gelegt wird, erhält die spätere Information eine größere Bedeutung als die frühere, die in der Zwischenzeit durch den Prozess entwertet worden ist (Jones & Welsh, 1971).
7. Der *Faktor Stabilität der Leistungsfähigkeit* wurde bereits beim Primacy-Effekt genannt. Nimmt der Beurteiler an, dass die Fähigkeit des zu Beurteilenden eher labil ist, entsteht ein Recency-Effekt (Jones et al., 1968; Jones & Welsh, 1971; McAndrew, 1981). Zu einer Eliminierung des Recency-Effekts kommt es in diesem Zusammenhang, wenn der Urteiler persönliche Verantwortung für das Urteil übernimmt (Webster et al., 1996).

Die bedeutende Rolle dieses Effektes wird wiederum am Beispiel des Vorstellungsgespräches deutlich. Wenn zwischen den beiden erwähnten Bewerbern eine Pause gemacht wird und die Entscheidung direkt nach der zweiten Rede erfolgt, stellt sich die Situation für den zweiten Kandidaten als vorteilhafter dar. Die Wahrscheinlichkeit für einen Recency-Effekt ist hoch und die Beurteiler erinnern das zweite Gespräch besser als das erste (Aronson et al., 2004, S. 239). Im Marketing wird dieser Effekt dazu genutzt, um bestimmte Werbebotschaften oder Argumente hervorzuheben. Der letzte Werbespot im Kino vor dem Film oder das letzte Argument im Verkaufsgespräch, indem der Verkäufer durch ein starkes Argument einen entscheidungsschwachen Kunden vom Kauf überzeugt, bekommen demzufolge eine besondere praktische Relevanz.

Obwohl serielle Reihenfolge-Effekte in den zu Beginn des Kapitels genannten Forschungsfeldern eine bedeutende Rolle spielen, besteht Uneinigkeit darüber, *welcher Reihenfolge-Effekt am meisten vorherrschend ist.* Empirische Untersuchungen haben bisher keine eindeutige Klärung geliefert. Eine dominante Stellung hat der Primacy-Effekt

beispielsweise bei der Intelligenz-Einschätzung. In angewandten Bereichen, wie in der Medizin oder im Gerichtswesen, konnten Belege für beide Effekte gefunden werden. So herrschen unterschiedliche Meinungen vor. Den Primacy-Effekt belegen Webster et al. (1996) im wirtschaftlichen Kontext sowie Pieters und Bijmolt (1997) im Bereich der Fernsehwerbung, während sich Costabile und Klein (2005) vor Gericht für die Dominanz des Recency-Effekts aussprechen. Nisbett und Ross (1980) sind der Meinung, dass, gelegentlich die Reihenfolge der Präsentation von Informationen keinen Effekt hat, es manchmal zum Recency-Effekt kommt, aber am häufigsten der Primacy-Effekt zu finden ist.

Eindeutig ist lediglich der Befund, dass bei der Präsentation mehrerer Informationen die zu Anfang (Primacy-Effekt) und die zu Ende (Recency-Effekt) vorgelegten besser erinnert werden als diejenigen, die in der Mitte dargeboten werden (u.a. Anderson & Hubert, 1963).

Über die *Stabilität* des Reihenfolge-Effekts im Kontext Sport gibt es lediglich eine Untersuchung. Eine hohe Stabilität beschreibt den Sachverhalt, dass obwohl der Beurteiler weiß, dass der Einfluss der Reihenfolge in der speziellen Urteilssituation denkbar ist und sein Urteil direkt davon betroffen werden kann, sich der Effekt vom Beurteiler nicht willentlich verhindern lässt. Die Untersuchung zum Reihenfolge-Effekt im Fußball hat genau das Gegenteil und damit eine geringe Stabilität des Effekts belegt (Greenlees, Hall, Filby, Thelwell, Buscombe & Smith, 2009). 146 Fußballtrainer schätzten anhand von Videoaufnahmen die sportliche Leistung zweier Spieler ein, die eine einfache Fußballbewegung durchführten. Der eine Spieler diente als Kontrollperson und allen Probanden wurden dieselben Videosequenzen gezeigt. Bei dem anderen Spieler sah eine Hälfte der VPn manipulierte Videosequenzen, in denen die Leistung des Athleten abnahm, also die Reihenfolge positiv-negativ gezeigt wurde. Die andere Hälfte der Probanden sah im Gegensatz dazu einen aufsteigenden Leistungsverlauf, negativ-positiv. Zusätzlich bekamen die Trainer entweder keine Warnung über die Gefahr des Reihenfolge-Effekts, eine Warnung bevor sie irgendeine Sequenz zu sehen bekamen, eine Warnung bevor sie die Sequenzen des Testspielers sahen oder bevor sie die Testperson bewerten sollten. Die Ergebnisse zeigen einen signifikanten Primacy-Effekt in den Bedingungen, in der die Probanden nicht gewarnt wurden und in der sie erst vor der Bewertung gewarnt wurden. Bei der Warnung vor der Sichtung der Testpersonsequenzen zeigt die Studie keinen Reihenfolge-Effekt. Diese Studie deckt einen kleinen bis mittleren Reihenfolge-Effekt auf (η^2 =

0,04). Daraus schlussfolgern die Autoren, dass der Reihenfolge-Effekt durch eine Warnung eliminiert werden kann.

Insgesamt hängt es immer von der Situation ab, welcher der beiden Effekte stärker ausgeprägt ist. Hogarth und Einhorn (1992) haben aus unterschiedlichen Untersuchungen zum Reihenfolge-Effekt eine Theorie entwickelt, die konkrete Vorhersagen macht, unter welchen Bedingungen welcher Effekt zu erwarten ist. Diese Theorie und die entsprechenden Vorhersagen werden in Kapitel 3.4 vorgestellt.

3.4 Vorhersage von Reihenfolge-Effekten

In der Vergangenheit wurde der Prozess der Urteilsbildung oft im Zusammenhang mit unterschiedlichen Phänomenen, wie dem der Entscheidungsfindung diskutiert. Dabei wurde der Versuch unternommen diesen Prozess zu beschreiben sowie ihn vorherzusagen (Slovic, Fischhoff & Lichtenstein, 1977). Die folgend aufgeführten drei Modelle sagen den Reihenfolge-Effekt voraus.

1. *‚Model for opinion change' von Anderson (1959)*

Das mathematische, lineare Modell von Anderson (1959) macht eine grundlegende Vorhersage bezüglich der Reihenfolge-Thematik. Bei zwei aufeinanderfolgenden Informationen bzw. hintereinander vortragenden Sprechern, die dasselbe Potential haben, eine bestimmte Meinungsänderung hervorzurufen, geht Anderson davon aus, dass immer der zweite Redner die begünstigte Position hat, und somit der Recency-Effekt vorherrschend ist. Da die erste Rede die Meinung der Zuhörer zu einem gewissen Grad ändert, erhöht sich währenddessen der Meinungsunterschied zwischen den Zuhörern und dem zweiten Redner. Wenn der zweite Redner nun ebenso effektiv wie der erste Redner vorträgt, entsteht der Effekt, eine größere Meinungsänderung zu produzieren, da eine größere Menge an Meinungsänderung abverlangt wird. Wenn die Informationen einen geringen abfallenden Verlauf (positiv – negativ) zeigen, sagt das Modell einen weniger starken Recency-Effekt voraus. Wenn der abfallende Verlauf aber stark ist, kommt es zu einem Primacy-Effekt. Der Primacy-Effekt und Unstimmigkeiten in den eigenen Ergebnissen werden durch eine Zweikomponenten-Hypothese erklärt. Entsprechend diesen Hypothesen hat die Meinung eines Individuums eine vordergründige Komponente, die der Vorhersage des Modells folgt und eine grundlegende Komponente, die, sobald sie einmal entstanden ist, größtenteils änderungsresistent ist.

Dieses Modell erklärt den in der vorliegenden Arbeit zu untersuchenden Reihenfolge-Effekt nicht detailliert genug. Außerdem gibt es kaum empirische Belege, die die getroffenen Vorhersagen unterstützen. Die Anwendbarkeit scheint, aufgrund des komplexen Charakters sportlicher Leistungsbewertung, nicht in ausreichendem Ausmaß gegeben zu sein.

2. Miller und Campbell Modell (1959)

Im Forschungsfeld der Kognitionswissenschaft entwickelten sich unterschiedliche Modelle, die die Funktionsweise des Gedächtnisses mit ihren verschiedenartigen Phänomenen zu erklären versuchen. Grundsätzlich lassen sich Mehrspeicher-Modelle (Atkinson & Shiffrin, 1968) von Einspeicher-Modellen (Craik & Lockhart, 1972) unterscheiden. Einspeicher-Modelle gehen von der Theorie der Verarbeitungstiefe aus und nehmen Folgendes an: je intensiver bzw. tiefer eine Information verarbeitet wird, desto länger kann sie behalten werden. Mehrspeicher-Modelle gehen davon aus, dass das Gedächtnis aus einem Ultrakurzzeitspeicher (UKZS), einem Kurzzeit- (KZS) und einem Langzeitspeicher (LZS) zusammengesetzt ist. Die einzelnen Gedächtnisspeicher sind über verschiedene Prozesse miteinander verbunden (3.1). Mithilfe des Mehrspeicher-Modells (Atkinson & Shiffrin, 1968) lässt sich der von Ebbinghaus (1885) entdeckte serielle Positions-Effekt erklären. Der KZS führt eher zum Recency-Effekt, während der LZS einen Primacy-Effekt wahrscheinlicher macht. Diese Unterscheidung konnte experimentell durch das von Hermann Ebbinghaus (1885) entwickelte Verfahren und das in der gedächtnispsychologischen Literatur bekannte Reproduktionsverfahren des freien Erinnerns belegt werden. Dabei erhalten Probanden eine Reihe von Items, meist Wörter, die im Anschluss in beliebiger Reihenfolge wiedergegeben werden sollen. Der Anteil der erinnerten Wörter in Beziehung zur präsentierten Position stellt das Ergebnis dar. Es zeigt die beiden Reihenfolge-Effekte, die wie folgt erklärt werden:

Der *Primacy-Effekt* nimmt eine größere Rolle ein, wenn am Ende einer Reihe von Informationen eine Gesamtbeurteilung getroffen wird. Die erstgenannten Informationen gehen leichter in den LZS über, da noch keine Information eingegangen ist, die mit dem Abspeicherungsprozess im LZS interferieren und ihn negativ beeinflussen könnte. Sie verweilen länger im Gedächtnis, was den Übergang in den LZS fördert und sie somit nicht vergessen werden. Die Wahrscheinlichkeit der Reproduktion der ersten Informationen fällt daher höher aus als bei nachfolgenden Items aus dem Mittelbereich der Präsentation. Infolgedessen stellt dieser Effekt eine Funktion des *LZS* dar.

Der *Recency-Effekt* ist hingegen ein Output des *KZS*, der nur eine kurzfristige Speicherleistung aufweist. Die zuletzt dargebotenen Items verbleiben als letzte Erinnerung im Gedächtnis, da sie nicht durch nachkommende Information überschrieben werden. Somit wird eine bessere kognitive Auseinandersetzung mit der Information möglich und es entsteht eine höhere Wahrscheinlichkeit, unmittelbar abgerufen zu werden. Durch die zeitliche Nähe zwischen Präsentation und Wiedergabe sind die letzten Informationen noch im KZS und können daher zuerst wiedergegeben werden, bevor sie vergessen werden.

Die Wörter aus dem mittleren Bereich bleiben am schlechtesten in Erinnerung, da sie bei der Reproduktion nicht mehr im KZS und auch noch nicht im LZS abgelegt worden sind.

Miller und Campbell (1959) unterstützen diesen Ansatz und haben den Faktor ‚Zeit' in die Reihenfolge-Effekt-Diskussion eingebracht. Sie beziehen sich auf die Vergessenskurve von Ebbinghaus (1885), die bezüglich der Zeitachse einen zunächst steilen und dann immer flacher absinkenden Verlauf des Vergessens beschreibt. Sie leiten daraus ab, dass ein Primacy-Effekt entsteht, wenn ein kurzes Zeitintervall zwischen zwei inkonsistenten Informationen und ein großes Intervall zwischen der zweiten Information und dem Urteil entsteht. Zum Recency-Effekt kommt es hingegen, wenn ein größeres Zeitintervall zwischen den beiden Informationen und lediglich ein kurzes Intervall zwischen der zweiten Information und dem Urteil vorzufinden ist. Kein Effekt wird erwartet, wenn die Informationen und das Urteil direkt aufeinanderfolgen. Auch, wenn zwischen den Informationen sowie zwischen der zweiten Information und dem Urteil eine längere Zeitspanne liegt, sollte nach den Vorhersagen der Autoren kein Effekt sichtbar werden.

Das Modell von Miller und Campbell (1959) macht, wie schon das vorherige Modell von Anderson (1959), keine detaillierten Vorhersagen und ist daher eher nicht geeignet, die vorliegende Fragestellung zu beantworten. Weiterhin scheint die Forderung bestimmte Zeitabstände einzuhalten, und im sportlichen Kontext des Gerätturnens kaum durchführbar zu sein, da die Urteile unter Zeitdruck gefällt werden und längere Zeitintervalle eher unerwünscht sind.

Eine detaillierte Herangehensweise und entsprechende Vorhersagen bietet das nachstehend beschriebene Modell. Es scheint sich hervorragend für das zugrundeliegende angewandte Forschungsfeld im Sport zu eignen. Es wurde in unterschiedlichen Forschungsbereichen, wie auch im Sport (Greenless et al., 2007) zur Klärung herangezogen und

die Vorhersagen finden weitreichende empirische Unterstützung (Hogarth & Einhorn, 1992).

3. ‚Belief-adjustment'-Modell von Hogarth und Einhorn (1992)

Im Forschungsfeld der Sozialpsychologie haben Hogarth und Einhorn (1992) insgesamt 76 Studien[13] aus den unterschiedlichsten Bereichen wie dem Gerichtswesen, der Psychophysik oder der Wirtschaft auf den Reihenfolge-Effekt hin überprüft. Davon zeigten 36 Studien einen Primacy-Effekt und 35 einen Recency-Effekt. In fünf Untersuchungen konnte kein Effekt nachgewiesen werden. Somit lässt sich nicht analysieren, welcher der beiden Effekte vorwiegend auftritt. Ersichtlich wird jedoch, dass Reihenfolge-Effekte eine bedeutende Rolle spielen. Aufgrund des recht uneinheitlichen Ergebnisses haben die Autoren das ‚belief-adjustment'-Modell entwickelt. Es bezieht sich auf den Prozess der Urteilsbildung und sagt voraus, welcher Reihenfolge-Effekt unter welchen Bedingungen zu erwarten ist.

So unterscheiden die Autoren einzelne Einflüsse für das Auftreten der Reihenfolge-Effekte. Sie gehen von vier Variablen aus, die das Zustandekommen des Effekts beeinflussen. Mithilfe dieser Bedingungen kann eine Vorhersage getroffen werden, wann es zu einem Primacy- und wann zu einem Recency-Effekt kommt. Die vier Variablen, Bewertungstyp, Komplexität der Aufgabe, Länge der Informationssequenz und die Konsistenz werden im Folgenden unterschieden:

1. Der *Bewertungstyp* kann sich in zwei Ausprägungen zeigen. Zum einen, wenn das Urteil mit einem Zwischenurteil (step by step - SbS) und zum anderen ohne ein Zwischenurteil (end of sequence - EoS) gefällt wird. Im Urteilbildungsprozess mit Zwischenurteil kann man davon ausgehen, dass die augenblickliche Meinung eines Subjekts eine Art Anker darstellt. Dieser Anker wird durch neu eingehende Informationen angepasst, woraus ein neues Meinungsbild entsteht. Dieses bildet dann den neuen Anker und kann wiederum durch Informationen verändert werden.
2. Die *Aufgabenkomplexität* unterscheidet einfache von komplexen Aufgaben. Als einfach wird eine Aufgabe definiert, wenn nur ein einziges Item, sei es eine Charaktereigenschaft oder eine Zahl, als Stimulus-Material verwendet wird. Komplexe Aufgaben zeichnen sich dadurch aus, dass die Informationsmenge pro Item groß ist

[13] Diese Untersuchungen beinhalten teilweise mehrere Experimente.

oder der dargebotene Stimulus unbekannt ist. Die Informationsmenge wird bezüglich der vorherrschenden Expertise der VPn (Hogarth & Einhorn, 1992, p. 4) relativiert, indem sie hinzufügen: "…an important component of expertise in a specific domain involves strategies for coping more effectively with amounts of information". Somit haben ihrer Auffassung nach Experten Strategien entwickelt, mit denen sie effektiver größere Mengen an Informationen bewältigen können. Die Autoren merken zudem an, dass bei der Festlegung der Aufgabenkomplexität Schwierigkeiten auftreten können.

3. Die *Länge der Informationssequenz*, also die Serienlänge der einzelnen Items wird von den Autoren unterteilt in Sequenzen die kurz oder lang sind. Als kurz wird eine Sequenz bezeichnet, wenn sie zwischen zwei und 12 Items beinhaltet, während eine lange Sequenz 17 und mehr Items umfasst. Mit steigender Itemanzahl werden zwei Effekte erwartet. Zum einen können die VPn ermüden, wenn sie viele Informationen verarbeiten sollen. Zum anderen kann es zu einer Desensibilisierung kommen. Dabei haben die neuen Informationen einen im Verlauf immer geringeren Einfluss auf die Meinung des Subjekts. Begründet wird dies damit, dass im Vergleich zu den bereits verarbeiteten Informationen, die neu eintreffenden eine zunehmend geringere Aussagefähigkeit besitzen.
4. Die *Konsistenz* bzw. die Inkonsistenz der Informationskomponenten beschreibt die Zusammensetzung der Informationen. Sind alle Informationen durchgehend positiv oder negativ, werden sie als konsistent bezeichnet, während Informationen, die aus positiven und negativen Anteilen bestehen, inkonsistent sind. Die vorliegende Arbeit thematisiert die Beurteilung von sportlicher Leistung in technisch-kompositorischen Sportarten, in der konsistente Informationen, also Übungen, die komplett fehlerfrei sind, eher eine Ausnahme darstellen und daher nicht von Interesse sind. Aufgrund dessen werden weiterhin nur inkonsistente Informationen und die damit verbunden Vorhersagen berücksichtigt.

Vorhersagen des Models

Je nachdem, in welcher Ausprägung und Kombination die drei erst genannten Bedingungen in der Urteilssituation zusammen auftreten, lässt sich einer der beiden Reihenfolge-Effekte, Primacy- oder Recency-Effekt, prognostizieren. Tabelle 1 enthält die Vorhersagen des Modells im Überblick.

Tabelle 1: Vorhersagen des ‚belief-adjustment'-Modells (angelehnt an Hogarth & Einhorn, 1992, p. 7)

	Bewertungstyp	
Serienlänge & Aufgabenkomplexität	Ohne Zwischenurteil (EoS)	Mit Zwischenurteil (SbS)
Kurz & einfach	Primacy (19)	Recency (16)
Kurz & komplex	Recency (7)	Recency (2)
Lang	Eher Primacy	Eher Primacy

Anmerkung: die Zahl in der Klammer stellt die Anzahl der Studien dar, die den entsprechenden Reihenfolge-Effekt nachgewiesen haben (vgl. Tabelle 2) (EoS = End of sequence; SbS = Step by step).

Das Modell sagt für den *Bewertungstyp* mit Zwischenurteilen (SbS) vorher, dass ein Recency-Effekt entsteht. Bei einem EoS-Prozess (Bewertungsform ohne Zwischenurteile und damit am Ende der Informationspräsentation) soll es hingegen zu einem Primacy-Effekt kommen.

Die meisten Studien (43 von 76 Fällen) haben *einfache Items in kurzen Serien* untersucht, indem sie beispielsweise zwischen drei und sechs Charaktereigenschaften verwendeten, um soziale Beurteilungen, wie die Sympathie, zu untersuchen. Bei dieser Art von Aufgabe und einem EoS-Prozess besagt die Theorie, da von 27 Fällen 19 ein signifikantes Ergebnis hervorbrachten, einen Primacy-Effekt voraus. Bei einem SbS-Prozess wird der Recency-Effekt prognostiziert (16 von 16 Fällen).

Vom Bewertungstyp unabhängig stellt sich in einfachen, langen Aufgaben ein Primacy-Effekt ein, während bei komplexen, kurzen Aufgaben ein Recency-Effekt beobachtet werden kann (in 9 von 11 Fällen).

Komplexe, lange Aufgaben zeigen in den durchgeführten Untersuchungen, die im Artikel von Hogarth und Einhorn (1992) berücksichtigt wurden, ein vergleichsweise uneindeutiges Bild. Nur sechs Untersuchungen wurden mit dem Aufgabentyp (komplex und lang) durchgeführt. In drei Fällen zeigte sich ein Recency-Effekt, zwei Studien belegten einen Primacy-Effekt und in einem Fall konnte kein Effekt nachgewiesen werden. Die Autoren prognostizieren im Fall einer komplexen Aufgabe (unabhängig von der Serienlänge und dem Bewertungstyp) einen Recency-Effekt. Tabelle 2 stellt die Untersuchungsergebnisse zum Reihenfolge-Effekt dar.

Tabelle 2: Klassifizierung der Ergebnisse von Untersuchungen zum Reihenfolge-Effekt (Hogarth & Einhorn, 1992, p. 5)

Aufgabenkomplexität	**Einfach**		**Komplex**		**Gesamt**
Bewertungstyp	EoS	SbS	EoS	SbS	
Kurze Serie					
Primacy	**19**	-	1	-	**20**
Recency	5	**16**	**7**	**2**	**30**
Kein Effekt	3	-	1	-	**4**
Lange Serie					
Primacy	12	2	2	-	**16**
Recency	2	-	1	2	**5**
Kein Effekt	-	-	1	-	**1**
Gesamt	**41**	**18**	**13**	**4**	

Bezüglich der Abhängigkeit des Urteils vom Bewertungstyp gibt es unterschiedliche Ansichten. Für die Abhängigkeit des Reihenfolge-Effekts vom Bewertungstyp und der Expertise des Urteilers sprechen sich Adelman und Kollegen aus (Adelman, Tolcott & Bresnick, 1993), die mit Untersuchungen im militärischen Kontext das Modell von Hogarth und Einhorn (1992) unterstützen. Im Personalmanagement konnten die Vorhersagen des Modells bestätigt werden (Highhouse & Gallo, 1997). Die Autoren sprechen sich ebenfalls für den Einfluss des Bewertungstyps auf den Reihenfolge-Effekt aus. Das Forscherteam um Jones (Jones et al., 1968) sowie McAndrew (1981) und Bruine de Bruin (2005, 2006) sprechen sich hingegen dafür aus, dass die Art der Bewertung das Urteil nicht beeinflusst.

Im Forschungsfeld Sport konnte die Untersuchung von Greenless et al. (2007) die Vorhersagen des theoretischen Modells nicht bestätigen. Die Autoren untersuchten im Fußball anhand von Videoaufzeichnungen, die ein mehrmaliges einfaches Passspiel zweier Athleten zeigten, inwieweit die Bewertung der Spielleistung dieser Athleten von der Reihenfolge der präsentierten Informationen, vom Bewertungstyp und von der Vertrautheit mit der Aufgabe abhängt. Einer Hälfte der Probanden wurde ein abnehmender Leistungsverlauf gezeigt, also zunächst erfolgreiche und danach nicht erfolgreiche Handlungen, während die zweite Hälfte zunächst die nicht erfolgreichen und dann die erfolgreichen Szenen zu sehen bekam. Ein nicht erwartungskonformer, signifikanter Primacy-Effekt zeigte sich in den Ergebnissen. Die berichtete Effektgröße von η^2 = 0,42 wird als sehr groß eingestuft (Clark-Carter, 1997). Allerdings

muss dabei das dreifaktorielle Design (2x2x3) beachtet werden, das im Vergleich zu zweifaktoriellen Designs (5.1) zu höheren Effektgrößen führt (7.4). Zwei mögliche Erklärungen geben die Autoren für den belegten Primacy-Effekt ihrer Untersuchung an. Zum einen, da die Fähigkeit entweder als stabiles oder als rigides Merkmal angesehen werden kann. Wenn es als stabiles Gebilde betrachtet wird, fördert es eher den Primacy-Effekt und wenn es als rigides Gebilde eingeschätzt wird, kommt es eher zum Recency-Effekt. Für den Sport heißt das, dass beachtet werden sollte, ob die wahrgenommene Leistung als stabil oder labil angesehen wird. Die Beobachter können als eher Primacy-Effekt-anfällige und Recency-Effekt-anfällige Beurteilungstypen eingestuft werden. Je nachdem welches Konzept sie für die Erklärung von Leistung annehmen, entweder ‚die Leistung ist als stabiles Persönlichkeitsmerkmal anzusehen' oder ‚die Leistung ist veränderbar und abhängig von der Anstrengung', werden sie einer Urteilstendenz folgen. Zum anderen erklären sie, dass die Aufgabe der Probanden fälschlicherweise als eine kurze definiert wurde. Die Theorie von Hogarth und Einhorn (1992) sagt aber bei langen und komplexen Aufgaben eine Tendenz in Richtung Primacy-Effekt voraus.

Kritik am Modell

Kritisiert werden kann, wenn man sich die Definitionen der beschriebenen Bedingungen einmal genauer betrachtet, dass einige von ihnen eher subjektiver Natur sind. Eine Ungenauigkeit in der Definition der ‚(Aufgaben-) Komplexität' kann nicht klar von dem Parameter ‚Länge (der Informationssequenz)' abgegrenzt werden. Weiterhin ist nicht konkret berichtet, was ein ‚ungewohnter Stimulus' ist. Einerseits könnte er für den Beurteiler einen Stimulus darstellen, den er noch nie zuvor gesehen hat. Anderseits könnte es sich auch um einen Stimulus handeln, der nur nicht zu seinem Expertisebereich gehört. Die Kategorisierung des Parameters ‚Länge (der Informationssequenz)' ist durch die Anzahl der Items beschränkt, aber die Operationalisierung eines einzelnen ‚Items' wird nicht vorgenommen. Demnach könnte es sich um ein Wort, eine Zahl, einen Satz oder ein sonstiges Item handeln.

Einordnung der Bewertungsaufgabe im Gerätturnen

Die Aufgabe der Bewertung einer Gerätturnübung kann, angelehnt an die Vorhersagebedingungen des Modells (Hogarth & Einhorn, 1992), wie folgt einordnet werden.

Der Bewertungstyp stellt in der ‚normalen' Wertungssituation einen SbS-Prozess dar. Die urteilenden Kampfrichter notieren sich während der Übungspräsentation die Abzüge, die sie für gewisse Fehler vornehmen (2.3). Durch spezielle Instruktionen könnte jedoch ein EoS-Prozess forciert werden, indem den VPn untersagt wird, Notizen zu machen und somit erst am Ende der Übungspräsentation ein Urteil zu fällen (5.1).

Die sehr hohe Komplexität der Urteilsaufgabe eines Kampfrichters wurde in Kapitel 2.4 beschrieben.

Die Einteilung unter dem Aspekt der Serienlänge erweist sich als etwas schwieriger, da wie bereits erläutert, keine eindeutige Definition eines Items vorliegt. Wenn man ein komplettes Turnelement als ein Item ansieht, ergibt sich für den Unparteiischen eine kurze Aufgabe, da eine Übung aus etwa zehn Elementen besteht und damit zwischen zwei und 12 Items lang ist. Wenn man allerdings bedenkt, dass jedes Element aus einer Vielzahl an Informationen zusammengesetzt ist, wie beispielsweise der Arm-Rumpf-Winkel, die Haltung der Arme oder die Körperspannung, ergibt sich eine sehr große Anzahl an Einzelinformationen, die der Kampfrichter zu verarbeiten hat. Daher kann die Wertungsaufgabe des Kampfrichters im Gerätturnen als lange Aufgabe definiert werden.

Die Urteilsfehler-Thematik lässt sich zusammenfassend dahingehend beschreiben, dass deren Entstehung insbesondere auf die Begrenztheit der Informationsverarbeitung zurückzuführen ist. Sehr unterschiedliche und teilweise kaum voneinander abgrenzbare Einflüsse führen dazu, dass die stets angestrebte objektive Beurteilung von Sportlern und deren Leistungen nicht in vollem Ausmaß realisierbar ist.

Der Reihenfolge-Effekt, als einer von vielen Einflüssen, verändert die Urteilsbildung dahingehend, dass die ersten – der Primacy-Effekt – oder entsprechend die letzten – der Recency-Effekt – Informationen in einer Serie einen enormen Stellenwert erhalten. Unterschiedliche Bedingungen, wie etwa die Geschwindigkeit der Informationspräsentation, führen eher zum einen oder zum anderen Effekt. Das ‚belief-adjustment'-Modell (Hogarth & Einhorn, 1992) beschäftigt sich mit dem Phänomen der Reihenfolge und trifft konkrete Vorhersagen darüber, wann es zu welchem Effekt kommt. Das Modell konnte in unterschiedlichen Bereichen erfolgreich getestet werden und soll auch für diese Arbeit die theoretische Grundlage bilden.

4 Forschungsfragen

Ausgehend von den theoretischen Grundlagen und Überlegungen der letzten Kapitel und vom Forschungsbereich der sozialen Kognitionspsychologie ausgehend, sollen nun die angestrebten Fragestellungen für die vorliegende Studie abgeleitet werden.

Die seit Anfang 2006 geltenden Wertungsvorschriften im Gerätturnen sind so ausgelegt, dass die ‚Übungsausführung' gegenüber dem ‚Übungsinhaltswert' extrem hoch in die Wertung eingeht (siehe 2.3). Da die Wertungsurteile des B-Kampfgerichts jedoch die meisten Diskussionen und Kontroversen liefern, kann vermutet werden, dass gerade die B-Note anfälliger für Urteilsverzerrungen und auch für den Einfluss des Reihenfolge-Effekts ist. Aufgrund dessen wird anhand der B-Note als abhängige Variable untersucht, ob ein Reihenfolge-Effekt nachweisbar ist. Ein weiterer eher pragmatischer Aspekt, der für die Verwendung der Ausführungsnote in den geplanten Untersuchungen spricht, bezieht sich auf die Manipulation der Untersuchungsvariablen. Die gewählte Methode des Videoschnitts (5.3) kann besser durchgeführt werden, wenn, aufgrund des geforderten fließenden Übungsablaufs, miteinander verbundene Elemente theoretisch getrennt werden können. Diese Verbindungen werden im Zusammenhang mit dem Übungsinhaltswert bewertet, sind für die Übungsausführung aber irrelevant. Aus diesen Überlegungen heraus soll alleinig über die B-Note bzw. die Abzüge für die Übungsausführung die vorliegende Fragestellung untersucht werden.

Wie bereits im Forschungsproblem geschildert (1) fehlt es bisher an entsprechenden Untersuchungen, deren Studiendesign Aussagen über die Bedeutung der Übungszusammenstellung im Gerätturnen, exemplarisch für andere ästhetische (technisch-kompositorische) Sportarten, zulassen. Unklar ist, ob der Reihenfolge-Effekt, im Zusammenhang schneller Bewegungen und damit geringen Zeitintervallen, innerhalb inkonsistenter Informationen und auch zwischen der letzen Information und dem endgültigen Urteil vorherrscht (3.3). Wirkt sich für die Wertung eines Athleten, ein Fehler zu Beginn der Übung schwerwiegender aus als einer am Ende? Verhält es sich vielleicht umgekehrt oder macht es gar keinen Unterschied, zu welchem Zeitpunkt der Athlet einen Fehler begeht?

Die Aufgabe der Kampfrichter, die Bewertung von Gerätturnübungen, ist trotz der durchgeführten Änderungen in den internationalen

Wertungsvorschriften (CdP) (2.3) eine sehr komplexe (2.4). Für die Untersuchungen, die dieser Arbeit zu Grunde liegen, wird das ‚belief-adjustment'-Modell (Hogarth & Einhorn, 1992) als theoretische Grundlage herangezogen. Das Modell sagt voraus, unter welchen Umständen es zu einem bestimmten Reihenfolge-Effekt kommt (3.4). Einerseits beeinflusst die Komplexität der Aufgabe die Vorhersage, andererseits ist bedeutend, auf welche Weise der kognitive Prozess der Urteilsbildung abläuft. Für komplexe Aufgaben sagt das Modell generell einen Recency-Effekt voraus (3.3). Dabei ist unbedeutend, ob das Urteil anhand eines SbS- oder eines EoS-Prozesses gebildet wird. Üblicherweise macht der Kampfrichter im Gerätturnen Zwischenurteile, die er auf einem Bewertungsbogen[14] festhält. Die Frage, ob sich ein Unterschied in der Bewertung einer Übung ergibt, wenn Kampfrichter nicht die Übungen bzw. die Abzüge mitschreiben dürfen, sondern ihr Urteil direkt aus dem Gedächtnis heraus abgeben müssen, steht zusätzlich zur Hauptfragestellung im Interesse dieser Arbeit.

Der Kampfrichter bewertet im Falle, dass ein Recency-Effekt vorliegt, das Ende einer Übung als wichtiger und ahndet wahrgenommene Fehler härter, als er das bei Fehlern am Anfang der Übung tun würde. Dadurch wirken sich Fehler in der zweiten Hälfte der Übung schwerwiegender aus, als Fehler, die in der ersten Hälfte geturnt werden. Zwar sind die Belege bei komplexen und einfachen Aufgabe zugunsten eines Recency-Effekts recht eindeutig, hingegen für komplexe und lange Aufgaben recht uneindeutig und wenig untersucht. Falls ein Primacy-Effekt in der speziellen Urteilssituation vorliegt, werden Übungen mit dem Fehler vorne, mit mehr Abzügen in der B-Note bestraft, da die ersten präsentierten Informationen stärker in das Urteil eingehen als die zuletzt dargebotenen. Somit bleibt spannend, inwieweit sich die Manipulation der Reihenfolge von schnellen Bewegungen auf die Wertung des Athleten auswirkt.

Weiterhin als sehr interessant für die vorliegende Fragestellung zeigt sich die getroffene Unterscheidung der Turngeräte nach der Bewegungsgeschwindigkeit des Athleten (Plessner, 1999) und damit der unterschiedlichen kognitiven Belastung für die beurteilenden Kampfrichter (2.4). Schnelle Geräte, wie das Pauschenpferd, der Sprung, oder das Reck fordern den Kampfrichter stärker als die vergleichsweise langsamen Geräte, Boden, Ringe und Barren. Vermutlich führen die schnellen

[14] Wettkämpfe im Gerätturnen werden für gewöhnlich mit Hilfe eines Bewertungsbogens beurteilt. Er dient der Protokollierung der gezeigten Elemente, indem die geforderten Elemente, Pluspunkte und auch die Abzüge für Fehler notiert werden.

Geräte eher dazu, anfällig für systematische Kontexteinflüsse wie der der Reihenfolge zu sein.

Daraus ergeben sich für die Studien folgende Fragestellungen:

- Ist die Vorhersage des ‚belief-adjustment'-Modells von Hogarth und Einhorn (1992) richtig? Kommt es wirklich in der entsprechenden Konstellation von Bedingungen zum vorhergesagten Reihenfolge-Effekt? Oder ergibt sich der gegensätzliche Reihenfolge-Effekt? Im konkreten Fall: Kommt es bei einem Fehler am Ende der Übung zu einer geringeren Wertung und damit zu einem Recency-Effekt, da die Kampfrichter eine komplexe Aufgabe zu bewältigen haben? Oder kommt es zu einem Primacy-Effekt und die Übung mit einem Fehler am Ende wird mit einer höheren Wertung belohnt?
- Ist der Reihenfolge-Effekt, so wie es das Modell vorhersagt, unabhängig von der Art der Bewertung? Ist es somit vollkommen unerheblich für den Ausgang der Beurteilung, ob diese ‚Schritt für Schritt' oder auf einmal am Ende der Präsentation der Informationen durchgeführt wird?
- Hat das Gerät, an dem die Übung gezeigt wird einen Effekt auf die Wertung der beurteilenden Kampfrichter? Ist somit die Urteilsverzerrung an einem schnellen Gerät stärker bzw. überhaupt vorhanden, während es am anderen, langsameren Gerät nur zu einem geringen oder zu überhaupt keinem Einfluss kommt?

5 Methode und Material

Um die Fragestellungen der Arbeit zu beantworten, wird die experimentelle Vorgehensweise gewählt. Die beiden Untersuchungen werden nicht im Feld, sondern anhand von Videoaufnahmen durchgeführt, was eine gewisse Einschränkung der externen Validität mit sich bringt (Mascarenhas, Collins & Mortimer, 2002; Plessner & Betsch, 2002; Puhl, 1980). Zur Umsetzung der gewählten Methode wird die Untersuchung zweigeteilt durchgeführt.

Das Erstexperiment dient der Sammlung von Erfahrungen im Zusammenhang mit dem Aufnehmen, Sichten und Erstellen von geeignetem Videodatenmaterial sowie der Organisation und Durchführung der experimentellen Untersuchung. Anhand dieser ersten Untersuchung soll überprüft werden, ob die Ergebnisse Tendenzen aufweisen, die zur Klärung der aufgestellten Fragestellungen (4) beitragen und aufzeigen, ob und welche Verbesserungen im Hinblick auf die zweite Untersuchung vorgenommen werden sollen.

Das Zweitexperiment wird anhand der Erkenntnisse aus dem Erstexperiment optimiert und dient der eigentlichen Erkenntnisgewinnung und Beantwortung der Forschungsfragen. In beiden Untersuchungen haben lizenzierte Kampfrichter die Aufgabe, verschiedene Videosequenzen von Gerätturnübungen zu bewerten, die in einer Präsentation organisiert sind. Die hierzu zusammengestellten Microsoft PowerPoint-Präsentationen, die alle Instruktionen und Videosequenzen enthalten, werden in Kapitel 5.3 vorgestellt. Die Präsentation bildet die erste von drei aufeinander abgestimmten Materialien, die für die Untersuchungen benutzten werden. Die von den Kampfrichtern ermittelten Wertungen für die Übungen werden auf einem sogenannten Bewertungsbogen festgehalten. Außerdem füllen die Kampfrichter einen Personenfragebogen aus, der neben den demographischen Daten, die Expertise auf dem Gebiet der Kampfrichterei durch einen spezifischen Fragenkatalog erhebt. Die ausführliche Beschreibung des Bewertungsbogens und des Personenfragebogens erfolgt in Kapitel 5.4.

5.1 Design

In beiden Experimenten, dem Erst- und dem Zweitexperiment, werden Kampfrichtern - weiterhin als Versuchspersonen (VPn) bezeichnet - manipulierte Videosequenzen gezeigt. Diese zeigen unterschiedliche Turnübungen, die von den VPn gesichtet und bewertet werden sollen.

Durch die beiden unabhängigen Variablen ergibt sich ein 2x2 Untersuchungsdesign, das anhand der Videosequenzen umgesetzt wird.

Die Videosequenzen, sind wie folgt aufgebaut: Die VPn bekommen sowohl Experimental- als auch Kontrollübungen an den Geräten Ringe und Reck zu sehen. Alle Videosequenzen eines Gerätes werden jeweils durch einen schwarzen Bildschirm getrennt, der ein Interstimulus-Intervall darstellt.

Die erste Übung am Gerät ist immer eine *Einwertübung (E)*, die dazu dient, die Kampfrichter auf das gezeigte Niveau der Übungen einzustellen. Durch dieses ‚lead-in' sollen die VPn das Leistungsniveau, die Schnelligkeit, die Perspektive und ähnliche Faktoren der Übungen kennenlernen. Diese Übung wird nicht manipuliert und allen VPn in gleicher Weise vorgeführt. Daraufhin folgt eine *Experimentalübung*, die anhand von zwei unabhängigen Variablen manipuliert wird. Die Kontrollübung (K), als dritte präsentierte Videosequenz, stellt eine nicht manipulierte Übung dar. Abschließend bekommen die Kampfrichter noch einmal eine Experimentalübung zu sehen.

Die zwei unabhängigen Variablen, *Reihenfolge* und *Bewertungstyp*, werden jeweils in zwei Ausprägungen präsentiert. Daraus ergibt sich ein 2x2 Design, das durch *vier Untersuchungsbedingungen* repräsentiert wird. Die VPn werden zufällig einer dieser Untersuchungsbedingung zugeordnet. Der erste Faktor *Reihenfolge* wird variiert, indem pro Experimentalübung die beiden Videoversionen eins und zwei entstehen. Diese unterscheiden sich dahingehend, dass eine Version den Fehler in der ersten Videohälfte (V) zeigt, während in der anderen Version dieser in der zweiten Videohälfte (H) zu sehen ist. Der zweite Faktor *Bewertungstyp* wird durch die zwei Ausprägungen ‚end of sequence' (EoS) und ‚step by step' (SbS) variiert. Bei der EoS-Untersuchungsbedingung sollen die VPn keine Notizen während der Untersuchung machen, während sie in der SbS- Untersuchungsbedingung dazu instruiert werden (5.2).

Durch die *Farben* – blau, gelb, grün und rot – werden die Untersuchungsbedingungen dargestellt. Die Farbgebung dient dazu, den Versuchsleiter in der Untersuchungsdurchführung (6.2) nicht explizit mit den speziellen Inhalten der Untersuchungsbedingung zu konfrontieren. Dadurch soll verhindert werden, dass er sein Verhalten speziell auf eine Untersuchungsbedingung hin verändert und damit eine Verzerrung der Ergebnisse hervorruft (Bortz & Döring, 2003).

Die beiden Untersuchungen gleichen sich strukturell größtenteils. Das Erstexperiment wird entsprechend der obigen Beschreibung aufgebaut. Beispielsweise haben die VPn der ersten Untersuchungsbedingung, blau, die Aufgabe das Video eins zu bewerten und dürfen dabei keinerlei Notizen machen. Das erste Video zeigt am Reck die Einwertübung, dann eine Übung, in der der Fehler vorne zu sehen ist, die Kontrollübung und schließlich eine weitere Experimentalübung mit dem Fehler hinten. Im Anschluss folgt das Gerät Ringe, das ebenfalls durch eine Einwertübung eingeleitet wird. Daraufhin bewerten die Kampfrichter eine Übung mit dem Fehler hinten, dann die Kontrollübung und zum Schluss eine Übung mit einem Fehler in der ersten Videohälfte. In entsprechender Weise sind die anderen Untersuchungsbedingungen, die in Tabelle 3 abgebildet sind, zu lesen.

Tabelle 3: Untersuchungsdesign des Erstexperiments

		Bewertungstyp (Faktor 2)							
		EoS				**SbS**			
Reihenfolge (Faktor 1)	**Video 1**	UB 1 - Blau				UB 2 - Gelb			
	Übung	1	2	3	4	1	2	3	4
	Reck	E	V	K	H	E	V	K	H
	Ringe	E	H	K	V	E	H	K	V
	Video 2	UB 3 - Grün				UB 4 - Rot			
	Übung	1	2	3	4	1	2	3	4
	Reck	E	H	K	V	E	H	K	V
	Ringe	E	V	K	H	E	V	K	H

Das Zeitexperiment ist beinahe identisch mit dem geschilderten Aufbau des Erstexperiments strukturiert. Zu den vier Versuchsbedingungen werden die VPn auch in dieser Studie zufällig zugeordnet. Das Design des Zweitexperiments unterscheidet sich von dem des Erstexperiments durch drei Punkte:

Erstens, werden pro Gerät sechs, statt vier Übungen gezeigt. Durch einen schwarzen Bildschirm getrennt, werden die unterschiedlichen Videosequenzen – Einwertübung (E), zwei Experimentalübungen, Kontrollübung (K), zwei Experimentalübungen – den Kampfrichtern zum Werten vorgelegt. Dieser Aufbau wird am zweiten Gerät Reck beibehalten.

Zweitens, wird die Reihenfolge der beiden gezeigten Geräte entsprechend olympischer Reihenfolge – Boden, Pauschenpferd, Ringe, Sprung, Barren und Reck – organisiert. Damit werden den VPn zunächst die Übungen an den Ringen und dann am Reck vorgeführt.

Und drittens besteht die Videoversion eins, anders als im Erstexperiment, nur aus Experimentalübungen, die beispielsweise den Fehler vorne (V) zeigen. Video zwei enthält nur Experimentalübungen mit dem Fehler hinten (H).

Somit ergibt sich für das Zweitexperiment ein übersichtliches Design (Tabelle 4). Die VPn, die beispielsweise zufällig zur vierten Untersuchungsbedingung (rot) zugeordnet sind, haben die Aufgabe, anhand des zweiten Videos jeweils sechs Übungen an den Ringen und sechs Übungen am Reck zu bewerten. Dabei werden sie aufgefordert, alle Abzüge zu notieren, die sie in den Übungen sehen. An den Ringen, und auch am Reck, sehen sie zunächst die Einwertübung (E), zwei Übungen mit dem Fehler hinten (H), die Kontrollübung (K) und erneut zwei Übungen mit dem Fehler hinten.

Tabelle 4: Untersuchungsdesign des Zweitexperiments

		Bewertungstyp (Faktor 2)											
		EoS						**SbS**					
Reihenfolge (Faktor 1)	**Video 1**	UB 1 - Blau						UB 2 – Gelb					
	Übung	1	2	3	4	5	6	1	2	3	4	5	6
	Reck	E	V	V	K	V	V	E	V	V	K	V	V
	Ringe	E	V	V	K	V	V	E	V	V	K	V	V
	Video 2	UB 3 - Grün						UB 4 – Rot					
	Übung	1	2	3	4	5	6	1	2	3	4	5	6
	Reck	E	H	H	K	H	H	E	H	H	K	H	H
	Ringe	E	H	H	K	H	H	E	H	H	K	H	H

5.2 Untersuchungsvariablen

Um die formulierten Forschungsfragen (4) zu untersuchen, wird ein Untersuchungsdesign mit zwei unabhängigen Variablen und einer abhängigen Variable organisiert (5.1). Dabei sehen und bewerten die VPn unterschiedliche Übungen an den beiden Turngeräten Ringe und Reck.

5.2.1 Die Geräteauswahl

In der Literatur führt die Anzahl an Elementen, die innerhalb einer Sekunde bewertet werden müssen, zu einer Unterscheidung zwischen schnellen und langsamen Geräten. Die Geräte Pauschenpferd, Sprung und Reck zählen zu den schnellen Geräten, während die Geräte Boden, Ringe und Barren den langsamen Geräten zugeordnet werden (Plessner, 1997, 1999). Bei den schnellen Geräten bieten sich den Wertungsrichtern in der Regel weniger als zwei Sekunden Zeit, um ein gesamtes Element in all seinen Aspekten zu erfassen und zu bewerten. So liegt die Vermutung nahe, dass das Wertungsurteil bei den schnellen Geräten besonders anfällig für verzerrende Einflüsse ist (Plessner, 1997, 1999; Wilson, 1976a) (2.1).

So hat Wilson (1976a) untersucht, wie objektiv die Wertungen im Gerätturnen sind und wie valide und reliabel der Gebrauch von Videoaufnahmen ist. Ein sehr hoher Grad an Objektivität wurde im Frauenturnen am Gerät Stufenbarren, allerdings nur moderate Objektivität am Gerät Sprung ermittelt. Bezüglich der Validität zwischen den Wettkampf-Wertungen und denen der Video-Wertungen und der Reliabilität der Video-Wertungen, wurden sehr gute Ergebnisse am Stufenbarren, jedoch keine signifikanten Ergebnisse am Sprung festgestellt.

Angelehnt an diese Unterscheidung, soll überprüft werden, ob sich die genannten Einflüsse auch unter den geplanten Untersuchungsbedingungen im männlichen Gerätturnen zeigen. Die Geräte werden im Folgenden (Abbildung 3) graphisch als Skala von langsam nach schnell dargestellt:

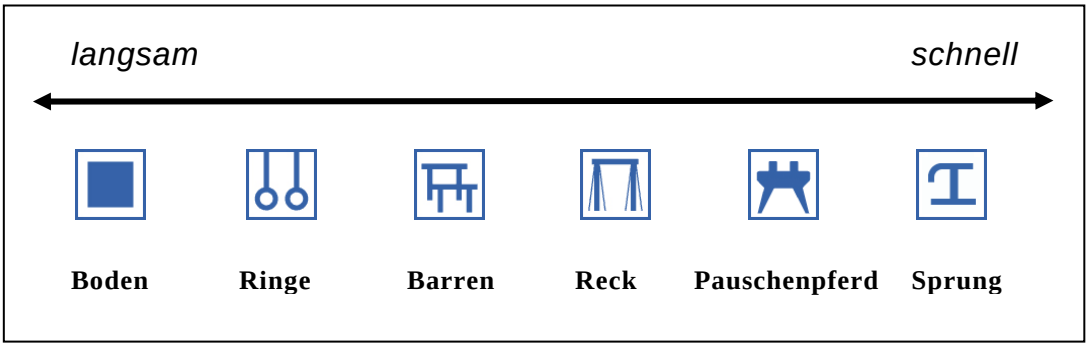


Abbildung 3: Geräte im männlichen Gerätturnen geordnet von langsam nach schnell (eigene Darstellung)

Das Gerät *Sprung* hat im Gerätturnen eine Sonderstellung. Jeder Turner absolviert beim Turnen dieses Gerätes nur ein einziges Element. Da sich dieser Sachverhalt für die Untersuchung von sequenziellen Effekten innerhalb von Turnübungen nicht eignet, wird im Weiteren diesem Gerät keine Beachtung mehr geschenkt.

In den Untersuchungen dieser Arbeit wird die Differenzierung in schnelle und langsame Geräte (Plessner, 1997, 1999) als Unterscheidungsmerkmal herangezogen. Dabei werden in den durchgeführten Untersuchungen die Geräte Ringe, als relativ langsames Gerät, und Reck, als relativ schnelles Gerät, ausgewählt. Die Begründung für diese Auswahl liegt in den speziellen Bedingungen, die eine Videoaufnahme erfüllen muss, um der beabsichtigten Manipulation der Videosequenzen mithilfe des Videoschnitts (5.3) zu genügen. Dabei ergeben sich die folgenden Vorraussetzungen:

1. Eine Kameraperspektive, die komplett identisch mit der der Kampfrichter ist. Diese erste Voraussetzung steht mit den beiden folgenden Punkten (2 und 3) im Widerspruch, weswegen sie nicht in optimaler Weise verwirklicht werden kann und angepasst werden muss. Es wird eine Kameraperspektive gewählt, die der des Kampfrichters recht ähnlich ist.
2. Keine Bewegung der Kamera, da ansonsten ein ständig wechselnder Hintergrund den möglichst unbemerkten Videoschnitt verhindert.
3. Eine maximal große Aufnahme des Athleten am Gerät während der kompletten Übung mit An- und Abgang, damit die gezeigten Bewegungen vom Kampfrichter gut erkannt werden.
4. Eine wiederkehrende Position des Körpers zum Gerät ist ausschlaggebend für einen möglichst verdeckten Videoschnitt. Man stelle sich vor, eine Person bewegt sich auf einer 5 Meter langen Strecke vom Ausgangspunkt A zum Endpunkt B, dreht sich um 180° um die Körperlängsachse und geht wieder zurück. Die Videoaufnahme dieser Bewegung, die keine Wiederholung einer Körperposition im Raum beinhaltet, ist mit der in Kapitel 5.3.2 beschriebenen und angewendeten Art der Videobearbeitung nicht geeignet.

Die Geräte Boden, Pauschenpferd und Barren können aufgrund der geschilderten Bedingungen nicht verwendet werden. Wobei hier vor allem die Bedingungen drei und vier ausschlaggebend sind.

5.2.2 Unabhängige Variablen

Wie in Kapitel 5.1 bereits kurz beschrieben, werden zwei unabhängige Variablen zur Prüfung des Reihenfolge-Effekts angenommen. Die Reihenfolge der präsentierten Informationen, im speziellen Fall der Turnelemente und die Art der Bewertung, die der Kampfrichter nutzt um sein Urteil über die gesehenen Videosequenzen zu bilden, stellen die unabhängigen Variablen dar. Diese beiden Faktoren werden nun ausführlicher beleuchtet.

1. Reihenfolge

Kapitel 5.1 schildert die Unterscheidung von Einwert- bzw. Kontrollübungen und Experimentalübungen. Eine Manipulation wird ausschließlich bei Experimentalübungen durchgeführt und zeigt sich anhand des Auftretens eines Fehlers zu Beginn bzw. am Ende der Übung. Aus einer ursprünglichen Originalübung werden zwei Experimentalübungen erstellt, die sich lediglich durch die Reihenfolge der Turnelemente, und damit dem Zeitpunkt, in dem ein Fehler im Übungsverlauf gemacht wird, unterscheiden. Die Videosequenzen werden per Videoschnitt so verändert, dass der identische Fehler nicht mehr vorne (bzw. hinten) in der Übung, sondern in der zweiten Hälfte (ersten Hälfte) zu sehen ist (5.3). In zwei von vier Untersuchungsbedingungen sehen VPn eine manipulierte Übung, in der ein größerer Fehler in der ersten Hälfte[15] der Übung gezeigt wird, und die in der zweiten Hälfte keinen entsprechenden Fehler zeigt. Diese Videosequenz hat demzufolge die Charakteristik ‚negativ-positiv' bzw. ‚Fehler vorne'. Die Videovariante ‚positiv-negativ' bzw. ‚Fehler hinten' zeigt die umgekehrte Reihenfolge und wird von den VPn der anderen beiden Untersuchungsbedingungen eingeschätzt.

2. Bewertungstyp

Die zweite Unterscheidung beruht auf der Formulierung der Aufgabe, die den VPn mittels Instruktion in einer Microsoft PowerPoint-Präsentation mitgeteilt wird (5.3.3). Der Bewertungstyp wird in zwei Ausprägungen untersucht:

Bewertungstyp A: Abzüge nicht mitschreiben (end of sequence - EoS)

Die VPn sollen das Videomaterial aufgrund ihres Gesamteindrucks, ohne das parallele Mitschreiben der Abzüge, am Ende der Übung in Form

[15] Die Mitte der Übung entspricht der Mitte der Zeitspanne vom Beginn der ersten Bewegung aus dem ruhigen Hang, bis zum Ende des Abgangs zum ruhigen Stand.

einer durch Punkte repräsentierten Wertung abgeben. Das Urteil wird auf Grundlage der aktuellen Wertungsvorschriften des Internationalen Turnerbundes gefällt. Diese Aufgabe besteht für die Hälfte der VPn, im konkreten Fall die Untersuchungsbedingung eins und drei. Dieser Bewertungstyp, ohne das Notieren von Abzügen, wird allerdings in den Kampfrichter-Ausbildungen nicht explizit gelehrt. Geraten wird ihnen, die Übungen lückenlos zu dokumentieren, eine Prüfung der Dokumentation findet allerdings nicht statt. Wie der B-Kampfrichter seine Note erstellt, bleibt somit komplett ihm überlassen. Vermutet werden kann, dass es Kampfrichter gibt, die sich wenige oder keine Notizen in Wettkämpfen machen. Für die meisten Kampfrichter allerdings dürfte die Anweisung, keinerlei Abzüge zu notieren, eine ungewohnte sein. Allerdings soll durch die Anweisung, keine Notizen zu machen, die im Modell von Hogarth und Einhorn (1992) geforderte Art der Bewertung EoS, also die Bildung des Urteils am Ende der gesamten Präsentation aller Informationen – ohne Zwischenurteil, forciert werden (3.4).

Bewertungstyp B: Abzüge mitschreiben (step by step - SbS)

Die VPn sollen die gezeigten Übungen wie im Gerätturnen gelehrt, anhand von Bewertungs- bzw. Protokollbögen bewerten. Dieses Urteil beruht ebenfalls auf den verbindlichen Wertungsvorschriften des Internationalen Turnerbundes. Die Hälfte der VPn, Untersuchungsbedingung zwei und vier, werden instruiert, diese Form der Bewertung zu verwenden. Ein Großteil der Kampfrichter wendet diese Art der Bewertung im Wettkampf an. Die Aufgabe stellt sich als eine bekannte und gewohnte dar. Aus diesem Grund kann die Aussagekraft der Ergebnisse und die Übertragbarkeit auf die Realsituation (Wertungssituation von Kampfrichtern in realen Gerätturn-Wettkämpfen) als besonders hoch eingestuft werden. Darüber hinaus, kann es als ein natürliches Inventar zur Messung der kognitiv ablaufenden Prozesse betrachtet werden (Taylor & Fiske, 1981). Diese Art der Bewertung soll eine Realisierung der Bewertungsform ‚mit Zwischenurteil' (Hogarth & Einhorn, 1992) darstellen (siehe 3.4).

Einwert- und Kontrollübungen werden nicht manipuliert. Sie werden lediglich von den VPn bewertet, um sicherzustellen, dass sich die Untersuchungsdaten der vier Versuchsgruppen nicht grundsätzlich unterscheiden, sondern nur aufgrund der unterschiedlichen Untersuchungsbedingungen, denen sie zufällig zugeordnet werden. Der ermittelte Unterschied der Wertungen der Kontroll- und der Experimentalübungen bzw. der Unterschiede der Experimentalübungen in den vier

Untersuchungsbedingungen wird als Folge der manipulierten Bedingungen und daher als Wirkung der unabhängigen Variablen angesehen.

5.2.3 Abhängige Variable

Die abhängige Variable stellt die Bewertung der B-Note einer Übung dar. Im Gerätturnen, aber auch in anderen technisch-kompositorischen Sportarten, wird die Bewertung der gezeigten Übungen in Form einer aus zwei Teilnoten gebildete Endnote getroffen. Diese beruht auf den verbindlichen Wertungsvorschriften des Internationalen Turnerbundes (FIG, 2006) und setzt sich aus der Schwierigkeit (A-Note) und der Ausführung (B-Note) einer Übung zusammen.

Die B-Note drückt die Höhe der technischen, haltungsmäßigen und kompositorischen Fehler in der Ausführung aus. Jegliche Abweichungen von der Idealausführung werden anhand festgelegter Kriterien geahndet und von der Maximalnote 10,0 Punkte in Abzug gebracht. Je weniger Fehler in einer Übung enthalten sind, desto geringer sind auch die vorgenommenen Abzüge. Dadurch ergibt sich, durch die Subtraktion der anfallenden Fehler vom Ausgangswert, eine höhere B-Note. Bei Übungen mit hohem Leistungsniveau[16] stellt die Höchstnote 10,0 Punkte diesen Ausgangswert dar. Die A-Note ist im Fall des Rope Skippings (Boen et al., 2006), aber auch in anderen ästhetischen Sportarten, wie dem Gerätturnen, mit einem geringeren Maß an Subjektivität behaftet (Ste-Marie & Lee, 1991) (2.3). Daraus folgernd wirken sich psychologische Effekte stärker auf die B-Note aus.

Da es sich im betrachteten Reihenfolge-Effekt um einen eher geringen bis mittleren Effekt handelt, scheint es sinnvoll, diesen anhand der B-Note zu untersuchen. Im Falle, dass er im speziellen Kontext überhaupt vorhanden ist, sollte er besser aufzudecken sein, als das mit der A-Note möglich wäre. Um mögliche Urteilsverzerrungen aufzudecken, sprechen sich auch Boen et al. (2006) für die Verwendung der B-Note aus. Für die Untersuchungen dieser Arbeit ist somit nur die B-Note von Bedeutung.

5.2.4 Störvariablen und der Einfluss anderer Effekte

Störvariablen

Die Durchführung von experimentellen Untersuchungen unterliegt einer Vielzahl von Störvariablen. Diese wirken sich manipulierend auf die erhobenen Daten und auf die interne Validität der Untersuchung aus.

[16] Übungen, die allen Anforderungen genügen, die im CdP gefordert werden (2.3).

Kontrolltechniken sollen zur Erhöhung der internen Validität beitragen. Personengebundene und untersuchungsbedingte Störvariablen lassen sich unterscheiden (Bortz & Döring, 2003, S. 525).

Unter *personengebundene Störvariablen* werden Unterschiede bei den Versuchsteilnehmern verstanden, die nicht auf eine unabhängige, sondern auf die abhängige Variable zurückzuführen sind. Die wichtigste Technik, Störvariablen möglichst gering zu halten, ist die Randomisierung, in der eine zufällige Zuweisung der Untersuchungsteilnehmer zu den Untersuchungsbedingungen durchgeführt wird. „Auf diese Weise werden auch Störvariablen neutralisiert, die man im Vorfeld gar nicht benennen könnte“ (ebenda, S. 526). Die Empfehlung der Autoren lautet, dass pro Untersuchungsbedingung mindestens 20 VPn akquiriert werden sollen. Auf Basis dieser Grundaussagen wird in den Untersuchungen eine Randomisierung durchgeführt.

Untersuchungsbedingte Störvariablen bilden eine weitere Ursache für mangelnde interne Validität (Bortz & Döring, 2003, S. 528). Kontrolltechniken sollen wiederum sicherstellen, dass die äußeren Rahmenbedingungen der Untersuchungsdurchführung für alle Stichproben identisch sind. Alle, nicht auf die unabhängigen Variablen zurückgehenden Unterschiede werden kontrolliert oder durch Ausschalten, Konstanthalten oder Registrieren (anhand eines Untersuchungsprotokolls) neutralisiert. Dadurch soll ein störungsfreier Ablauf der Untersuchung ermöglicht werden. Außerdem sind Unterschiede in den Versuchsdurchführungen zu vermeiden, wie beispielsweise bei den Räumlichkeiten, in denen die Untersuchung stattfindet oder die Instruktionen, die standardisiert ablaufen sollen. Falls darüber hinaus Störungen auftreten, werden diese nach Art und Intensität protokolliert, wie das beispielsweise bei störenden Geräuschen, unerwarteten Zwischenfragen, Instruktionsfehlern oder technischen Problemen der Fall ist. Die Versuchssituation wird durch den annähernd identischen Aufbau der Untersuchungsbedingungen konstant gehalten und soll dadurch zu einer Maximierung der internen Validität führen. Das jeweilige komplette Experiment wird zu diesem Zweck unter weitgehend standardisierten Bedingungen durchgeführt. Ein separater Raum, in dem die VPn an je einem Laptop das Experiment durchführen, bietet die nötige Standardisierung. Sämtliche Instruktionen werden auf dem Laptop gezeigt und die Darbietung der Untersuchung ist für alle VPn identisch. Die Versuchsleiterin ist in allen Untersuchungen dieselbe und hat vor und während der Untersuchung einen so geringen Kontakt mit den VPn wie nur möglich, um die Motivierung der Subjekte möglichst identisch zu halten.

Durch die Messungen selbst kann es ebenfalls zu störenden Einflüssen kommen. Dazu kann beispielsweise die *Sensibilisierung*[17] oder die Ermüdung der VPn gerechnet werden. Der Sensibilisierung kann entgegengewirkt werden, indem den VPn vorher nicht gesagt wird, welchen Zweck das Experiment verfolgt. Die sogenannte Coverstory verhilft dazu, dass die VPn unvoreingenommen die Untersuchung durchführen. Die *Ermüdung* der VPn lässt sich nicht gänzlich durch den Versuchsleiter kontrollieren, da er keinen Überblick über den Wachheitszustand zum Teilnahmezeitpunkt hat. Alleine durch die Beachtung der Länge der Untersuchung kann darauf geachtet werden, dass die VPn nicht zu sehr durch die Tätigkeit im Experiment ermüden.

Der Einfluss anderer Effekte

Um zu verhindern, dass ein anderer Effekt die Messdaten verzerrt, werden mögliche beeinflussende Faktoren so gering wie möglich gehalten oder der Versuch unternommen, sie ganz auszuschalten. Die in Kapitel 3.2 erläuterten Einflussgrößen auf Kampfrichterurteile könnten sich konfundierend auf die Studien dieser Arbeit auswirken. Im Folgenden werden Maßnahmen beschrieben, die diese Einflüsse kontrollieren sollen. Einflüsse, die nicht verhindert werden können, werden konstant gehalten. Die Fehlerquellen werden nach den Informationsverarbeitungsstufen angeordnet.

1. Wahrnehmung

Um *Wahrnehmungsprobleme* zu kontrollieren, wird in den durchgeführten Experimenten eine Videoperspektive gewählt, die zum einen der Kampfrichterperspektive sehr ähnlich ist. Zum anderen wird nur diese Sicht auf die Übungen für alle Kampfrichter identisch präsentiert. Dadurch können sich Urteilsfehler, die aufgrund unterschiedlicher Blickperspektiven entstehen, nicht auftreten. Eine wiederholte oder verlangsamte Wiedergabe wird nicht eingeräumt und ist damit identisch mit der Realbedingung im Wettkampf.

2. Kategorisierung

Der *Hof-Effekt* wird kontrolliert, indem jeder Athlet jeweils nur eine Übung turnt und sich somit charakteristische Eigenschaften nicht auf

[17] Unter Sensibilisierung versteht man die verstärkte Reaktion einer VPn nach der Darbietung eines bekannten Reizes.

ein anderes Gerät übertragen können. Weitere Auswirkungen dieses Einflusses können nicht direkt kontrolliert werden.

Um einer *Stereotypisierung* vorzubeugen, werden keinerlei Angaben über ethnische Zugehörigkeit, Nation oder andere Merkmale der Video-Personen an die VPn weitergegeben. Das Geschlecht beeinflusst nicht die Untersuchungen, da nur Männer gezeigt werden, und auch der Körperbau ist bei allen Turnern sehr ähnlich. Mögliche Annahmen der VPn, die sich aufgrund der äußeren Gegebenheiten der Turner ergeben könnten, werden dadurch auf ein Minimum reduziert.

Bezüglich der *Bekleidung* der Video-Personen müssen die beiden Untersuchungen separat betrachtet werden. Die Anzüge des Erstexperiments variieren in ihren Farben, wodurch die Nation zu erahnen ist, wobei die Nationalitätsaufnehmer aber nicht erkennbar sind. Das deutsche Trikot ist vermutlich am bekanntesten. Da die unterschiedliche Bekleidung im Erstexperiment leider nicht vermieden werden konnte, werden zwei Athleten mit dem Deutschland-Trikot eingesetzt, einmal als Kontroll- und einmal als Experimentalvideo. Dadurch kann ein Nationalitäts-Effekt, wenn er denn auftritt, erkannt werden, da er sich auch in der Kontrollübung zeigen müsste. Im Zweitexperiment scheidet dieser Einflussfaktor aus, da die Athleten durch die Farbe der Kleidung, die sie tragen, nicht eindeutig einem Land zugeordnet werden können. Alle Athleten sind identisch gekleidet und tragen keinen Nationalitätsaufnehmer (5.3).

Ebenso wissen die VPn nichts über die potentiellen Leistungen der Turner oder andere Sachverhalte, die ihre Urteile beeinflussen könnten. Die Turner sind ihnen mit hoher Wahrscheinlichkeit nicht bekannt, so dass *erwartungsbedingte Verzerrungen* reduziert werden können. Im Erstexperiment werden Junioren-Athleten gezeigt, die im Allgemeinen nicht so bekannt sind. Im Zweitexperiment sind es erwachsene ausländische Athleten. Durch diese Maßnahmen ist auch der Einfluss der *Reputation* reduziert. Die Videos enthalten keinerlei Informationen, die Auskunft über die Athleten beinhalten. So gibt es keine Einblendung auf Anzeigetafeln oder sonstige Bekanntmachung des Namens oder der Nation (5.3).

Die gezeigten Videosequenzen stellen auch keinen Mannschaftswettkampf dar, so dass die VPn weniger durch den *Positions-Effekt* oder den *Reihungs-Effekt* beeinflusst werden. Die Anordnung der Turner gleicht eher einem Geräte-Finale, in dem die Start-Reihenfolge vorher aufgrund der Ergebnisse der Qualifikation von der Organisation bzw. des Fachverbandes bestimmt wird. Außerdem turnen relativ wenige

Athleten, so dass durch die Untersuchung selbst keine ermüdungsbedingten Einschränkungen entstehen.

3. Gedächtnisprozesse

Um den *Erinnerungs-Effekt* auszuschalten und somit zu verhindern, dass bereits beobachtete oder bewertete Leistungen eines Athleten das Kampfrichterurteil beeinflussen, werden unbekanntere Athleten präsentiert, die jeweils nur einmal im gesamten Experiment bewertet werden.

4. Urteils- und Entscheidungsprozesse

Die Geräuschkulisse bzw. die Interaktion der Wertungsrichter mit den *Zuschauern* kann das Urteil der Kampfrichter beeinflussen. Kontrolliert wird diese Variable durch den kompletten Verzicht auf Akustik während der Präsentation der Übungen.

Der *Heimvorteil* stellt keine Beeinflussung der Messdaten dar. Es sind weder Zuschauer zu hören, noch zu sehen und es gibt keine Mannschaften oder Athleten, die sich derartig unterscheiden, dass man den Eindruck haben könnte, es handle sich um einen Wettkampf vor heimischen Kulissen.

Die *Nationalität* der Turner kann, wie bereits erläutert (vgl. ‚Bekleidung'), anhand der unterschiedlichen Turnanzüge im Erstexperiment von den VPn erahnt werden. Im Zweitexperiment besteht diese Möglichkeit nicht. Die Untersuchungen erhalten keinerlei Informationen über die Nationalität der gezeigten Athleten.

Das *Hot-hand-Phänomen* ist in den durchgeführten Studien nicht relevant, da keine Wettkampfatmosphäre im eigentlichen Sinne entsteht, in der eine Spannung von Übung zu Übung aufgebaut wird. Durch den nicht thematisierten Mannschaftswettkampf stellt sich dieser Effekt als unbedeutend dar.

Dem Einfluss *anderer Kampfrichter* wird entgegengewirkt, indem jede VPn die gestellte Aufgabe alleine bearbeiten muss und jegliche Kommunikation mit Kollegen untersagt wird. In den gezeigten Videos sind auch keine Wertungen anderer Kampfrichter enthalten, die die Bewertung der VPn beeinflussen könnten.

Um den *Sequenz-Effek*t und damit Vergleichsprozesse mit vorherigen Turnern zu verhindern, könnte man die Videosequenzen in unterschiedlicher Reihenfolge präsentieren um so diesen Effekt herauszumitteln. Diese Methode wird allerdings nicht angewendet, um das Design schlicht zu halten. Die *Tendenz zur Einseitigkeit* sowie die *zentrale*

Tendenz stellen Einflüsse dar, die kaum zu verhindern sind. Extreme Urteile werden in der Datenaufbereitung berücksichtigt (7.3).

5.3 Videomaterial und Präsentation

Der Einsatz von Videomaterial in Untersuchungen im Sportbereich ist eine viel genutzte und bevorzugte Methode (Frank & Gilovich, 1988; Scheer & Ansorge, 1979; Ste-Marie & Valiquette, 1996). Die Manipulation von Videomaterial ist in diesem Forschungsbereich ein gängiges Vorgehen, das Vorteile aufweist. Viele Kampfrichter können ein und dieselbe Szene sichten und bewerten, wodurch direkte Vergleiche möglich werden. Darüber hinaus können einzelne Merkmale systematisch konstant gehalten bzw. variiert werden (Plessner & Raab, 1999). Angemerkt werden muss hierbei, dass Untersuchungen, die die Real-Situation mit Videoaufnahmen vergleichen, eine größere Übereinstimmung zwischen den Kampfrichtern ermittelt haben (Puhl, 1980; Taha et al., 1991; Wilson, 1976a & b). Für den Kampfrichter beinhaltet die Wertungssituation im Labor weniger Druck. An sich sollten derartige Untersuchungen in der Realsituation durchgeführt werden, was sich bei der vorliegenden Fragestellung als eher schwierig darstellt.

Als Untersuchungsmaterial dienen Videosequenzen verschiedener männlicher Gerätturner an den Geräten Ringe und Reck. Die folgenden drei Kapitel beschreiben das Zustandekommen des Videomaterials und die Art der Präsentation.

5.3.1 Auswahl der Video-Personen

Die Sportler, die in den Videos des Erstexperiment und Zweitexperiment zu sehen sind, müssen bestimmte Eigenschaften haben. Im Folgenden werden drei Bedingungen formuliert, die bei der Videoauswahl bzw. -erstellung beachtet werden. Die Auswahl der Sportler bzw. der gezeigten Videosequenzen erfolgt aufgrund der ersten beiden Anforderungen.

1. *Kein Wiedererkennungseffekt*

Die Turner sollen den teilnehmenden Wertungsrichtern nicht bekannt sein, um keine Erwartungen bezüglich der bevorstehenden Leistung zu wecken (vgl. Erwartungsbedingte Verzerrungen, Erinnerungs-Effekt, Gruppe und Nationalität & Reputation in 3.2 & 5.2.4).

Für das *Erstexperiment* werden Videosequenzen von der Europameisterschaft im griechischen Volos im Jahre 2006 verwendet. Sinnvoll scheint die Verwendung einer Wettkampfsituation als Videoquelle, da

man hier von einer gewissen Ernsthaftigkeit und der definierten Anwendung des Reglements ausgehen kann. Da es sich um Aufnahmen handelt, die zum Zeitpunkt der Untersuchung (6.1) bekannt sein könnten, werden Junioren-Turner ausgewählt. Ein wesentlicher Grund für die Wahl der Junioren ist ihr geringer Bekanntheitsgrad außerhalb der eignen Nation, wodurch verhindert werden soll, dass die VPn ihre Wertungen aufgrund des Vorwissens über die Leistung oder aufgrund der Reputation vergeben (auch Vanden Auweele et al, 2004). Im *Zweitexperiment* werden auf der Grundlage dieses Kriteriums Videos von Senioren-Turnern verwendet, die im Ausland aktiv sind. Die Videos wurden bei den Rumänischen Meisterschaften im Juli 2007 in Ploiesti aufgenommen. Auch hier wird die Wettkampfsituation aus den genannten Gründen als Videoquelle benutzt. In den Experimentalübungen sind Turner zu sehen, die entweder einen geringen Bekanntheitsgrad haben, unauffällig (Hof-Effekt in 3.2 & 5.2.4) sind oder in den Aufnahmen schlecht zu erkennen sind. Zur Kontrolle des Kriteriums, ob die Athleten den Kampfrichtern bekannt sind, wird dies über den Personenfragebogen (5.4) erhoben.

2. *Leistungsniveau*

Ein weiteres Augenmerk liegt auf dem Leistungsniveau, auf dem sich die Sportler bewegen. Um eine abwechslungsreiche Übung, angelehnt an die Forderungen der aktuellen Wertungsvorschriften, zeigen zu können, müssen die Athleten in der Lage sein, gewisse Schwierigkeiten an den Geräten zu turnen. Zudem sind Übungen auf höherem Leistungsniveau bezüglich der hier geprüften B-Abzüge meist klarer zu bewerten als solche, in denen sich die Fehler stark häufen. Daraus resultiert, dass die Überlagerung von Fehlern in hochklassigen Übungen geringer ist. Folglich fällt es dem Kampfrichter leichter, die einzelnen Fehler wahrzunehmen und so einen guten Überblick über die Übungsausführung zu erhalten (2.3).

Zur Sicherstellung des Leistungsniveaus wird daher für das *Erstexperiment* Videomaterial der Europameisterschaft und nicht eines weniger hochwertigeren Wettkampfs verwendet. Die Junioren-Athleten sind gut auf diesen für sie wichtigsten Wettkampf der Saison[18] vorbereitet und zeigen hochklassige Übungen, was sich positiv auf die Bewertung der B-Note auswirkt (2.3). Das *Zweitexperiment* zeigt Senioren-Turner auf hohem und nicht zu stark voneinander abweichendem Leistungsniveau.

[18] Für Junioren stellt die Europameisterschaft, die jedes zweite Jahr stattfindet, den Saisonhöhepunkt dar. Weltmeisterschaften werden nur für Senioren organisiert.

3. *Kleidung*

Durch die Kleidung können Rückschlüsse auf die Nationalität der Turner gezogen werden, die zur Verzerrung der Wertung führen. Die Turner sollten daher bei den Videoaufnahmen dieselbe Kleidung, ohne spezielle Nationalitäts-Kennung oder sonstige Merkmale tragen (vgl. Trikotfarbe in 3.2 & 5.2.4).

Im Rahmen des *Erstexperiments* konnte diesem Kriterium keine Rechnung getragen werden, da bereits bestehende Aufnahmen einer internationalen Meisterschaft verwendet werden. Für das *Zweitexperiment* führt dieses Kriterium dazu, dass die Videos selbst aufgezeichnet wurden.

5.3.2 Erstellung der Videoversionen

Für die Untersuchung des Reihenfolge-Effekts werden in den beiden Studien unterschiedliche Videosequenzen Kampfrichtern zum Bewerten gegeben. Der grundsätzliche Aufbau einer Untersuchung sieht vor, dass die VPn sowohl Kontroll- als auch Experimentalübungen sichten (5.1).

Als *Einwert- und Kontrollübung* bekommen alle VPn unveränderte Übungen zur Bewertung vorgelegt. Die Videoerstellung für die *Experimentalvideos* lässt sich jeweils in vier identische Stufen, Rekrutierung und Aufnahme, Auswahl und Vortest, Videoschnitt und Videonachbearbeitung einteilen. Diese werden im Folgenden separat beleuchtet.

1. *Rekrutierung und Aufnahme*

Zunächst werden die Video-VPn ausgewählt. Im Erstexperiment werden bereits aufgenommene Videosequenzen verwendet. Im Zweitexperiment folgt nach der Kontaktaufnahme mit den jeweiligen Organisatoren und Trainern eine kurze Erläuterung des Studienvorhabens, bevor die benötigten Bedingungen für die Untersuchung genau vereinbart werden. Die für die geplanten Aufnahmen wichtigen Kriterien, wie beispielsweise die Filmperspektive oder die Nähe zum Gerät wird im Vorfeld besprochen. Außerdem wird die anonyme Behandlung der Aufnahmen, nur für den Zweck der experimentellen Untersuchung versichert. Diese Organisationsschritte werden im Rahmen der *Rekrutierung* durchgeführt. Die ausgesuchten Athleten turnen an zwei unterschiedlichen Geräten (Ringe und Reck) ihre Übung. Die Übungen werden mit einer Videokamera mit Stativ aufgezeichnet und stellen die Phase der *Aufnahme* dar. Durch die Aufnahme möglichst vieler Sportler soll

gewährleistet werden, dass ausreichend Datenmaterial zur Experimentalvideoerstellung vorhanden ist.

2. *Auswahl und Vortest*

Das Datenmaterial wird auf verschiedene Kriterien hin gesichtet und einer Vorauswahl unterzogen. Diese Kriterien sind beispielsweise die Qualität der Videosequenzen (bspw. Abstand zum Gerät, Winkel zum Gerät – im Fall des Erstexperiments), Auftreten von ungewollten Vorkommnissen, Bildqualität, Perspektive und sonstige Besonderheiten (*Auswahl*). Die Erstellung der verschiedenen Videos ist ein aufwendiges Verfahren und bedarf einer vorherigen Sichtung (*Vortest*). In diesem Prozess bewerten Wertungsrichter die aufgenommenen Szenen, um geeignete Übungen für den folgenden Schritt des Videoschnitts zu ermitteln. Dabei ist den Wertungsrichtern gestattet, die Videosequenzen mehrmals zu betrachten, um den Grad der Reliabilität zu erhöhen (Ste-Marie & Lee, 1991) (2.1). Eine Zeitlupenanalyse hingegen wurde nicht gewährt, da die Möglichkeit besteht, dass eine derartig zeitliche Verzögerung die Wahrnehmung und damit die Bewertung beeinflusst (Neumaier, 1988, S. 230). Die Sichtung orientiert sich dabei an den folgenden Kriterien:

- Eine *Perspektive*, die dem des Kampfrichters sehr ähnlich ist. Eine mit der Originalperspektive des Oberkampfrichters des entsprechenden Gerätes identische Aufnahme ist nicht möglich, da ansonsten das Aufnahmegerät zu nah am Gerät stehen muss. Hierbei kann der Turner nur durch eine Kamerabewegung eingefangen werden. Diese in Fernsehübertragungen übliche Technik macht jedoch einen Videoschnitt unmöglich. Die Aufstellung der Kamera in einem gewissen Abstand hinter den Kampfrichtern als Alternative, ergibt ein sehr kleines Bild, was für die folgende Bewertung schlecht ist. Folglich wird eine leicht schräge Kameraposition gewählt, die geringe Unterschiede zur originalen Kampfrichterperspektive aufweist und für alle Aufnahmen identisch ist.
- Gute *Qualität der Videosequenzen*. Während der gesamten Übung steht der betreuende Trainer des Athleten neben dem Gerät. Bei Flugelementen (am Reck) oder wenn der Turner zu stürzen droht, versucht der Trainer so gut wie nur möglich den Sportler zu sichern. Manchmal wechselt er die Seite des Gerätes, um sich in eine entsprechend günstige Position zu bringen. Genau dieser Wechsel bewirkt in einigen Aufnahmen, dass der Athlet teilweise verdeckt wird. Der Trainer als zusätzliche Quelle der Bewegung muss bei der Auswahl der Übungen berücksichtigt und vermieden werden.

- *Turner stürzt nicht* – dieses Kriterium wird gewählt, da nicht bekannt ist, wie sich ein Sturz und eine damit verbundene Unterbrechung der Übung bis zu einer halben Minute auf die Bewertung der Kampfrichter auswirkt. Eine Studie aus den 60er Jahren hat gezeigt, dass die Übereinstimmung der Wertungen bei nicht kompletten Übungen, also mit Sturz, im Vergleich zu kompletten Übungen schlechter ausfällt (Faulkner & Loken, 1962, zitiert nach Landers, 1970).
- *Besonderheiten am Gerät Ringe*: Eine Übung besteht aus Halte- und Schwungelementen. In den meisten Fällen werden die Halteelemente zu Beginn gezeigt, da sie sehr kraftraubend sind. Sie haben die Eigenschaft, dass sie orthogonal betrachtet zwischen den Ringen geturnt werden. Bei Schwungelementen bewegt sich der Körper ober- und unterhalb der Ringe. Diese werden oftmals gegen Ende der Übung geturnt. Ein möglichst unbemerkter Videoschnitt kann allerdings nur gemacht werden, wenn nach dem Element eine identische Körperposition eingenommen wird. Beim oben beschriebenen Übungsaufbau ist das allerdings sehr schwer, wenn überhaupt möglich.
- *Besonderheiten am Gerät Reck*: Übungen an diesem Gerät beinhalten oftmals viele Drehungen um die Körperlängsachse. Außerdem wechseln die Turner innerhalb der Übung die Drehrichtung um die Reckstange, so dass sie aus Kampfrichterperspektive (neben dem Gerät) zum Teil im und zum Teil gegen den Uhrzeigersinn turnen. Darüber hinaus verändert sich die Handposition im Laufe der Übung, zum einen, was den gewählten Griff betrifft, und zum anderen, was den Abstand der Hände von den Reckstangen-Enden und auch zueinander betrifft. Diese Tatsachen erschweren die Videoschnittarbeit.
- *Auftreten eines deutlichen Fehlers*: die Sichtung erfolgt mit dem Ziel, die Schwere sowie die zeitliche Position der jeweils in den Originalvideos gezeigten Fehler aufzudecken. Damit wird die gezielte weitere Bearbeitung der Videos hin zu zwei unterschiedlichen Experimentalvideos, die die UV Reihenfolge abbilden, ermöglicht. Die optimale Übung zur weiteren Bearbeitung enthält daher einen großen Fehler in der ersten oder zweiten Hälfte der Übung. Die restliche Übung ist weitgehend fehlerfrei bzw. beinhaltet lediglich vereinzelte, kleine Fehler.

3. *Videoschnitt*[19]

- Die Videosequenzen werden jeweils zu *zwei Experimentalvideoversionen* zusammengestellt: Das Experimentalvideo eins ist dadurch gekennzeichnet, dass die Übung des Turners einen deutlich größeren Fehler in der ersten Hälfte zeigt. Die zeitliche Mitte des Videos beschreibt dabei auch die Mitte der Übung, wobei der An- und Abgang, also der definierte Beginn und das Ende keine groben Fehler aufweisen dürfen. Somit ist nicht nur der Sturz im Übungsverlauf, sondern auch am Ende der Übung in einer Experimentalübung ausgeschlossen. Experimentalvideo zwei unterscheidet sich von Video eins dadurch, dass ein größerer Fehler in der zweiten Hälfte der Übung zu sehen ist. Der komplette Videoschnitt wird mit dem Programm Adobe Premiere Elements 2.0 durchgeführt. Abbildung 4 zeigt eine schematische Darstellung der Videobearbeitung am Beispiel einer 75 Sekunden andauernden Übung. Diese Übung besteht inhaltlich aus unterschiedlichen Abschnitten, wie dem Angang, unterschiedlicher Elemente und dem Abgang. Entsprechend dieser Inhalte können dann geeignete Stellen ausgewählt werden, an denen man das Gesamtvideo in Abschnitte trennen kann. Diese Abschnitte werden in ihrer Reihenfolge vertauscht und wieder zu einer neuen manipulierten Videoversion zusammengesetzt. Damit erhält man ein manipuliertes Video (nach Videoschnitt), das bezüglich Inhalt und Dauer dem Originalvideo (vor Videoschnitt) gleicht. Die eigentliche Schwierigkeit dieser Videoschnittmethode ergibt sich durch die Suche der geeigneten Schnittstellen, so dass nach der Zusammenführung der Einzelabschnitte die Manipulation möglichst unbemerkt bleibt. In Kapitel 5.2.1 wird auf eine wichtige Voraussetzung für den verdeckten Videoschnitt hingewiesen: Die wiederkehrende Körperposition des Turners zum Gerät. Beispielhaft kann dies an Abbildung 4 gezeigt werden. Die Körperposition an der Schnittstelle vor bzw. nach Abschnitt 2 muss jeweils der Körperposition vor bzw. nach Abschnitt 4 gleichen. Dabei können die beiden Positionen vor dem Abschnitt sich von denen nach dem Abschnitt unterscheiden. Beispielsweise könnte der Athlet einmal die gestreckte Hangposition am Gerät zu Beginn des Fehlerabschnitts (Abschnitts 2)

[19] Aufgrund der Wahl der Videoschnitt-Methode und der in diesem Abschnitt beschriebenen Vorgehensweise können die Männergeräte ‚Boden', ‚Barren' und ‚Pauschenpferd' nicht verwendet werden. Der ‚Sprung' fällt als Gerät heraus, da die gewählte Fragestellung der ‚Reihenfolge' hier keinerlei Bedeutung hat (5.2.1). Alle Frauengeräte (Sprung, Stufenbarren, Schwebebalken und Boden) eignen sich aus den genannten Gründen ebenfalls nicht.

einnehmen und zum Ende dieses Abschnitts in der gestreckten Handstandposition sein. Dann muss das Austauschstück (Abschnitt 4) ebenfalls zu Beginn eine gestreckte Hangposition und zum Schluss die gestreckte Handstandposition beinhalten. Zu beachten ist hierbei auch, ob der Athlet von vorne oder von hinten zu sehen ist. Einen bedeutenden Stellenwert nehmen hierbei die Besonderheiten der einzelnen Geräte ein, die allerdings im Zusammenhang mit der zweiten Stufe der Videobearbeitung ‚Auswahl und Vortest' bereits beschrieben wurden.

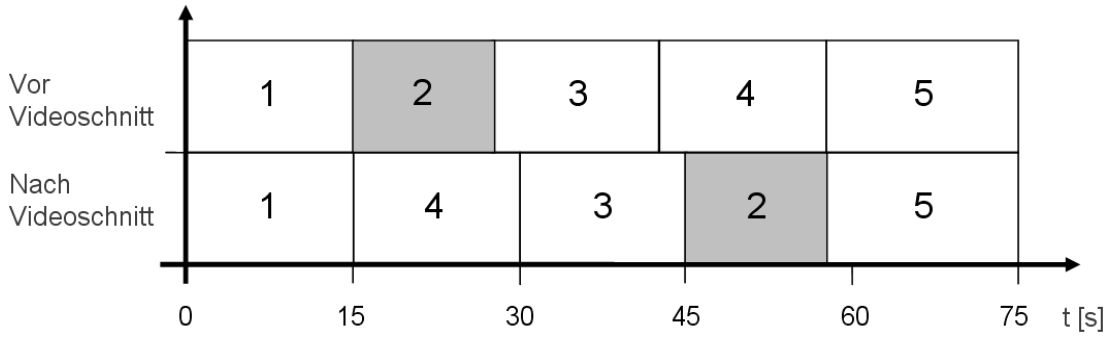


Abbildung 4: Videosequenz unterteilt in einzelne Abschnitte vor und nach dem Videoschnitt

4. Videonachbearbeitung

Die Überarbeitung der geschnittenen Videosequenzen erfolgt mit dem Programm Adobe After Effects 7.0. Ziel dieser Bearbeitungsstufe ist die Überdeckung von eventuell vorhandenen Indizien in den geschnittenen Videos, die darauf hinweisen, dass ein Schnitt stattgefunden hat. Beispielsweise wird in einigen Videos der sich bewegende Hintergrund mit einem Standbild überblendet.

Die *Kontrollübungen* stellen ungeschnittene Originalaufnahmen dar, die als Vergleichsstandard genutzt werden. In allen Untersuchungsbedingungen werden die identischen Videosequenzen gezeigt. Je nach Untersuchungsbedingung bekommen die VPn neben diesen Kontrollübungen entweder Experimentalübungen mit dem Fehler vorne oder hinten gezeigt. Die jeweilige Anordnung kann dem Untersuchungsdesign der jeweiligen Untersuchung entnommen werden (5.1).

5.3.3 Präsentation

Um ein weitgehend standardisiertes Vorgehen (5.2.4) sicherzustellen, werden alle notwendigen Informationen und die zu bewertenden

Videosequenzen in eine einheitliche Form gebracht. Zur Präsentation des Untersuchungsmaterials wird eine gebräuchliche PC-Anwendung (Microsoft PowerPoint) verwendet. Dabei erweist sich vor allem die unabhängige Untersuchungsdurchführung ohne Eingreifen des Versuchsleiters als vorteilhaft. Die geläufige Oberfläche der Anwendung unterstützt die Bereitschaft der VPn, an der Studie teilzunehmen. Darüber hinaus ermöglicht diese Form der Darstellung eine unkomplizierte Art der Kombination von schriftlichen Informationen und Videomaterial.

Insgesamt werden vier unterschiedliche Präsentationen entsprechend den vier Untersuchungsbedingungen erstellt. Den einzelnen Bedingungen werden die Farben - blau, gelb, grün und rot - zugeteilt (5.1). Innerhalb einer Präsentation sind alle für die Untersuchung notwendigen Informationen enthalten.

Nach einer kurzen Begrüßung der VPn folgt der jeweilige Instruktionstext. Danach werden alle Übungen des ersten Gerätes präsentiert, die jeweils mit einem schwarzen Bildschirm getrennt sind. Als nächstes folgen der Instruktionstext des zweiten Gerätes und wiederum alle Übungen. Am Schluss jeder Präsentation werden die VPn aufgeklärt und gebeten, sich nicht mit Kollegen über den Ablauf oder die Inhalte der Untersuchung auszutauschen. Ob die VPn dieser Bitte nachgegangen sind, wurde nicht kontrolliert. Danach füllen die VPn den Personenfragebogen aus (5.4).

5.4 Bewertungs- und Personenfragebogen

Um die Bewertung der Videosequenzen der entsprechenden Präsentation festhalten zu können, wurde ein *Bewertungsbogen* erstellt. In der realen Wettkampfsituation notieren sich die Kampfrichter des B-Kampfgerichts die individuellen Abzüge auf einem Blatt Papier, auf dem die Gesamtabzüge der zu bewertenden Übung zusammengezählt werden. Dieser Abzug wird dem Oberkampfrichter gegeben, der die B-Note ermittelt und durch die Addition zur A-Note des A-Kampfgerichts die Endnote für die Übung berechnet. Da für die Ermittlung der B-Note kein spezielles Hilfsmittel üblich ist, wird in den Experimenten ebenfalls nur ein entsprechender Platz zur Verfügung gestellt, auf dem die VPn ihre Abzüge notieren bzw. im Falle der Bedingung EoS nur ihre Gesamt-Abzüge aufschreiben können. Die Unterschiede der Bewertungsarten werden in den Instruktionen und den Bewertungsbögen umgesetzt. Es gibt zwei unterschiedliche Bewertungsbögen. Der eine mit einer speziellen Spalte für ‚notierte Abzüge' und einer Spalte ‚B-Abzüge' (Gesamtabzüge in der B-Note) und der andere ohne die Spalte ‚notierte

Abzüge'. Die unterschiedlichen Bewertungsbögen am Beispiel der ersten Untersuchung können im Anhang (A 1 & 2) eingesehen werden.

Die demographischen Daten sowie Daten, die Aufschluss über die Expertise der VPn geben sollen, werden über eine Befragung mittels eines selbst entwickelten *Personenfragebogens* sichergestellt. Das Alter, das Geschlecht sowie der Beruf werden erhoben. Auffällig dabei ist, dass Frauen im männlichen Turnen durchaus eine Lizenz erwerben können, es aber eher ungewöhnlich ist, dass eine Frau im Kampfgericht sitzt. Die aktuell vorhandene Lizenz, sowie die Anzahl der Jahre an Wertungsrichtertätigkeit, Trainertätigkeit bzw. aktiver Turnlaufbahn als mögliche Belege für die Expertise der VPn werden mit erhoben. Weiter abgefragt wird die durchschnittliche Anzahl der Einsätze als Kampfrichter des vorangegangenen Jahres. Der Personenfragebogen ist ebenfalls im Anhang (A 3a & b) enthalten.

6 Untersuchungsteilnehmer und Durchführung

6.1 Untersuchungsteilnehmer

Die Population der vorliegenden Untersuchungen bilden deutschsprachige Kampfrichter des Gerätturnens der Männer mit Wertungslizenzen unterschiedlicher Kategorien. Die Lizenzstufen, die entsprechend ihres Qualifizierungsgrades eingestuft werden, reichen von der internationalen Kampfrichterlizenz, die von der FIG (Fédération Internationale de Gymnastique - Internationaler Turnerbund) vergeben wird und als sehr hoher Qualifizierungsgrad angesehen wird, bis hin zu weniger hoch qualifizierten Lizenzen (2.2).

Sowohl weibliche, als auch männliche Kampfrichter können an den Studien dieser Arbeit teilnehmen. Frauen können Wertungslizenzen im männlichen Gerätturnen erwerben, stellen allerdings die Ausnahme dar. Somit ergibt sich ein sehr geringer Frauenanteil, der aber der repräsentativen Verteilung im männlichen Gerätturnen entspricht, vor allem auf höheren Kampfrichterlizenz-Leistungsebenen.

Bei den gewählten Stichproben der beiden Experimente handelt es sich nicht klassisch um Zufallsstichproben. Die Versuchspersonen (VPn) wurden aus einer Vorauswahl (Kampfrichter mittlerer bis hoher Kampfrichterlizenz) zufällig ausgewählt und mittels Randomisierung den einzelnen Untersuchungsbedingungen zugeordnet.

6.1.1 Rekrutierung der Studienteilnehmer

Potentielle Teilnehmer der Studien wurden im Rahmen verschiedener Veranstaltungen, wie Meisterschaften oder Kampfrichterschulungen akquiriert. Die gesamte Untersuchung des Erstexperiments und des Zweitexperiments, samt dem angehängten Personenfragebogen wurde anonym durchgeführt.

Bei der Rekrutierung der VPn wurde angestrebt, möglichst gut ausgebildete Kampfrichter für die Untersuchung zu gewinnen. Es sollte gewährleistet werden, dass die VPn eine bestimmte Expertise haben und die gezeigten hochklassigen Videosequenzen problemlos bewerten können. Die Expertise eines Kampfrichters kann anhand der erworbenen Lizenz eingestuft werden. Vermutet wird, dass eine ‚zu geringe' Expertise eventuell die Untersuchungsdaten verzerren könnte und damit der Rückschluss auf die manipulierten Variablen nur sehr wage möglich wird. Außerdem nehmen Experten mehr Fehler wahr als

Novizen (Bard et al., 1980). Ziel war daher die Rekrutierung möglichst vieler Kampfrichter mit internationaler Lizenz und nationaler Lizenz A und B. Dadurch sollte die Reliabilität und die interne Validität der Untersuchung erhöht werden. VPn niedrigerer Lizenzstufe stellen insgesamt einen sehr geringen Prozentsatz der Stichproben dar, wurden somit nicht von der Untersuchung ausgeschlossen.

Neben der zu geringen Expertise könnte sich auch die fehlende Erfahrung nachteilig auf die Untersuchungsergebnisse auswirken. Die Expertise ist nämlich keine ausreichende Garantie für eine hohe Beurteilungsqualität (Haase, 1972; Landers, 1970). Thomas (1978, S. 266) geht davon aus, dass sich langjährige eigene Erfahrung des Beurteilers als aktiver Sportler, Trainer oder Kampfrichter in der jeweiligen Sportart positiv auf die Beurteilungsfähigkeit auswirkt. Eine sehr ausgeprägte Bewegungsvorstellung erlaubt in Verbindung mit Erfahrungen aus der Eigenrealisation, dass ein Kampfrichter Ausführungsdetails selbst dann aus Folgeerscheinungen ableiten kann, wenn er das Detail an sich nicht gesehen hat.

Im Vorfeld jeder Untersuchung wurden die Organisatoren sowie die für die Veranstaltung eingesetzten Oberkampfrichter telefonisch oder schriftlich über die Studie informiert und in vorheriger Absprache wurde diskutiert, ob und welche organisatorischen Probleme bei der Akquise entstehen könnten und wie man diesen Schwierigkeiten entgegen wirken kann.

1. Das Erstexperiment

Für die erste Erhebung wurde im Dezember 2006 das Finale der Deutschen Turnliga in Heidelberg ausgewählt. Bei dieser Veranstaltung wurden aufgrund organisatorischer Gegebenheit für die Auswertung des Erstexperiments nicht genügend VPn untersucht. Ende Januar 2007 fand daher eine weitere Erhebung statt. Bei einem Kampfrichterlehrgang in Karlsruhe konnten die restlichen, für eine statistische Auswertung benötigten Teilnehmer untersucht werden. Insgesamt wurden somit für das Erstexperiment insgesamt 51 Teilnehmer rekrutiert.

Persönliche Gespräche mit den VPn und einige Anmerkungen im Personenfragebogen bezüglich der Bildqualität der gezeigten Videosequenzen führten zur Überlegung, ein weiteres Experiment mit verbesserter Bildqualität durchzuführen. Bei allen Videosequenzen wurde infolgedessen die Qualität verbessert, nicht aber der Inhalt der Videosequenzen verändert. Mit diesen neuen Videos wurden im Februar 2007

weitere 59 VPn bei einer Weiterbildungsveranstaltung in Ostfildern/ Ruit untersucht.

Da sich die Sequenzen inhaltlich nicht unterscheiden, werden alle erhobenen Datensätze aufgrund der Übersichtlichkeit einer Untersuchung untergeordnet. Somit haben insgesamt 110 VPn das experimentelle Videomaterial des Erstexperiments gesichtet und bewertet. Selbstverständlich wird aufgrund der unterschiedlichen Videomaterialqualität eine adäquate Behandlung der beiden untersuchten Stichprobengruppen im Rahmen der Auswertung und der Interpretation der Daten erfolgen.

2. Das Zweitexperiment

Die empirischen Erhebungen des Zweitexperiments fanden in den Monaten November 2007 und Januar sowie Februar des Jahres 2008 statt. Die Besonderheit dieser Untersuchung stellt die lizenzierten Kampfrichter aus den drei deutschsprachigen Nationen Deutschland, Schweiz und Österreich dar. Um dabei mögliche Verzerrungen durch eine Übersetzung zu vermeiden, beschränkte sich die Auswahl der VPn auf Nationen aus dem deutschsprachigen Raum.

In Anbetracht des zur Verfügung stehenden personellen, zeitlichen, finanziellen und organisatorischen Rahmens war frühzeitig antizipierbar, dass es nicht möglich sein würde, eine Vollerhebung der Kampfrichter aus der Bundesrepublik Deutschland und den deutschsprachigen Nachbarländern durchzuführen. Um einen Nationenvergleich ziehen zu können, benötigt man eine entsprechend große Stichprobe je Nation. Nationale Vergleiche können daher nur in sehr geringem Maße realisiert werden. Die akquirierte Stichprobe wird als Gesamtstichprobe angesehen. Wie in Kapitel 2.3 erläutert, stellt der CdP (Code de Pointage – internationale Wertungsvorschriften) die Grundlage jeder Bewertung dar, wodurch eine Zusammenführung der Untersuchungsdaten möglich wird.

Das gesamte Datenmaterial wurde in vier Etappen eingeholt. Die erste Erhebung fand Anfang November 2007 im österreichischen Rif/ Hallein bei den 61. Österreichischen Staatsmeisterschaften statt, wobei 25 VPn teilnahmen. Die zweite Erhebung mit identischen Rahmenbedingungen konnte Ende November 2007 im Rahmen des in Heidelberg ausgetragenen Finales der Deutschen Turnliga durchgeführt werden. An dieser Veranstaltung konnten 17 VPn experimentell untersucht werden. Weitere 29 VPn bewerteten das Untersuchungsmaterial bei einer Kampfrichterfortbildung im deutschen Bartholomä im Januar 2008. Die letzte Datenerhebung fand im Februar 2008 statt. Daran beteiligten sich im

Rahmen eines Kampfrichterlehrgangs im schweizerischen Magglingen 35 lizenzierte Kampfrichter.

6.1.2 Beschreibung der Stichprobe

Für die *erste Untersuchung* liegen die Daten von 110 lizenzierten Kampfrichtern des männlichen Gerätturnens vor. Die nachfolgenden Auswertungen beziehen sich auf 102 lizenzierte VPn. Die Personendaten wurden anhand eines Personen-Fragebogens erhoben. Dieser wird in Kapitel 5.4 näher erläutert. In Kapitel 7.3 wird detailliert beschrieben, aus welchen Gründen acht VPn aus der Auswertung genommen wurden. Tabelle 5 gibt einen Überblick über die demographischen Angaben der Personenstichprobe des Erstexperiments.

Tabelle 5: Demographische Angaben der Erstexperiment-Teilnehmer

		Geschlecht	Alter	Kampfrichter-Tätigkeit in Jahren	Aktuelle Lizenz	Turner-Tätigkeit	Turner-Tätigkeit in Jahren	Trainer-Tätigkeit	Trainer-Tätigkeit in Jahren	Anzahl der Wettkämpfe 2006
N	Gültig	102	102	102	102	102	96	102	80	102
	Fehlend	0	0	0	0	0	6	0	22	0
M		1,99	38,24	13,59	2,47	1,05	19,38	1,22	13,19	8,64
SD		,10	14,29	10,60	1,01	,217	9,95	,41	11,49	6,18
Minimum		1	16	0	1	1	4	1	1	0
Maximum		2	73	39	5	2	59	2	49	30

Die Gruppe der VPn besteht aus 101 männlichen (99%) und einer weiblichen Kampfrichterin, die zwischen 16 und 73 Jahre alt sind (M = 38,24, SD = 14,29). Die Kampfrichtertätigkeit üben die untersuchten VPn durchschnittlich 13,59 Jahre (SD = 10,6) aus. Über kein ganzes Jahr an Erfahrung berichten 5 Personen (4,9%). 43% (44) aller VPn des Erstexperiments geben an bis zehn Jahre dieser Tätigkeit bereits nachzugehen, während 30,4% (31) angeben, 11 bis 20 Jahre als Kampfrichter an Wettkämpfen teilzunehmen. Zwischen 21 bis 30 Jahre führen 13,8% der Teilnehmer diese Tätigkeit aus. Über bereits mehr als 30 Jahre Erfahrung als Kampfrichter haben 7,8% (8) der Teilnehmer.

Angegeben werden verschiedenartigste Berufsgruppen, von Ingenieuren (10 Personen), über Handwerker unterschiedlicher Bereiche bis hin zu Lehrern (6). Hinzu kommen Schüler (6), Studenten (10) und Rentner (8).

Zum Untersuchungszeitpunkt waren oder sind 97 VPn (95,1%) über durchschnittlich 19,38 Jahre (SD = 9,95) selbst als Turner aktiv im Gerätturnen. Lediglich 5 VPn geben an, keine eigene Erfahrung als Turner gesammelt zu haben. Auf die Frage einer früheren oder aktuellen Trainertätigkeit antworten 80 (78,4%) VPn positiv und geben an durchschnittlich 13,19 Jahre (SD = 11,49) dieser Tätigkeit nachgegangen zu sein bzw. nachzugehen. 22 Personen verneinen diese Frage und haben somit keine Trainertätigkeit bis zum Zeitpunkt der Untersuchung ausgeübt.

Die durchschnittliche Anzahl an Wettkämpfen, die die getesteten Kampfrichter im Jahr 2006 bewertet haben, liegt bei durchschnittlich 8,64 Einsätzen (SD = 6,18). Die Bandbreite ist dabei weit gestreut und reicht von 0 bis hin zu 30 Einsätzen. Keinen Kampfrichtereinsatz hatten 5 Personen (4,9%).

Alle Kampfrichter besaßen zum Zeitpunkt der Untersuchung eine gültige Kampfrichterlizenz auf unterschiedlichen Niveaus. Die höchste, die internationale Lizenz haben 20 Personen (19,61%). 31 (30,39%) VPn besitzen die nationale Lizenzstufe A, 36 (35,29%) die Lizenz B, 13 (12,75%) Lizenz C, während lediglich 2 (1,96%) VPn die Lizenzstufe D erworben haben. Die durchschnittliche Lizenzhöhe liegt bei einem Wert von 2,47 (SD = 1,01), wobei der Wert 1 der internationalen Lizenz zugeordnet ist und der Wert 5 der Lizenzstufe D (Abbildung 5).

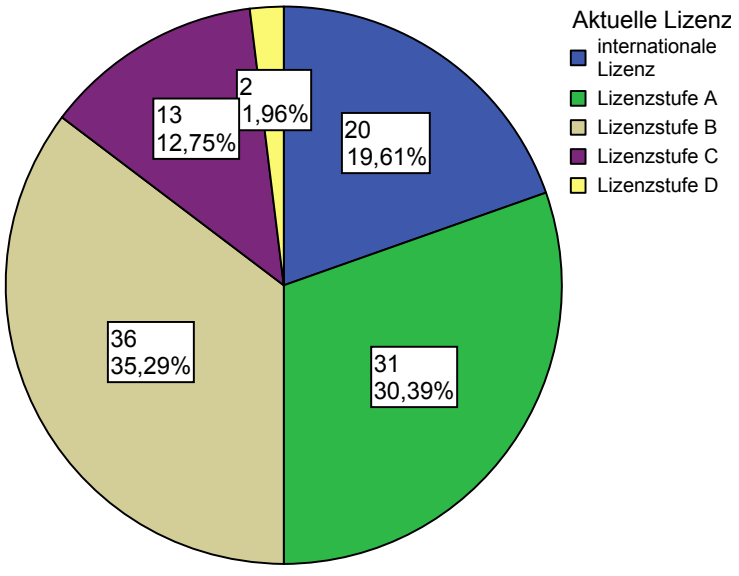


Abbildung 5: Aktuelle Lizenzhöhe der VPn des Erstexperiments

Insgesamt konnten für die *zweite Untersuchung* 106 VPn-Datensätze erhoben werden. Eine Datenbereinigung, wie sie im Erstexperiment

durchgeführt wurde, musste hier nicht angewendet werden, da aufgrund der vorgenommenen Verbesserungen keine Verletzungen der Kriterien stattgefunden hat (7.3).

104 (98,1%) männliche und 2 (1,9%) weibliche VPn, im Alter zwischen 18 und 77 Jahren (M = 38,68, SD = 13,48), haben am Zweitexperiment teilgenommen. Die Nationen Österreich (AUT), Deutschland (GER) und die Schweiz (SUI) sind in unterschiedlicher Kampfrichteranzahl vertreten. Eine Kampfrichtertätigkeit üben die untersuchten VPn durchschnittlich 14,54 Jahre (SD = 10,39) aus. 45,1% (48) aller VPn geben an, zwischen zwei und zehn Jahre Erfahrung als Kampfrichter zu besitzen. 34,3% (36) aller VPn des Zweitexperiments geben an, 11 bis 20 Jahre dieser Tätigkeit bereits nachzugehen, während 12,8% (13) angeben, 21 bis 30 Jahre als Kampfrichter an Wettkämpfen teilzunehmen. Über 30 Jahre Erfahrung als Kampfrichter haben 7,8% (8) der Teilnehmer.

Wenn man sich die Lizenzstufen getrennt nach nationaler Zugehörigkeit der Kampfrichter betrachtet, ergibt sich folgendes Bild: Insgesamt vermerken 25 Personen (23,6%) für das Land Österreich Kampfrichtertätigkeiten auszuführen, während 46 VPn (43,4%) in der Bundesrepublik Deutschland tätig sind und 35 Versuchsteilnehmer (33%) angeben, die Schweiz national zu vertreten.

Auch im Zweitexperiment gehören die angegebenen Berufsgruppen den unterschiedlichsten Branchen und Status an. Tabelle 6 veranschaulicht die demographischen Daten der Zeitexperiment-Stichprobe.

Tabelle 6: Demographische Angaben der Zweitexperiment-Teilnehmer

		Geschlecht	Alter	Kampfrichter-Tätigkeit in Jahren	Nation	Aktuelle Li-zenz	Turner-Tätigkeit	Turner-Tätigkeit in Jahren	Trainer-Tätigkeit	Trainer-Tätigkeit in Jahren	Anzahl der Wettkämpfe 2007
N	Gültig	106	106	102	106	104	106	101	106	86	106
	Fehlend	0	0	4	0	2	0	5	0	20	0
M		1,98	38,68	14,54	2,09	1,94	1,02	15,35	1,18	14,70	8,47
SD		,14	13,48	10,39	,75	,97	,137	6,31	,39	10,53	5,75
Minimum		1	18	2	1	1	1	3	1	1	0
Maximum		2	77	45	3	4	2	33	2	41	30

Als Turner waren oder sind 104 VPn (98,1%), über durchschnittlich 15,35 Jahre (SD = 6,31), selbst aktiv. Die Angaben reichen von drei bis 33 Jahre aktive Turnertätigkeit. Lediglich zwei VPn weisen keine eigene Erfahrung als Turner vor. Somit sind fast alle der untersuchten Kampfrichter selbst einmal aktiv Turner gewesen. Frühere oder aktuelle Trainertätigkeit gehen 87 (82,1%) Teilnehmer nach. Sie geben an, durchschnittlich 14,70 Jahre (SD = 10,53) dieser Tätigkeiten nachgegangen zu sein bzw. nachzugehen. 19 Personen verneinen diese Frage und haben somit keine Trainertätigkeit bis zum Zeitpunkt der Untersuchung ausgeübt.

Durchschnittlich geben die VPn an, im Jahr 2007 bei 8,47 (SD = 5,75) Wettkämpfen als Kampfrichter eingesetzt worden zu sein.

Die untersuchten Kampfrichter geben eine durchschnittliche Lizenzhöhe von 1,94 (SD = 0,97) an. Die internationale Lizenz wurde von 42,3% (44 VPn) angegeben. Weiterhin geben 30 Personen (28,8%) an, die höchste nationale Lizenzstufe A zu besitzen. 22 (21,1%) VPn besitzen die nationale Lizenzstufe B und 8 (7,7%) die Lizenz C (Abbildung 6). Von den untersuchten VPn gibt keiner an, die Lizenzstufe D zu besitzen.

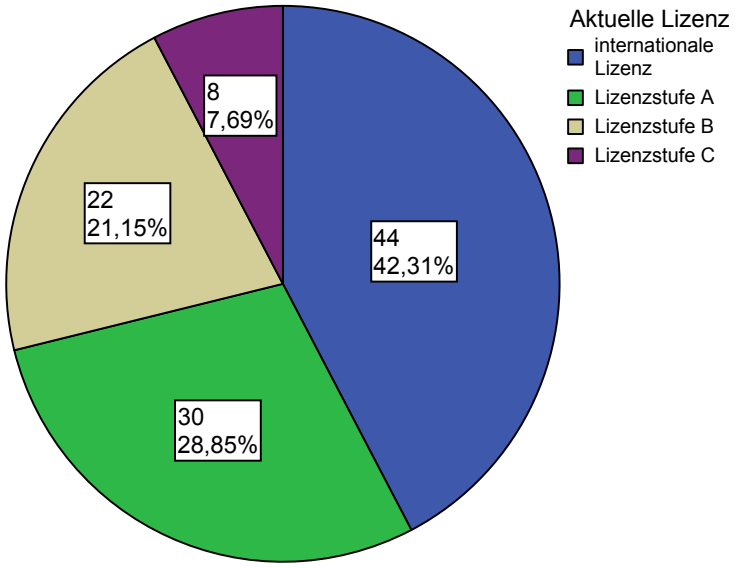


Abbildung 6: Aktuelle Lizenzhöhe der VPn des Zweitexperiments

2 VPn geben leider keine Lizenzstufe an. Allerdings kann davon ausgegangen werden, dass diese entsprechend vergleichbar mit den anderen ausgebildet sind, da sie ansonsten nicht an der Veranstaltung bzw. an der Fortbildung hätten teilnehmen können. Aus diesem Grund werden sie nicht aus der Auswertung gestrichen. Die Auswahl der

Veranstaltungen hat selbstverständlich dieses Ergebnis forciert und war nicht anders zu erwarten.

Unter den 25 VPn österreichischer nationaler Herkunft geben 11 Personen an, die internationale Lizenz zu besitzen. Weitere 8 Personen haben die national höchste Zulassung für Kampfrichter erworben. Ein Kampfrichter gibt an, die zweithöchste Lizenz zu besitzen, während fünf österreichische Unparteiische die dritthöchste nationale Lizenz erworben haben. Keine Nennung fällt auf die vierthöchste und damit niedrigste Lizenzstufe. Bei den deutschen Kampfrichtern ergibt sich ein ähnliches Bild. Von den 46 untersuchten Personen haben acht die internationale, 16 die höchste und 19 die zweithöchste nationale Lizenzstufe erreicht. Weitere drei VPn geben an, die dritthöchste Lizenz zu besitzen. Bei den schweizerischen Kampfrichtern fällt besonders auf, dass von 35 untersuchten Personen 25 (71,4%) VPn eine internationale Lizenz erworben haben. Sechs VPn geben an, die höchste nationale Lizenz zu besitzen, während zwei die zweithöchste Lizenzstufe errungen haben. Darunter liegende Lizenzen werden von keiner VPn angegeben.

6.2 Durchführung der Untersuchung

Die Studienteilnehmer wurden im Rahmen ihrer Kampfrichtertätigkeit an verschiedenen Veranstaltungen aufgesucht. Die Durchführung der Untersuchung erfolgte, um den Ablauf nicht zu stören, meist nach dem Ende der besuchten Veranstaltung. In einigen Fällen wurden auch größere Pausen genutzt, um den Versuch durchzuführen.

Bei der Durchführung der Untersuchung wurde auf die Erzeugung einer möglichst wettkampfnahen Situation geachtet. Die einzigen Unterschiede für den Kampfrichter, im Vergleich zur Real-Situation im Wettkampf, sind der separate Raum, die Darbietung der Übungen auf einem Bildschirm, die leicht veränderte Perspektive auf den turnenden Athleten, die keine Blicksprünge erforderlich macht, und bei einigen der VPn die Beurteilung der präsentierten Übungen ohne bzw. mit dem Notieren der vorzunehmenden Abzüge.

Obwohl der Versuchsraum aufgrund der unterschiedlichen Veranstaltungen nicht konstant gehalten werden konnte, waren die Orte der Durchführung ähnlich. Es handelte sich um Seminarräume oder Räume, die zur Unterbringung von Tischen, Stühlen und sonstigen Gegenständen dienten. Diese waren durch Türen verschließbar, so dass sie räumlich und akustisch von anderen Geschehnissen abgegrenzt waren.

Außerhalb des Versuchsraumes wurde ein Schild angebracht, das die Benutzung von Handys während der Studie untersagt, um Störungen dieser Art zu vermeiden.

Während der Durchführung befanden sich nur die Versuchsleiterin und die Versuchsteilnehmer im Untersuchungsraum. Dadurch konnte ein konzentriertes Arbeiten der Kampfrichter sichergestellt werden. Im Versuchsraum war jeweils genügend Platz für die Studie vorhanden. Jede VPn hatte einen Arbeitsplatz mit Laptop und einen Stuhl zur Verfügung. Meistens bearbeiteten mehrere Kampfrichter das Experiment zur gleichen Zeit, wobei jeweils darauf geachtet wurde, dass sich die Versuchsteilnehmer weder behindern, noch beeinflussen konnten.

Ein standardisierter Ablauf der Untersuchung wurde durch identische technische Bedingungen, dieselbe Versuchsleiterin, die einheitliche (nicht zu lange) Bearbeitungszeit, die Aufklärung über den Sinn und Zweck der Untersuchung sowie die Anweisungen an die Kampfrichter, weder über den Ablauf noch über den Inhalt der Untersuchung mit Kollegen zu sprechen, sichergestellt.

Der Ablauf der Untersuchung sieht wie folgt aus (vgl. Abbildung 7): Im Vorfeld der Untersuchung wurde jedem zusammenhängenden Papierbogen, bestehend aus einem Bewertungsbogen (Anhang, A 1 & 2) und einem Personenfragebogen (Anhang, A 3a & b), eine VPn-Nummer zugeordnet. Den Nummern wurde in einem weiteren Schritt, in zufälliger Anordnung, einer von vier Farbpunkten – blau, gelb, grün, oder rot - zugeteilt. Jeder Farbpunkt repräsentiert eine der vier Untersuchungsbedingungen (5.1). Somit ist jeder Papierbogen mit einer VPn-Nummer und einem Farbpunkt versehen.

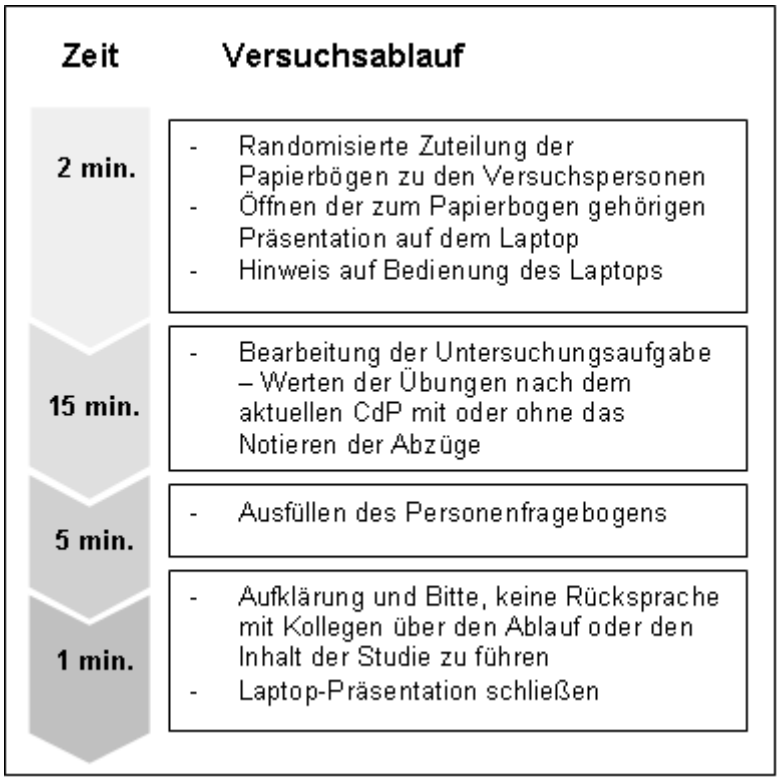


Abbildung 7: Schematische Darstellung des Versuchsablaufs

Von einem Stapel an Papierbögen nimmt die Versuchsleiterin den obersten Bogen, der beispielsweise mit einem blauen Punkt versehen ist. Je nach Farbpunkt des Bewertungsbogens wird die dazugehörige Präsentation (hier: blau) geöffnet, die den Namen der Farbe trägt und die Kriterien der entsprechenden UB, in diesem Fall der Untersuchungsbedingung eins, (5.3) beinhaltet. Der Papierstapel wird neben den Laptop gelegt, wo sich auch ein Stift zum Ausfüllen der Papiere befindet. Somit wird sichergestellt, dass an einem Laptop jede beliebige VPn in jeder beliebigen Untersuchungsbedingung die Studie durchführen kann. Durch die randomisierte Zuteilung der Nummern zu den Papierbögen erfolgt auch eine Randomisierung der VPn, da diese in zufälliger Reihenfolge an der Untersuchung teilgenommen haben.

Die VPn betritt das Versuchszimmer und wird nach der Begrüßung direkt zu einem Laptop begleitet. Er wird darauf hingewiesen, dass alle weiteren Informationen in der Präsentation enthalten sind und er mit der Bedienung der Leertaste des Laptops zur nächsten Folie kommt. Als zusätzliche Information wird ihm gesagt, dass die Präsentation unterschiedliche Text- und Videosequenzen beinhaltet. Außerdem wird an das Ausfüllen des angehängten Personenfragebogens erinnert, der auf keinen Fall vergessen werden soll. Das Experiment wird gestartet.

Die Aufgabe der VPn besteht darin, die dargestellten Videosequenzen nach den aktuellen, verbindlichen Wertungsvorschriften der FIG (Fédération Internationale de Gymnastique - internationaler Turnerbund), dem CdP, zu bewerten (2.3). Sie müssen die Übungen von jeweils vier bzw. sechs Turnern an den Geräten Ringe und Reck beurteilen. Die seit Anfang 2006 geltenden Wertungsvorschriften sind so ausgelegt, dass der Faktor ‚Übungsausführung' (B-Note) gegenüber dem ‚Übungsinhaltswert' (A-Note) extrem hoch in die Wertung eingeht. In der durchgeführten Studie übernehmen die VPn genau diese wichtiger gewordene Aufgabe des B-Kampfgerichts. Dabei müssen sie die Abweichungen in der Ausführung einer Übung von den Idealnormen im CdP feststellen und mit entsprechenden Abzügen ahnden. Je nach Untersuchungsbedingung sollen die VPn entweder ihr Urteil mit Hilfe des Mitschreibens der Abzüge oder ohne dieses Mittel fällen (5.2.2).

Die vom Kampfrichter durchgeführten Abzüge bilden die abhängige Variable. Die B-Note stellt die Höhe der ‚Übungsausführung' dar und berechnet sich als Differenz aus dem Punktewert 10,0 und den vorgenommenen Abzügen (2.3). In der Realsituation eines Wettkampfes wird diese Note zum ‚Übungsinhaltswert' addiert, um die Endnote der entsprechenden Übung zu bilden. Da für die Untersuchung die A-Note irrelevant ist, wird sie auch nicht erhoben bzw. keine Endnote gebildet.

Die zu bewertenden Übungen werden im Experiment nur einmal und in Echtzeit gezeigt, um so nah wie möglich an den Realbedingungen des Wertens zu sein. Der Faktor Zeitdruck wird bewusst nicht in die Experimente integriert. Dieser resultiert aus der im CdP explizit vermerkten Forderung, dass die B-Kampfrichter 10 Sekunden nach Beendigung der Übung die B-Note fertig gestellt haben müssen (2.3). Allerdings wird in der Wertungspraxis dieser Vorschrift, im Gegensatz zu anderen, keine Beachtung geschenkt. Wenn die Note fertig ist, ist sie fertig!

Während die Kampfrichter die Übungen bewerten, wird bei Bedarf von der Versuchsleiterin ein Protokollbogen ausgefüllt, indem Anmerkungen besonderer Vorkommnisse, wie beispielsweise technische Probleme, notiert werden. Sobald die VPn mit dem Experiment fertig ist, wird überprüft, ob der Personenbogen komplett ausgefüllt wurde. Falls dieser Lücken hat oder sogar leer ist, wird die VPn gebeten dies nachzuholen. Abschließend wird der Laptop wieder in den Ausgangszustand versetzt, in dem die Präsentation geschlossen wird, um dieselben Rahmenbedingungen für die nächste Person bereit zu stellen.

6.3 Vorgenommene Änderungen im Zweitexperiment

Aus den im Erstexperiment gemachten Erfahrungen wurden Konsequenzen für eine verbesserte Durchführung des Zweitexperiments gezogen. Erläutert werden im Folgenden alle Änderungen die im Zweitexperiment durchgeführt wurden. Dabei sind auch Verbesserungen enthalten, die sich im Rahmen der Datenbereinigung (7.3) ergeben haben.

Eine erste durchgeführte Veränderung im Vergleich zum Erstexperiment, stellt der Zugang zu den Videosequenzen dar. Im Erstexperiment wurden Übungen verwendet, die an einem internationalen Wettkampf gefilmt wurden (5.3.1). Für das Zweitexperiment war es nicht möglich geeignete Videos mit beispielsweise den Kriterien Qualität und (Kampfrichter-)Perspektive in benötigter Anzahl zu organisieren. Um möglichst wenig Kompromisse bezüglich derartiger Kriterien einzugehen bzw. die geplante Untersuchung durchführen zu können, wurden die *Videosequenzen selbst aufgenommen*. Dadurch konnte eine verbesserte Auswahl und eine striktere Einhaltung der Kriterien erreicht werden. Durch die Wahl einer nationalen Meisterschaft in Rumänien und dem knappen Zeitraum zwischen Aufnahme (Ende Juli 2007) und erster Präsentation (Anfang November 2007) wurden hauptsächlich folgende Vorteile sichtbar. Zum einen war es sehr unwahrscheinlich, dass die für die Untersuchung in Frage kommenden deutschsprachigen Kampfrichter die Meisterschaft gesehen haben, da eine Fernseh-Übertragung nur im rumänischen Fernsehen zu sehen war und diese meist nur die bekanntesten Turner zeigt. Im Internet kursieren Übungen von aktuelleren Wettkämpfen fast ausschließlich aus dem populäreren, weiblichen Turnen und eher selten aus dem männlichen Bereich. Außerdem kommt es selten vor, dass die dort gezeigten Übungen aus Kampfrichterperspektive aufgezeichnet werden. Auch medienwirksame Kamerabewegungen und Nahaufnahmen sind für Untersuchungen wie diese nicht geeignet. Zum anderen kann das Risiko fast ausgeschlossen werden, dass diese Kampfrichter die Übungen bereits in einer Kampfrichterausbildung besprochen hatten, da hierfür Videos bei den entsprechenden Ausbildungsleitern vorliegen müssten. Eine vorherige Bewertung hätte vermutlich zu einer veränderten Einschätzung der Übungen geführt.

Weiterhin wurden die Videosequenzen von einem zusätzlichen *Vorrater* bewertet, um das eventuelle Risiko der Nichterkennung oder Fehlbewertung von Fehlern auf ein Minimum zu reduzieren.

Im Zweitexperiment konnten aufgrund der vermehrten Auswahl an Videosequenzen *mehrere Experimentalübungen* angeboten werden. Es

wird davon ausgegangen, dass dadurch detaillierte Aussagen über einen eventuell vorhandenen Effekt möglich werden.

Im Untersuchungsplan wurde auch eine relevante Änderung durchgeführt. Die in der Untersuchung präsentierte *Videoversion eins* besteht nun komplett aus Übungen in denen der *‚Fehler vorne'* zu sehen ist und damit einer Ausprägung der UV Reihenfolge angehört. Die Videoversion zwei besteht entsprechend aus der zweiten Ausprägung der UV und beinhaltet nur Übungen, die einen groben Fehler am Ende zeigen. Durch diesen Wechsel kann vermutet werden, dass die anschließende Auswertung zu umfangreicheren Aussagen befähigt.

Wie detaillierter in Kapitel 7.3 erläutert, wurde versucht, durch eine *Komprimierung der Videosequenzen* kleinere Dateien zu produzieren. Der Arbeitsspeicher der Laptops wurde daher weniger stark in Anspruch genommen und damit konnte das Risiko technischer Probleme während der Untersuchungsdurchführung stark reduziert werden.

Im Erstexperiment konnte die Versuchsleitung zu Beginn bzw. im Verlauf der Experimentalphase bei einigen VPn beobachten, dass nicht ganz eindeutig zu sein schien, wie der Laptop zu bedienen ist. Die VPn mussten, um zur nächsten Text- bzw. Videosequenz zu gelangen, nur eine Taste betätigen. An diese wurde im Rahmen des Zweitexperiments ein *visueller Hinweis* angebracht, um aufkommende Verunsicherung bei den VPn zu verhindern.

Um während der gesamten Untersuchung einen guten Überblick über die bereits bewerteten und die noch folgenden Übungen an den beiden Geräten zu behalten (7.3), wurde folgende Verbesserung durchgeführt. Damit sich die VPn vor jeder zu bewertenden Übung davon überzeugen konnten, nicht versehentlich eine Übung übersprungen zu haben, wurden die eingeblendeten schwarzen Zwischenfolien mit dem *Gerät und der Nummer des folgenden Turners* versehen.

Als letzte Verbesserung im Vergleich zum Erstexperiment, wurde im Zweitexperiment ein Wechsel der präsentierten Gerätereihenfolge hin zur üblichen *olympischen Reihenfolg*e, zuerst Ringe, dann Reck, vorgenommen. Damit sollte ein möglichst natürlicher Ablauf simuliert und vermieden werden, dass bei den Kampfrichtern Fragen bezüglich dieser Reihenfolge aufkommen.

7 Hypothesen und statistische Auswertung

7.1 Hypothesen

Das ‚belief-adjustment'-Modell (Einhorn & Hogarth, 1992) stellt Vorhersagen zum Reihenfolge-Effekt auf. Um sich daran orientieren zu können, muss die Bewertungsaufgabe eines Kampfrichters einer der Kategorien (siehe 2.4) der Aufgabenschwierigkeit zugeordnet werden. Bei komplexen Aufgaben, wie der Bewertungsaufgabe von Kampfrichtern, besagt das Modell, dass es unabhängig vom Bewertungstyp zu einem Recency-Effekt kommt.

Auf dieser Grundlage basierend und aufgrund theoretischer Überlegungen zur sozialen Urteilsbildung und den bisherigen Befunden zum Positions-Effekt und zum Reihenfolge-Effekt können a priori folgende unspezifische Unterschiedshypothesen formuliert werden. Sie sind aus den wissenschaftlichen Hypothesen abgeleitet, die in Kapitel 4 beschrieben sind. In den formulierten Hypothesen stellt μ den Mittelwert[20] der erhobenen Kampfrichterwertungen dar und die Zahl entspricht der jeweiligen Untersuchungsbedingung[21] (5.1).

1. Hypothese zum Reihenfolge-Effekt

Ein Kampfrichter im Gerätturnen nimmt geringe Abzüge in der Übungsausführung (B-Note) vor, wenn eine Übung gut geturnt wurde. Die Bewertung von Kampfrichtern für Übungen mit Fehlern in der ersten Hälfte (Fehler vorne - V) unterscheidet sich von der Bewertung von Übungen mit Fehlern in der zweiten Hälfte (Fehler hinten - H). Somit gibt es einen Reihenfolge-Effekt. Entweder gehen die zuerst dargebotenen Informationen stärker in die Bewertung der Übung ein oder die zuletzt dargeboten Informationen (Tabelle 7).

Ho-1: $\mu 1$ (V) = $\mu 3$ (H) **H1-1**: $\mu 1$ (V) ≠ $\mu 3$ (H)

Ho-2: $\mu 2$ (V) = $\mu 4$ (H) **H1-2**: $\mu 2$ (V) ≠ $\mu 4$ (H)

[20] Der Erwartungswert ‚μ' bildet den arithmetischen Mittelwert der Grundgesamtheit ab. Der Stichprobenmittelwert hingegen wird durch das arithmetische Mittel ‚M' beschrieben und stellt einen Schätzwert für den tatsächlichen Erwartungswert dar.

[21] Im Rahmen des Erstexperiments muss bei der Betrachtung der 1. Hypothese zum Reihenfolge-Effekt die Besonderheit berücksichtigt werden, dass nicht alle Übungen in Untersuchungsbedingung eins oder zwei (bzw. drei oder vier) einen Fehler vorne (bzw. hinten) zeigen. In der Datenauswertung wird diese Besonderheit selbstverständlich berücksichtigt.

Tabelle 7: Mögliche Reihenfolge-Effekte bei der Beurteilung von Turnübungen

Reihenfolge (Videoversion)	Effekt
Abzüge für Fehler vorne (V) < Abzüge für Fehler hinten (H)	Recency
Abzüge für Fehler vorne (V) > Abzüge für Fehler hinten (H)	Primacy

2. Hypothese zum Bewertungstyp

Der Reihenfolge-Effekt entsteht unabhängig vom Bewertungstyp. Erwartet wird somit, dass kein Unterschied bezüglich des Bewertungstyps aufgedeckt wird. Es ist unerheblich für die Wertung einer Übung, ob ein Kampfrichter die Abzüge mitschreiben darf oder nicht.

Ho-3: $\mu 1 = \mu 2$ **H1-3**: $\mu 1 \neq \mu 2$

Ho-4: $\mu 3 = \mu 4$ **H1-4**: $\mu 3 \neq \mu 4$

3. Hypothese zur Interaktion der beiden Faktoren

Die beiden Faktoren Reihenfolge (R) und Bewertungstyp (B) interagieren nicht. Vermutet wird, dass sich die beiden Faktoren nicht gegenseitig bedingen und es so nicht beispielsweise zu einer höheren Wertung kommt, wenn der Kampfrichter nicht mitschreiben darf und der Fehler vorne ist. Alle Erwartungswerte in den Zellen setzen sich additiv aus den jeweils dazugehörigen Randerwartungswerten, verringert um den Erwartungswert der Gesamtpopulation, zusammen.

Ho-5: $\mu RB = \mu R + \mu B - \mu$ **H1-5**: nicht Ho

4. Hypothese zum Geräte-Effekt

Der Reihenfolge-Effekt (Effektgröße partielles η^2) ist an dem Gerät, an dem ohne Übungshalte geturnt wird (schnelles Gerät, Reck) größer, als an dem Gerät mit Übungshalten (langsames Gerät, Ringe).

Ho-6: $\eta^2_{Ringe} = \eta^2_{Reck}$ **H1-6**: $\eta^2_{Ringe} < \eta^2_{Reck}$

7.2 Stichprobenplanung

Zentraler Bestandteil der Planungsphase ist die Durchführung einer Teststärkenanalyse. Sie ist relevant, wenn man quantitative Studien durchführen möchte und das Ziel anstrebt, generalisierbare Aussagen zu treffen. Diese Generalisierbarkeit kann jedoch bei kleinen Fallzahlen unter einer geringen Teststärke leiden. In vier aufeinander folgenden Schritten wurde die Teststärkenberechnung durchgeführt:

Im ersten Schritt wird a priori die Effektgröße, die für die geplante Untersuchung prognostiziert wird, bestimmt. Durch die Festlegung einer Effektgröße spezifiziert man neben dem H0-Parameter den H1-Parameter. Damit wird bei einem nicht signifikanten Ergebnis (H0 annehmen) die β-Fehler-Wahrscheinlichkeit bzw. bei einem signifikanten Ergebnis (H1 annehmen) die α-Fehler-Wahrscheinlichkeit kalkulierbar. Grundsätzlich kann die Effektgröße auf zweierlei Weise generiert werden. Einerseits besteht die Möglichkeit, den Vorschlägen von Cohen (ebenda) für kleine, mittlere und große Effekte zu folgen und sich für die konkrete Studie zu überlegen, welcher Effekt erwartet werden kann. Andererseits kann man ähnliche Untersuchungen analysieren und die dort berichteten Effekte als Schätzwert annehmen. Die Effektgröße ist ein ausschlaggebender Faktor zur Ermittlung der erforderlichen Stichprobengröße. Kleine Effekte können nur mit entsprechend großen Stichprobengrößen nachgewiesen werden.

Somit klärt der *zweite Schritt* die Frage „Wie viele Versuchspersonen (VPn) müssen überhaupt untersucht werden, um einen entsprechenden schwachen, mittleren oder gar starken Effekt signifikant (alpha < 5%) zu bekommen? Diese Frage ist deshalb relevant, weil bei einer zu kleinen Stichprobengröße die Wahrscheinlichkeit steigt, dass ein Effekt – obwohl er vorhanden ist – nicht nachgewiesen werden kann. Sehr große Stichproben bewirken hingegen, dass relativ kleine Unterschiede zu signifikanten Ergebnissen führen. Um praktisch bedeutsame Effekte statistisch absichern zu können, benötigt man eine *Mindestanzahl bzw. eine optimale Anzahl an VPn*. Bei den Untersuchungen dieser Arbeit wird jede VPn nur einer Untersuchungsbedingung zufällig zugeordnet. Die entstehenden unabhängigen Stichproben erfordern in der Regel größere Stichproben als vergleichbare abhängige Stichproben. Um die optimale Anzahl notwendiger VPn zu bestimmen, wird nach dem Verfahren von Bortz und Döring (2003) vorgegangen oder mit Hilfe der Power-Analyse mit dem Softwarepaket GPOWER 3.0.10 nach Faul, Erdfelder, Lang und Buchner (2007) berechnet.

Die Forderung nach gleich großem α- (5%) und β- (.95) Fehler würde für die vorliegenden Studien nach eigenen Berechnungen mit GPOWER bedeuten, dass man insgesamt 112 VPn untersuchen muss, um einen großen Effekt aufzudecken. Bei einem mittleren Effekt sollten 280 VPn und bei einem kleinen Effekt gar 1720 VPn an der Studie teilnehmen. Die Grundgesamtheit der lizenzierten Kampfrichter im deutschsprachigen Raum würde schätzungsweise eine entsprechende Anzahl an Personen liefern. Allerdings ist es aus personellen, zeitlichen, finanziellen

und organisatorischen Gründen nicht möglich, eine derartige Vollerhebung im Rahmen einer Dissertation durchzuführen.

Wenn aufgrund inhaltlicher Überlegungen ein größeres β-Fehlerrisiko toleriert werden kann, was in den meisten sozialwissenschaftlichen Fragestellungen zutrifft, plädieren Bortz und Döring (2003, S. 603) dafür, ein α-β-Verhältnis von 1:4 zu wählen. Bei üblichem Fehler 1. Art von 5% würde ein β = 20% und eine Teststärke von 80% resultieren. Ein erhöhtes β-Fehlerrisiko bedeutet, dass aufgrund der Stichprobendaten die Nullhypothese als richtig bzw. die Alternativhypothese als falsch deklariert wird, obwohl die Alternativhypothese richtig ist. Bei einer Teststärke von 1 (und einem β von 0%) wird die Wahrscheinlichkeit, einen entsprechenden Unterschied in den Daten nachzuweisen, maximal. Für die aktuellen Untersuchungen wird ein Fairness-Kriterium von β/α = 2 festgelegt, womit das Verhältnis von α und β ausgedrückt wird. Dabei kann ein vorhandener großer Effekt mit einer Versuchspersonenzahl von 93, ein mittlerer mit 231 und ein kleiner mit 1422 VPn aufgedeckt werden. Da es sich bei ähnlichen bisherigen Untersuchungen zum Reihenfolge-Effekt (Greenless et al., 2009) bzw. zum Positions-Effekt (Plessner, 1999) im Sport eher um kleine bis mittlere Effekte handelt, müssen Stichprobenumfänge in genannter Größenordnung untersucht werden. Aus bereits genannten Gründen ist eine Stichprobengröße von ca. 230 lizenzierten Kampfrichtern nicht realisierbar.

Schlussfolgernd ist es möglich nur einen geringeren Anteil an VPn zu untersuchen. In diesem Fall wird auf eine Kompromiss-Analyse in GPOWER (Faul et al., 2007), als *dritten Schritt*, zurückgegriffen. Mit dieser Methode der α-Adjustierung wird ermittelt, welches α-Niveau angenommen werden soll, um bei einem schwachen, mittleren oder einem starken Effekt ein faires α-Niveau einzuhalten. Um sicher zu stellen, dass die Entscheidung über die Falsifizierung der Nullhypothese unter fairen Bedingungen getroffen wird, werden die Effektgröße und das α-β-Fehler-Verhältnis festgelegt und das anzunehmende α-Fehlerniveau berechnet. Das Signifikanzniveau α sagt aus, welche Irrtumswahrscheinlichkeit akzeptiert wird, dass bei einem positiven Ergebnis (der Effekt wird statistisch bestätigt) tatsächlich aber kein Effekt vorhanden ist. Resultierend kann festgehalten werden, dass aus Sicherheitsgründen ein höheres Signifikanzniveau gewählt wird. Je höher das gewünschte Signifikanzniveau, umso größer muss die untersuchte Stichprobe sein. Bei gegebenem Fairness-Kriterium β/α = 2 und einer geschätzten maximal großen Stichprobe von ca. 100 VPn auf 4 Gruppen aufgeteilt, ergibt sich bei einem kleinen Effekt ein α = 0,29. Die

dabei geschätzte Teststärke von (1-β) = 0,42 ist sehr gering. Ein mittlerer Effekt liefert ein α = 0,14 (1-β von 0,71) und ein großer Effekt ein α = 0,04 (1-β von 0,91).

Angaben aus der Literatur lassen einen kleinen bis mittleren Effekt vermuten (Greenless et al., 2009; Plessner, 1999). Für die Kampfrichteruntersuchung wird, aufgrund der berichteten Berechnungen, ein Signifikanzniveau von 10 % (α = 0,1) gewählt.

Da die geringe, nur sehr schwer steigerbare Versuchspersonenzahl und die Schätzung eines kleinen bis mittleren Effekts zu einer geringen Power führt, wird in einem *vierten Schritt* eine post hoc Analyse durchgeführt. Um nun den gefundenen Effekt mit dem in Schritt drei festgelegten α-Fehlerniveau aufdecken zu können, wird das erträgliche β-Fehlerniveau berechnet (8).

7.3 Datenbereinigung und -aufbereitung

Datenbereinigung

In Kapitel 6.2 wird erläutert, dass während der Testdurchführung vom Versuchsleiter ein Protokollbogen geführt wird, der über auftretende außerplanmäßige Vorkommnisse berichtet. Auf Grundlage der gemachten Anmerkungen in diesem Bogen und der Betrachtung der Bewertungsbögen werden einige VPn aus der Auswertung ausgeschlossen. Die Rohdatenbereinigung kann dabei unterschiedlich begründet werden.

(1) Ein erstes Kriterium schließt alle VPn aus, die sich nicht an die vorliegende Instruktion halten. Diese Nichteinhaltung der gestellten Aufgabe tritt sowohl bei der EoS- als auch bei der SbS-Bedingung auf. Es gibt Personen die entgegen der Instruktion sowohl ihre Abzüge mitschreiben bzw. auch nicht mitschreiben. Es gibt auch Einzelfälle, die erst beim zweiten Gerät die Instruktion befolgen. Das Erstexperiment besteht aus zwei Teilerhebungen, die sich durch die Bildqualität, nicht aber inhaltlich unterscheiden (6.1). Aus der Stichprobe der ersten Teil-Erhebung, mit der relativ gesehen schlechten Bildqualität wird eine VPn ausgeschlossen. In der zweiten Teil-Untersuchung, mit besserem Videodatenmaterial, werden drei Personen aufgrund einer fehlerhaften Bearbeitung der Aufgabe aussortiert.

(2) Ein zweites Kriterium, das zur Nichtberücksichtigung der Versuchspersonendaten führt, ist der Verlust des Überblickes über die Experimentalübungen. Bei der Durchführung des ersten Teil-Experimentes

klagt eine VPn am Ende seiner Wertungsaufgabe, dass er noch eine Übung zu werten hat, allerdings keine Videosequenz mehr zu sehen ist. Vermutlich hat die VPn zu schnell auf die ‚Weiter-Taste' gedrückt und damit eine Sequenz übersprungen.

(3) Auch technische Probleme sind ein Kriterium, aufgrund dessen die entsprechenden Daten aussortiert werden. Diese Probleme führen bei drei VPn des zweiten Teil-Experiments zur Unterbrechung bzw. Teil-Wiederholung des entsprechenden Videos. Da eine wiederholte Sichtung, die explizit in der Instruktion untersagt wird, zur veränderten Wahrnehmung bzw. Beurteilung führen kann, werden diese Daten aus der Auswertung genommen.

Abgeleitete Konsequenzen für das Zweitexperiment

Durch diese Erfahrungen bei der Durchführung des Erstexperiments werden Konsequenzen und damit Verbesserungen für das Zweitexperiment abgeleitet (6.3). Das erste Kriterium (1) ‚Nichteinhaltung der Instruktion' kann nicht in speziellem Maße verbessert werden. In den Untersuchungen wird auf eine standardisierte Durchführung Wert gelegt, sodass nur die nötigsten Informationen im Vorfeld des Experiments durch die Versuchsleiterin an die VPn weitergegeben werden. Dabei wird darauf hingewiesen, dass alle notwendigen Informationen einschließlich der Aufgabe in Text- und Videosequenzen präsentiert werden. Dem zweiten Kriterium (2) ‚Verlust des Überblicks' wird entgegengewirkt, indem eine Nummerierung der schwarzen Folien vorgenommen wird. Damit werden das Gerät und die Nummer der Übung vor der entsprechenden Videosequenz sichtbar. Dabei können die VPn die Angaben auf dem Bildschirm mit denen auf den Wertungsbögen vergleichen. (3) Technische Probleme können immer wieder Grund für eine Unterbrechung der Untersuchung sein. Eine Reduzierung der Dateigröße bei gleichbleibender Qualität soll diese Probleme minimieren.

Keines der beschriebenen Kriterien führt im Zweitexperiment zum Auftreten von Problemen und damit zum Ausschließen von Versuchspersonendaten. Weder die ‚Instruktionsmissachtung' noch die Problematik ‚Verlust des Überblicks' bestehen im Zweitexperiment. Technische Probleme sind im Rahmen der zweiten Untersuchung ebenso nicht beobachtet worden. Somit können alle erhobenen Datensätze zur Berechnung genutzt werden.

Datenaufbereitung – Umgang mit Ausreißern

Da einzelne grobe Datenausreißer die Ergebnisse stark beeinflussen können, wird im Rahmen der Datenaufbereitung, vor der entsprechenden Ergebnisdarstellung einer Untersuchung, überprüft, ob derartige Ausreißer vorhanden sind.

Grund für diese erste Annäherung an die Daten, ist die spezielle Art der Herangehensweise im Originalsetting. Im Gerätturnen sowie in anderen technisch-kompositorischen Sportarten wird zur Ermittlung der B-Note das gestutzte Mittel verwendet (2.3). Dabei fließt die niedrigste und die höchste Bewertung überhaupt nicht in die Endnote ein. Kommt es gar zu definierten, größeren Abweichungen zwischen den Wertungen, werden die entsprechenden Kampfrichter sanktioniert. Somit ist zu begründen, dass Noten die zu stark von der errechneten Mitte abweichen, also statistisch als Ausreißer oder entsprechend als Extremwerte auffällig werden, aus der Bewertung der speziellen Übung ausgeschlossen werden. Um diese auffälligen Bewertungen herauszufiltern, werden in der ersten Betrachtung der Daten keine arithmetischen Mittelwerte und das Dispersionsmaß der Standardabweichungen in Graphiken dargestellt. Stattdessen bietet sich die Darstellung mittels sogenannter Boxplots an. Sie beruhen auf Rangmaßzahlen und verzerren damit weit weniger, als Maße der zentralen Tendenz[22] (Sedlmeier, 1996). Die graphische Darstellung mittels Boxplot gestattet die Aufdeckung von Ausreißern aus der entsprechenden Stichprobe.

In einem weiteren, einmalig durchgeführten Schritt werden die Ausreißer jeder Übung in der Datendatei markiert und anhand eines ‚Missing Values' repräsentiert. Jede Übung wird nach Ausreißern hin getestet und die Werte daraufhin ausgeschlossen. Die weiteren Wertungen des entsprechenden VPn werden beibehalten. Anderweitige fehlende Werte gibt es in den Untersuchungsdaten nicht, sodass alle anfallenden fehlenden Daten die nicht berücksichtigten Wertungen darstellen.

Das gestutzte Mittel, also der Ausschluss einer bestimmten Anzahl an Wertungen, die ober- und unterhalb der Mitte liegen, wird nicht

[22] Der waagerechte Strich eines Boxplots stellt den Median der Stichprobe dar. Die Box zeigt, durch die jeweiligen Quartile begrenzt, etwa 50% der Werteverteilung an. Die Länge der Interquartilabstände ist unabhängig von Extremwerten. Die horizontalen Striche, über und unter der Box, geben den höchsten und den niedrigsten Wert an, nicht jedoch die extremen. Ausreißer werden in Form von Kreisen entsprechend oberhalb des 75%-Perzentils oder unterhalb des 25%-Perzentils angeben. Sie liegen anderthalb bis drei Boxhöhen außerhalb der Box. Werte die mehr als das dreifache der Boxhöhe von der Box entfernt sind, werden als Extremwerte gekennzeichnet und mit einem Sternchen markiert (Bühl & Zöfel, 2005, S. 681).

berechnet. Es ist nicht klar, ob die Werte an den Rändern vielleicht mit den untersuchten unabhängigen Variablen zusammenhängen und daher relevant für die ermittelten Ergebnisse sind.

7.4 Statistische Datenauswertung

Die statistische Datenauswertung erfolgt mit dem Programmpaket *SPSS für Windows (13.0).* Eine praxisorientierte Darstellung der gerechneten Verfahren gibt Bortz und Döring (2003). Die exakten mathematischen Grundlagen der entsprechenden Verfahren finden sich bei Bortz (2005).

Die Datenanalyse erfolgt zunächst mittels *deskriptiver Verfahren* (absolute und/oder relative Häufigkeit, Mittelwert, Standardabweichung). Um eine inferenzstatistische Überprüfung der Daten vornehmen zu können und parametrische Testverfahren zur Prüfung von Unterschiedshypothesen durchführen zu dürfen, müssen unterschiedliche Voraussetzungen berücksichtigt und erfüllt sein.

Eine erste Voraussetzung fordert *intervallskalierte* Daten. Diese Bedingung kann als gegeben betrachtet werden, da es sich bei den B-Abzügen um Daten handelt, die der Struktur nach einem Punktesystem gleichen und anhand dessen ermittelt und angegeben werden (2.3). Eine stetige Verteilung dieses Punktesystems kann angenommen werden.

Die zweite Voraussetzung bezieht sich auf die Verteilung der Daten in der Gesamtstichprobe. Zahlreiche statistische Verfahren setzen voraus, dass die zu untersuchenden Daten in der Grundgesamtheit normalverteilt sind. Die Prüfung auf Normalverteilung kann entweder graphisch oder mathematisch durchgeführt werden (Bühl & Zöfel 2005). Bei den vorliegenden intervallskalierten Daten erfolgt die Überprüfung auf Normalverteilung mittels des Kolmogorov-Smirnov-Tests. Bei kleinen Stichproben ($n < 30$) kann die Normalverteilung der Grundgesamtheit vorausgesetzt werden (Bortz, 2005, S. 149).

Die dritte Bedingung ‚*Unabhängigkeit der Stichproben*' gibt vor, dass die VPn zufällig zu den verschiedenen Bedingungen zugewiesen und die VPn nicht unter verschiedenen Treatmentstufen untersucht werden. Diese Bedingung ist in der Untersuchungsdurchführung (6.2) berücksichtigt worden und kann als gegeben betrachtet werden.

Die vierte Voraussetzung ist erfüllt, wenn es sich um *varianzhomogene Gruppen* handelt. Der Levene-Test überprüft, ob die Varianzen der Gruppen gleich sind und somit aufgedeckte Unterschiede zwischen den

Gruppen nur aufgrund der unabhängigen Variablen und nicht aufgrund von bestehenden, generellen Unterschieden zustande kommen.

Wenn diese Voraussetzungen nicht erfüllt sind, bieten sich verteilungsfreie, nicht-parametrische Verfahren an. Parametrische Verfahren sind recht robust gegen Verletzungen der Normalverteilung oder der Varianzhomogenität. Aufgrund der größeren Aussagekraft der Berechnungsmethoden wird auf parametrische Verfahren zurückgegriffen. Liegen schwerwiegende Verletzungen der Voraussetzungen vor, wird dies bei der Interpretation der Ergebnisse berücksichtigt. Bei zwei zu vergleichenden Gruppen wird der t-Test für unabhängige Stichproben verwendet (Bühl & Zöfel, 2005, S. 112).

Die bei der Varianzanalyse (VA) im Mittelpunkt stehende Frage lautet: Gibt es einen statistisch signifikanten Varianzanteil der abhängigen Variable, der allein durch die unabhängige(n) Variable(n) erklärt werden kann? Die parametrische Methode der VA hat gegenüber dem nicht-parametrischen t-Test zwei Vorteile. Zum einen führen bei 100 durchgeführten t-Tests fünf zu einer zufälligen Signifikanz[23]. Zum anderen sind mittels VA komplexere Vergleiche als die Paarvergleiche bei t-Tests möglich.

Aufgrund dieser Vorteile soll die Hauptfragestellung dieser Arbeit mit den Werkzeugen der zweifaktoriellen VA ohne Messwiederholung beantwortet werden. Mithilfe der univariaten mehrfaktoriellen VA können nicht nur die Effekte der einzelnen Faktoren (Haupteffekte), sondern auch gemeinsame Effekte (Interaktionseffekte) geprüft werden. Dabei ist die Prüfung von Hypothesen über spezifische Effekte einzelner Faktoren nicht immer ohne weiteres möglich. Sind signifikante Interaktionen vorhanden, ist die Interpretation der Haupteffekte problematisch.

Zur Annahme der jeweiligen Alternativhypothese wird bei allen Testverfahren, nach Fairnesskriterium, eine *Irrtumswahrscheinlichkeit* von α = 10% a priori festgelegt (7.2). Weiterhin wird als Effektgröße das partielle η^2 angegeben. Effektgrößen zwischen 0,001 und 0,058 werden als klein, zwischen 0,059 und 0,137 als mittel und größer als 0,138 als groß klassifiziert (Clark-Carter, 1997). Die Effektstärke stellt eine dimensionslose Zahl dar, die unabhängig von der Maßeinheit der Ursprungsdaten und der Stichprobengröße ist. Im Gegensatz zur klassischen Teststatistik ist das ein wichtiges Kriterium, da die erhobenen

[23] Die sogenannte α-Fehler-Inflation wird durch die Bonferoni-Korrektur etwas verbessert, indem die Wahrscheinlichkeit für ein signifikantes Ergebnis gesenkt wird.

Daten aus einer relativ geringen Population entstammen. Signifikanz und Effektstärke hängen insoweit zusammen, dass eine geringe Effektstärke eine größere Versuchsgruppe erfordert, damit das Ergebnis statistisch signifikant wird. Die Faktoren Stichprobengröße, Effektgröße und Alpha- und Betafehler hängen zusammen, so dass man bei einer fehlenden Größe, diese durch die anderen drei Faktoren berechnen kann. Das Effektgrößenmaß partielles η^2 bezieht sich auf die Quadratsumme des erklärbaren Faktors und der dazugehörigen Fehlerquadratsumme. Je mehr Faktoren vorhanden sind, desto geringer erweisen sich die Fehlerquadratsummen und somit ergibt sich eine höhere Effektgröße. Da das partielle η^2 somit vom versuchsplanerischen Design abhängig ist, sollten möglichst nur diejenigen Studien anhand dieses Maßes miteinander verglichen werden, die die gleichen Faktoren und die gleiche Anzahl an Stufen vorweisen.

Im Mittelpunkt des Interesses steht der Haupteffekt für den ersten Faktor Reihenfolge zwischen den Untersuchungsgruppen. Aufgrund vorheriger Studien zum Positions- bzw. Reihenfolge-Effekt im Sport, wird eher ein kleiner Effekt vermutet. Der Bewertungstyp sollte anhand theoretischer Vorhersagen keinen Haupteffekt zeigen. Weiterhin sollte auch kein Interaktionseffekt zu finden sein. Vermutet wird im Zusammenhang mit der Reihenfolge der Informationen, dass das zu bewertende Gerät einen Einfluss auf die Höhe des Effekts hat.

8 Ergebnisse

In den beiden organisierten Experimenten wird der Reihenfolge-Effekt bei schnellen Bewegungen am Beispiel von Kampfrichterurteilen untersucht. In diesem Zusammenhang wird ferner überprüft, ob die Art der Bewertung und das zu beurteilende Gerät unterschiedliche Bewertungen bei den Kampfrichtern hervorrufen.

8.1 Erstexperiment

8.1.1 Betrachtung der beiden Teil-Stichproben

Da der ermittelte Gesamtdatensatz des Erstexperiments aus zwei unabhängigen Stichproben zusammengestellt wurde (6.1), werden zunächst die beiden Personengruppen separat betrachtet.

Die erste Gruppe umfasst die Befragten, 49 Versuchspersonen (VPn) [24], die an der ersten Untersuchung im Dezember 2006 (Heidelberg) und im Januar 2007 (Karlsruhe) an einem Experiment mit relativ gesehen schlechter Bildqualität teilgenommen haben. Die zweite Gruppe umfasst 53[25] VPn, die an einer Untersuchung im Februar 2007 mit relativ gesehen guter Bildqualität teilnahmen. Der einzige Unterschied zwischen den beiden Studien stellt die Bildqualität der Videosequenzen dar.

Um zu klären, ob die gezeigten Experimentalübungen aufgrund der Bildqualität zu unterschiedlichen Wertungen führen und diese somit einen überzufälligen Einfluss hat, werden die Daten einem t-Test unterzogen. Angenommen wird, dass zwischen den Mittelwerten der Teil-Untersuchungs-Stichproben kein Unterschied besteht.

Vor dieser Prüfung werden die einzelnen Experimentalübungen auf Normalverteilung geprüft. Da sich je Untersuchungsbedingung eine Stichprobengröße ergibt, die kleiner als 30 ist, kann die Normalverteilungsannahme als gegeben angesehen werden (Bortz, 2005, S. 149). Der Levene-Test auf Varianzhomogenität weißt bei einer Übung – zweite Ringe-Übung ($p = 0{,}032$) – einen signifikanten Wert auf. Wenn die Stichprobegröße der einzelnen Zellen annähernd gleich verteilt ist,

[24] Ursprünglich haben 51 VPn an dieser Teiluntersuchung teilgenommen. Zwei VPn wurden aber aus der Datenauswertung ausgeschlossen (vgl. 6.1).

[25] An der Studie haben sich insgesamt 59 VPn beteiligt, wobei sechs VPn im Rahmen der Datenbereinigung und -aufbereitung aussortiert wurden (6.1).

erweist sich eine etwaige Inhomogenität als irrelevant[26]. Der t-Test für unabhängige Stichproben ergibt ein nicht signifikantes Ergebnis.

Dadurch wird angenommen, dass sich die Teil-Stichproben des Erstexperiments nicht überzufällig voneinander unterscheiden. Für die weiteren Berechnungen werden daher alle VPn, die am Erstexperiment teilgenommen haben, als eine Stichprobe angesehen.

8.1.2 Deskriptive Datenauswertung

Wie in Kapitel 7.3 erläutert, erfolgt zunächst die Aufdeckung von *Ausreißerwerten*. Diese können sich stark auf die Ergebnisse niederschlagen und damit zu Verzerrungen führen. Um extreme Werte in den Daten sichtbar zu machen, werden die einzelnen Übungen, getrennt nach den vier Untersuchungsbedingungen, denen die VPn zufällig zugeordnet sind, separat mittels Boxplot-Darstellung untersucht. Alle in diesem Schritt ermittelten Ausreißer werden darauffolgend aus den Daten ausgeschlossen und als fehlende Werte dargestellt.

Insgesamt werden 19 Einzelwertungen aus den weiteren Bewertungen ausgeschlossen. Am langsamen Gerät Ringe sind neun Wertungen (sechs Wertungen fallen auf die Einwertübung und je eine auf die übrigen) und am schnellen Gerät Reck zehn Wertungen (eine Wertung bei Übung zwei, ansonsten je drei Wertungen) vom Ausschluss betroffen.

Weitere Berechnungen werden auf der neu gewonnenen Datenbasis durchgeführt. In den unterschiedlichen Untersuchungsbedingungen zeigen sich die einzelnen Übungen, trotz kleiner Stichprobengröße und anzunehmender Normalverteilung, nach dem Kolmogorov-Smirnov-Test normalverteilt. Vergleicht man die Histogramm-Darstellungen mit den eingeblendeten Normalverteilungskurven können die Übungen als annähernd normalverteilt bezeichnet werden.

Tabelle 8 und 9 enthalten die Benotungen der Einwert- (E), Kontroll- (K) und Experimentalübungen (V oder H) an den Geräten Reck und Ringe[27], jeweils nach den Untersuchungsbedingungen getrennt aufgeführt. Dabei sind die reinen Abzüge in der B-Note und nicht die errechnete B-Note dargestellt. Die Untersuchungsbedingungen unterscheiden sich

---

[26] Das gilt, wenn das Verhältnis zwischen der größten und der kleinsten Zellen-Stichprobengröße kleiner als 1,5 ist (Stevens, 1999, p. 75).

[27] Die Präsentation der Übungen erfolgte in dieser Geräte-Reihenfolge. Die Ergebnisdarstellung wird aufgrund besserer Vergleichsmöglichkeit mit dem Zweitexperiment in umgekehrter, olympischer Geräte-Reihenfolge präsentiert.

in zwei Faktoren, der Reihenfolge der Informationspräsentation und der Art der Bewertung.

Faktor eins, die Präsentations-Reihenfolge der Informationen, variiert je nach Übung und besteht in zwei Ausführungen, positiv-negativ (entspricht H – Fehler hinten) oder negativ-positiv (entspricht V – Fehler vorne) (5.2.2). Untersuchungsbedingung eins und zwei enthalten unterschiedliche Videosequenzen, die teilweise den Fehler vorne und teilweise den Fehler hinten in der Übung zeigen. Die Videosequenzen der beiden anderen Untersuchungsbedingungen (drei und vier) enthalten Videosequenzen, die bis auf die Position des Fehlers identisch sind. Faktor zwei, der Bewertungstyp, kommt dadurch zum Ausdruck, dass die Untersuchungsbedingung eins und drei nach dem EoS-Prozess bewertet werden und die VPn ihre Abzüge nicht während der Videopräsentation mitschreiben dürfen. In Untersuchungsbedingung zwei und vier haben sie der SbS-Aufgabe zu folgen und müssen alle wahrgenommenen Abzüge simultan mitschreiben.

Tabelle 8: Deskriptive Statistik der Erstexperiment-Übungen (Untersuchungsbedingung eins und zwei)

EoS							SbS						
UB 1 - blau	**N**	**M**	**SD**	**Min**	**Max**	**Range**	**UB 2 - gelb**	**N**	**M**	**SD**	**Min**	**Max**	**Range**
Ringe 1, E	29	1,73	**0,52**	1,00	3,00	1,6	Ringe 1, E	19	1,77	**0,59**	0,80	3,20	0,9
Ringe 2, H	30	**1,20**	0,41	0,50	2,30	2,0	Ringe 2, H	20	**1,03**	0,30	0,50	1,40	1,0
Ringe 3, K	29	0,85	0,26	0,40	1,40	1,3	Ringe 3, K	20	0,80	0,29	0,40	1,50	1,4
Ringe 4, V	29	0,79	0,34	0,30	1,60	1,6	Ringe 4, V	20	0,64	0,27	0,30	1,20	1,5
Reck 1, E	29	**1,10**	0,37	0,40	2,00	**2,0**	Reck 1, E	20	**0,86**	0,30	0,50	1,40	**2,4**
Reck 2, V	30	**1,73**	**0,57**	0,80	2,80	1,8	Reck 2, V	20	**1,32**	0,35	0,80	1,80	0,9
Reck 3, K	29	1,82	0,36	1,20	2,50	1,0	Reck 3, K	19	1,72	0,39	1,20	2,60	1,1
Reck 4, H	30	1,44	0,44	0,50	2,10	1,3	Reck 4, H	19	**1,34**	0,42	0,80	2,30	0,9

Tabelle 9: Deskriptive Statistik der Erstexperiment-Übungen (Untersuchungsbedingung drei und vier)

EoS							SbS						
UB 3 - grün	**N**	**M**	**SD**	**Min**	**Max**	**Range**	**UB 4 - rot**	**N**	**M**	**SD**	**Min**	**Max**	**Range**
Ringe 1, E	21	1,68	0,30	1,10	2,40	1,4	Ringe 1, E	27	1,63	**0,53**	0,80	3,00	1,4
Ringe 2, V	24	1,18	0,40	0,30	2,00	1,5	Ringe 2, V	27	1,19	0,45	0,40	2,10	1,4
Ringe 3, K	25	0,92	0,29	0,50	1,60	1,7	Ringe 3, K	27	0,82	0,29	0,30	1,40	1,6
Ringe 4, H	25	**0,83**	0,31	0,40	1,60	1,3	Ringe 4, H	27	**0,66**	0,26	0,30	1,30	1,3
Reck 1, E	23	0,90	0,36	0,30	1,70	1,3	Reck 1, E	27	1,01	0,37	0,40	1,80	**2,2**
Reck 2, H	24	**1,52**	0,39	0,80	2,30	1,7	Reck 2, H	27	1,47	0,35	0,80	2,20	1,7
Reck 3, K	25	1,92	0,47	1,20	2,90	1,1	Reck 3, K	26	1,87	0,38	1,10	2,70	1,1
Reck 4, V	25	**1,35**	0,41	0,70	2,00	1,2	Reck 4, V	25	**1,53**	0,35	0,90	2,20	1,0

Bei der Betrachtung der Mittelwerte muss in diesem speziellen Untersuchungsfeld angemerkt werden, dass die absoluten Werte in Form der vorgenommenen Abzüge, wie sie in Tabelle 8 und 9 aufgeführt sind, in bestimmter Art und Weise interpretiert werden müssen. Da die gezeigten Experimentalübungen unterschiedlich hohe Absolutbewertungen aufgrund unterschiedlicher Schwierigkeitsgrade aufweisen, stellt sich der Vergleich verschiedener Übungen über die absoluten Höhen der Abzüge als nicht sinnvoll dar. Nur innerhalb verschiedener Untersuchungsbedingungen einer Übung sollten daher Wertungen direkt verglichen werden. Für die Darstellung und Interpretation der Untersuchungsergebnisse sind die Differenzen zwischen den Mittelwerten der Abzüge in den einzelnen Untersuchungsbedingungen und die Streuung der Werte entscheidend. Aufgrund dessen erfolgt eine geräteweise Auswertung des Datenmaterials.

Auffällige Werte

Bei der Betrachtung der einzelnen Übungen über die Geräte und Untersuchungsbedingungen hinweg werden Mittelwertunterschiede größer oder gleich 0,17 Punkten[28], eine Standardabweichung größer gleich 0,49 Punkten[29] und eine Spannweite (Range) von 1,9[30] und mehr

[28] Der Richtwert entspricht 13% des Mittelwerts aller Übungen (aller Untersuchungsbedingungen).

[29] Der Richtwert entspricht 130% der Mittelwerte aller Standardabweichungen.

Punkten innerhalb einer Untersuchungsbedingung als auffällig bezeichnet und in Tabelle 8 und 9 hervorgehoben.

Die Sichtprüfung zeigt, dass die erste Ringe-Übung in Untersuchungsbedingung eins (SD = 0,52), zwei (SD = 0,59) und vier (SD = 0,53) im Vergleich zu Untersuchungsbedingung drei (SD = 0,30) erhöhte Standardabweichungen aufweist. Da es sich hierbei um die Einwertübung handelt, war dieses Ergebnis erwartbar. Die Einwertübung am zweiten Gerät Reck hingegen zeigt keine auffälligen Streuungen. Reck-Übung eins zeigt sich auffällig durch den Mittelwertunterschied der ersten beiden Untersuchungsbedingungen von 0,24 Punkten und mit erhöhten Spannweiten mit Werten zwischen 2,0 und 2,4 Punkten in allen Untersuchungsbedingungen, bis auf Untersuchungsbedingung drei (Range = 1,3). Die Einwertübungen werden nicht mit den Verfahren der Inferenzstatistik geprüft (8.1.3), da sie lediglich dem Einwerten der Kampfrichter dienen.

Die zweite Übung an den Ringen, zeigt in Untersuchungsbedingung eins (M = 1,2) und zwei (M = 1,03) mit dem Fehler in der zweiten Hälfte der Übung einen Mittelwertunterschied von 0,17 Punkten. Die Übungen in Untersuchungsbedingung drei und vier (Fehler vorne) werden mit 0,01 Punkten Unterschied, und damit mit einem annähernd identischen Abzug bewertet. Die Übungen in Untersuchungsbedingung eins und zwei, wie Untersuchungsbedingung drei und vier, unterscheiden sich lediglich im Bewertungstyp voneinander.

In Ringe-Übung vier lässt sich in Untersuchungsbedingung drei (M = 0,83) und vier (M = 0,66), mit dem Fehler hinten, ein Mittelwertunterschied von 0,17 Punkten erkennen.

Die zweite Übung am Reck weist einen auffälligen Mittelwertunterschied von 0,41 Punkten zwischen Untersuchungsbedingung eins (M = 1,73; SD = 0,57) und zwei (M = 1,32), Fehler vorne sowie ein Streuungsmaß von SD = 0,57 Punkten in Untersuchungsbedingung eins auf. Weiterhin unterscheiden sich die Mittelwerte von Untersuchungsbedingung eins und drei (M = 1,52) um 0,21 Punkten. Wenn dieses Ergebnis auch statistisch nachgewiesen wird spricht das für einen Reihenfolge-Effekt (8.1.3). Da die Übung mit dem Fehler vorne mit mehr Abzügen bestraft wird, als die Übung mit dem Fehler hinten, würde das bedeuten, dass ein Primacy-Effekt besteht.

[30] Für diese Festlegung wurden 130% des Mittelwerts aller Spannweiten berechnet.

Die letzte Übung mit dem Fehler vorne (Reck, vier) wird von den VPn der Untersuchungsbedingung drei (M = 1,35) und vier (M = 1,53) um durchschnittlich 0,18 Punkte unterschiedlich bewertet. Ein weiterer Unterschied von 0,19 Punkten ergibt sich zwischen Untersuchungsbedingung zwei (M = 1,34) und vier. Somit geben Kampfrichter, die mitschreiben dürfen, unterschiedliche Wertungen für eine Übung ab, die sich nur durch die Position des Fehlers unterscheidet. Auch bei dieser Übung wird die Übung mit dem Fehler vorne mit höheren Abzügen bewertet und damit ein Primacy-Effekt angedeutet.

Kontrollübungen

Die Kontrollübung (K) am jeweiligen Gerät wird durch eine Übung repräsentiert, die für alle Untersuchungsbedingungen identisch ist und keine Besonderheiten in Form von groben Fehlern oder Stürzen beinhaltet. Die Übungen unterscheiden sich nicht im ersten Faktor Reihenfolge, sondern lediglich im Bewertungstyp.

Bei der speziellen Betrachtung der Kontrollübungen fällt auf, dass die über die Untersuchungsbedingungen gemittelte Standardabweichung der Ringe-Kontrollübung (Übung drei) mit einem Wert von SD = 0,28 geringer ausfällt als die der dritten Reck-Übung (SD = 0,4). Im Vergleich zu den Experimentalübungen, mit einem Wert von SD = 0,34 an den Ringen und von SD = 0,41 am Reck, ergibt sich eine vergleichbare Größenordnung und auch die Tendenz zur größeren Streuung der Wertungen am Reck ist dieselbe.

Zusammenfassend lässt sich festhalten, dass das Datenmaterial auf den ersten Blick keine klar erkennbaren Trends aufweist. Ein Reihenfolge-Effekt, und damit ein Unterschied zwischen den Videosequenzen mit dem Fehler vorne und denen mit dem Fehler hinten, zeigt sich nach visueller Prüfung nur bei den beiden Reck-Übungen, nicht am Gerät Ringe. Die Art der Bewertung, also Unterschiede zwischen Untersuchungsbedingung eins und zwei, bzw. drei und vier, scheint vermehrt Unterschiede in den Mittelwerten der Übungen hervorzubringen und von Bedeutung zu sein. Eine Besonderheit scheint es hierbei zu geben: Lediglich die Übungen mit dem Fehler hinten, aber nur am Gerät Ringe, zeigen einen Unterschied im zweiten Faktor Bewertungstyp. Am Reck ergibt sich ein entgegengesetztes Bild, da nur Übungen mit dem Fehler vorne, Mittelwertunterschiede von 0,17 Punkten und mehr zeigen. In diesem Zusammenhang ergibt sich bei allen Experimentalübungen, zwei Ringe- und zwei Reck-Übungen, ein (nach obiger Definition)

auffälliger Unterschied. Ob die Unterschiede nach inferenzstatistischer Prüfung überzufällig sind, zeigt das folgende Kapitel.

8.1.3 Inferenzstatistische Auswertung

Die Auswertung erfolgt für jede Turnübung separat. Dabei wird die Einwertübung (E) nicht ausgewertet, da sie lediglich dazu dient, das präsentierte Leistungsniveau einmal vorzuführen, damit die Kampfrichter wissen, was sie in der Untersuchung erwartet. Die Übungen können nicht zusammengenommen werden, da sie unterschiedlich hohe Absolutbewertungen aufweisen, und es somit zu mehrgipfligen Verteilungskurven kommt. Diese Art der Herangehensweise wirkt sich bei der Standardabweichung und den damit zusammenhängenden Varianzen innerhalb einer Experimentalgruppe, im Vergleich zu denen unterschiedlicher Gruppen, vergleichbar aus. Vermutlich würde sich eine nicht realistische Abbildung der Datenlage ergeben, so dass beispielsweise große Varianzen innerhalb einer Gruppe und vergleichsweise kleine Varianzen zwischen den Gruppen die Folge wären. Daher werden die Übungen einzeln ausgewertet und dargestellt.

Für die Benotung der Kontroll- und Experimentalübungen werden geräteweise, univariate zweifaktorielle VA (ANOVA) für unabhängige Stichproben mit den Faktoren Reihenfolge und Bewertungstyp gerechnet.

Die a posteriori Teststärkenanalyse hat im Erstexperiment eine Teststärke (1-β) von 0,194 ermittelt, wenn man von einem kleinen Effekt, einer Stichprobengröße von N = 102, einem Freiheitsgrad von 3 und einem α-Fehlerniveau von 10% ausgeht. Das β-Fehlerniveau von 0,8 und ein kritischer F-Wert von 2,141 können somit angenommen werden.

1. Hypothese: Reihenfolge-Effekt

Erwartet wird, dass die Videosequenzen mit dem Fehler vorne (V) in der Übung unterschiedlicher bewertet werden als Übungen mit dem Fehler hinten (H) und somit ein Reihenfolge-Effekt aufgedeckt wird (μV $\neq$ μH) (7.1). Somit sollte es zu einem Haupteffekt des Faktors Reihenfolge kommen. Die Kontrollübungen sollten den Erwartungen zufolge keinen Reihenfolge-Effekt zeigen, da auch keine Manipulation dieses Faktors stattgefunden hat.

Für den Faktor Reihenfolge kann in keiner Übung ein Haupteffekt aufgedeckt werden, sondern nur in Form einer Interaktion mit dem zweiten Faktor Bewertungstyp. Der Faktor Reihenfolge scheint keinen überzufälligen Einfluss auf die vorgenommenen Abzüge an den Experimental-

übungen zu haben. Die Mittelwertunterschiede zeigen sich uneinheitlich, allerdings nicht signifikant (vgl. Ergebnisse zu Hypothese 3).

Betrachtet man die Effektstärken in Form der partiellen Eta-Quadrate für den Faktor Reihenfolge, wird kein oder nur ein geringer Effekt geschätzt (Werte zwischen η^2 = 0,001 und 0,024). Die Wahrscheinlichkeit, diesen Effekt mit der vorhandenen Anzahl an VPn zu finden, falls er wirklich existiert, beläuft sich zwischen 6% und 32,8%. Die absoluten Unterschiede der Abzüge, berechnet aus der Differenz der gemittelten Wertungen der EoS- und der SbS-Bedingung einer Übung mit dem Fehler vorne und mit dem Fehler hinten, belaufen sich auf bis zu 0,113 Punkte am Reck und bis zu 0,098 Punkte an den Ringen.

Die erste Forschungshypothese kann somit nach den Ergebnissen der ersten Untersuchung nicht bestätigt und damit nicht angenommen werden. Die Bewertungen der Kampfrichter unterscheiden sich nicht in überzufälligem Ausmaß. Somit kann die Nullhypothese, es bestehen keine Unterschiede bezüglich der Reihenfolge, beibehalten werden.

2. Hypothese: Effekt des Bewertungstyps

Angenommen wird, dass der Reihenfolge-Effekt, entsprechend der ersten Hypothese, unabhängig von der Art der Bewertung entsteht. Es existiert somit kein Unterschied in den Wertungen, ob ein Kampfrichter die Abzüge einer Übung mitschreibt oder nicht. Erwartet wird, dass kein Haupteffekt des Faktors Bewertungstyp sichtbar wird.

Die Experimentalübungen Reck, Übung zwei und Ringe, Übung vier zeigen einen Haupteffekt auf dem Faktor Bewertungstyp. Im ersten Fall wird dieser neben einem Interaktionseffekt ‚Reihenfolge x Bewertungstyp' sichtbar (vgl. Ergebnisse zu Hypothese 3). Im zweiten Fall, Ringe-Übung vier, zeigt sich ein signifikanter Haupteffekt auf dem Faktor Bewertungstyp ($F(0{,}740) = 7{,}282$; $p = 0.008$, $\eta^2 = 0{,}07$). Für diesen Haupteffekt schätzt das η^2 einen mittleren Effekt. Bei beiden Faktoren unterscheiden sich die Mittelwerte in beiden Stufen. In den Interaktionsdiagrammen lassen sich dieselben Trends erkennen. Der Mittelwert der Abzüge von der SbS-Bedingung (μ = 0,65) liegt konsistent unter dem der EoS-Bedingung (μ = 0,81) (Tabelle 10). Das macht einen durchschnittlichen Unterschied der Bewertungstypen von etwa 0,16 Punkten.

Tabelle 10: Abzüge für die Experimentalübung vier an den Ringen

Reihenfolge	Bewertungstyp	M	SD	N
Video 1 (V)	EoS	,793	,338	29
	SbS	,640	,270	20
	Total	,731	,318	49
Video 2 (H)	EoS	,828	,306	25
	SbS	,659	,256	27
	Total	,740	,291	52
Total	EoS	,809	,321	54
	SbS	,651	,260	47
	Total	,736	,303	101

Somit deutet sich an, dass sowohl bei dem Video mit dem Fehler vorne als auch bei dem mit dem Fehler hinten das Notieren der Abzüge zur besseren Wertung führt. Sind die Linien der Interaktionsdiagramme absolut parallel, kann man von keinerlei Interaktion zwischen den Faktoren sprechen. Im vorliegenden Fall könnte eine Interaktion vorhanden sein, die allerdings kaum praktische Relevanz hat und damit nicht interpretiert wird.

Die Effektstärken des Faktors Bewertungstyp belaufen sich auf ein partielles Eta-Quadrat mit Werten zwischen η^2 = 0,003 und 0,067. Die Teststärke ergibt Werte zwischen 5% und 74,5%. Die absoluten Unterschiede der Abzüge, berechnet aus der Differenz der durchschnittlichen EoS-Wertungen und der SbS-Wertungen, weisen Werte bis zu 0,065 Punkte am Reck und bis zu 0,158 Punkte an den Ringen auf.

Mit Blick auf Forschungshypothese zwei darf die Alternativhypothese angenommen werden, da sich die Mittelwertunterschiede der Wertungen der beiden Bewertungstypen unterscheiden. Damit wird die Nullhypothese, es besteht kein Unterschied zwischen den Wertungen die mittels des EoS-Prozess und denen mittels des SbS-Prozess gefällt werden, nicht beibehalten. Der Bewertungstyp erweist sich im Erstexperiment als der Faktor, der teilweise überzufällige Ergebnisse liefert und daher in der Urteilspraxis beachtet werden sollte.

3. Hypothese: Interaktionseffekt ‚Reihenfolge x Bewertungstyp'

Vermutet wird, dass sich die beiden Faktoren Reihenfolge und Bewertungstyp nicht gegenseitig bedingen und es so nicht beispielsweise zu einer höheren Wertung kommt, wenn der Kampfrichter nicht

mitschreiben darf und der Fehler vorne ist. Somit lässt sich kein Interaktionseffekt erwarten.

Das Datenmaterial zeigt nicht erwartungskonforme Interaktionen bei beiden Experimentalübungen des schnellen Geräts *Reck*. Die *Übung 2*, zeigt nach statistischer Überprüfung einen signifikanten Interaktionseffekt ($F(3,670) = 4{,}462$, $p = 0.037$, $\eta^2 = 0{,}044$). Zugleich weist der Faktor Bewertungstyp einen Haupteffekt auf ($F(3,670) = 6{,}991$, $p = 0.01$, $\eta^2 = 0{,}067$). Für den Interaktionseffekt schätzt das η^2 einen kleinen Effekt und für den Haupteffekt einen mittleren Effekt.

Die Betrachtung der Mittelwerte zeigt, dass sich die Kampfrichter der unterschiedlichen Bewertungstypen, EoS und SbS, bei Video eins mit dem Fehler vorne (V) in ihren Abzügen uneiniger sind als bei Video zwei mit dem Fehler hinten (H). Sowohl bei Video eins (V) als auch bei Video zwei (H) führt die SbS-Bedingung (Abzüge notieren) stets zu geringeren Abzügen ($\mu = 1{,}41$) als in der EoS-Bedingung ($\mu = 1{,}64$). Der absolute Unterschied der Abzüge beläuft sich auf einen Wert von 0,23 Punkten. Das arithmetische Mittel, die Standardabweichung und die Stichprobengröße der beiden Videoversionen der zweiten Reck-Übung lässt sich Tabelle 11 entnehmen.

Tabelle 11: Abzüge für die Experimentalübung zwei am Reck

Reihenfolge	Bewertungstyp	M	SD	N
Video 1 (V)	EoS	1,73	,571	30
	SbS	1,32	,353	20
	Total	1,57	,533	50
Video 2 (H)	EoS	1,52	,392	24
	SbS	1,47	,349	27
	Total	1,50	,367	51
Total	EoS	**1,64**	,506	54
	SbS	**1,41**	,356	47
	Total	1,53	,456	101

Graphisch veranschaulicht werden die beiden Interaktionsdiagramme dieser Übung in Abbildung 8. Durch die vorliegende hybride Interaktion darf der Haupteffekt global interpretiert werden (Bortz & Döring, 2003, S. 535).

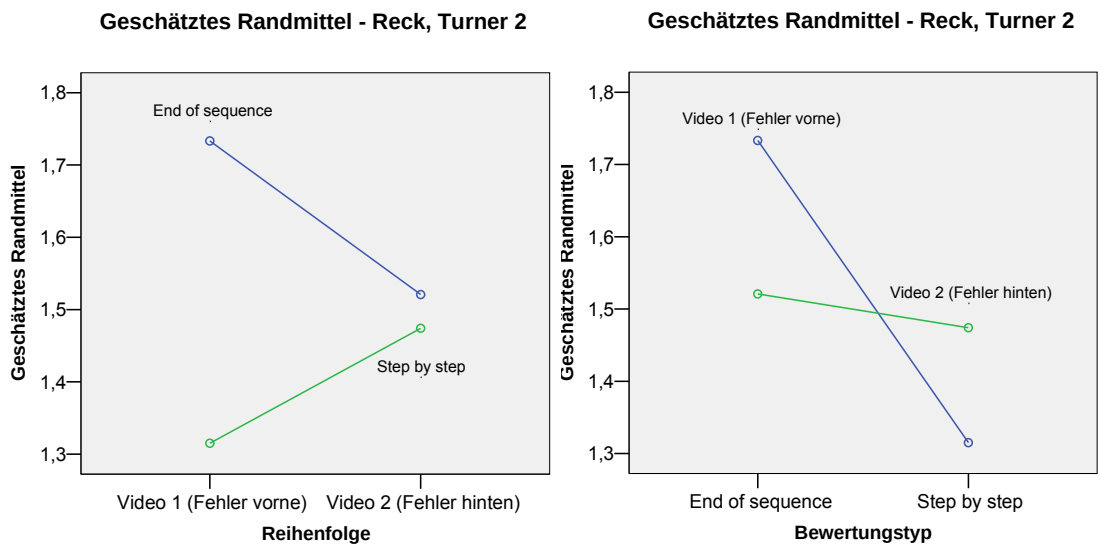


Abbildung 8: Interaktionsdiagramme der zweiten Übung am Reck (Reihenfolge - links und Bewertungstyp - rechts)

Ein Haupteffekt Reihenfolge liegt hier nicht vor. Es ist kein klarer Mittelwert-Unterschied zwischen Video eins (V) und Video zwei (H) erkennbar. Die beiden Verbindungslinien im Interaktionsdiagramm des Faktors Reihenfolge zeigen einen gegenläufigen Verlauf.

Die *Übung vier am Reck* weist eine nach Fairnesskriterium (7.2) bedeutende Interaktion auf ($F(0{,}653) = 2{,}789$; $p = 0.098$, $\eta^2 = 0{,}029$). Für den Interaktionseffekt schätzt das η^2 einen kleinen Effekt. Es bestehen keinerlei Haupteffekte, die aufgrund der bestehenden disordinalen Interaktion auch nicht interpretiert werden dürften (ebenda, S. 535).

Es unterscheiden sich weder die Mittelwerte der Abzüge des Bewertungstyps EoS ($\mu = 1{,}40$) und SbS ($\mu = 1{,}45$) signifikant voneinander, noch die von Video eins (H) ($\mu = 1{,}40$) und Video zwei (V) ($\mu = 1{,}44$). Die Zuordnung der VPn zu den Untersuchungsbedingungen bestimmt alleinig die Auswirkungen der Abzüge. Dabei zeigen die VPn, die der SbS-Bedingung zugeordnet sind und den Fehler vorne sehen, die größten vorgenommen Abzüge. Vergleichsweise geringe Abzüge nehmen die Kampfrichter vor, die ebenfalls Notizen machen können, aber den Fehler in der zweiten Hälfte zu sehen bekommen. Abzüge in vergleichbarer, geringer Höhe nehmen die Kampfrichter der EoS-Bedingung vor, die den Fehler in der ersten Hälfte präsentiert bekommen (Tabelle 12).

Tabelle 12: Abzüge für die Experimentalübung vier am Reck

Reihenfolge	Bewertungstyp	M	SD	N
Video 1 (H)	EoS	1,44	,440	30
	SbS	1,34	,422	19
	Total	1,40	,431	49
Video 2 (V)	EoS	1,35	,410	25
	SbS	1,53	,353	25
	Total	1,44	,390	50
Total	EoS	1,40	,425	55
	SbS	1,45	,391	44
	Total	1,42	,409	99

Die differenzierte Betrachtung der einzelnen Zellenmittelwerte lässt folgende Darstellung der Ergebnisse zu: Wird ein Fehler in der ersten Übungshälfte präsentiert, bewerten Kampfrichter die Übung mit höheren Abzügen und damit schlechter, wenn sie dazu angehalten werden, die Ausführungsfehler zu notieren (μ = 1,53). Wenn sie keine Abzüge notieren, bewerten sie die identische Übung besser (μ = 1,35). Bei einem Fehler in der zweiten Übungshälfte ergibt sich das gegenteilige Bild. Kampfrichter, die dem EoS-Prozess folgen, ziehen mehr Fehler ab (μ = 1,44), als diejenigen, die Notizen machen (μ = 1,34).

Die dritte Forschungshypothese, es besteht eine Interaktion der beiden Faktoren, darf ebenso angenommen werden. Die Nullhypothese muss verworfen werden, da am Reck beide Experimentalvideos Interaktionseffekte ‚Reihenfolge x Bewertungstyp' aufweisen.

4. Hypothese: Geräteeffekt

Der Reihenfolge-Effekt ist an schnellen Geräten stärker ausgeprägt als an langsamen Geräten. Am schnellen Gerät Reck kommt es zu zwei Interaktionseffekten, während es am langsamen Gerät zu einem Haupteffekt Bewertungstyp kommt. Die Reihenfolge der präsentierten Informationen scheint keinen eindeutigen Einfluss zu haben.

Auf Grund dessen darf die Forschungshypothese vier nicht angenommen werden und die Nullhypothese ist beizubehalten.

Tabelle 13 bildet die Ergebnisse des Erstexperiments ab. Geräteweise werden die überzufälligen, signifikanten oder nach Fairnesskriterium auffälligen Unterschiede sowie die anhand des partiellen η^2-Wertes geschätzten Effekte dargestellt. Die letzte Spalte gibt die Reihenfolge-

Trends der Wertungen wider und verbildlicht, unabhängig von der Höhe der Abzüge in der B-Note, ob die Übung mit dem Fehler vorne oder die mit dem Fehler hinten zu einem größeren Abzug geführt hat.

Tabelle 13: Erstexperiment – Überblick zum inferenzstatistischen Vergleich der Experimentalübungen (Einwert- und Kontrollübungen nicht aufgeführt; HE – Haupteffekt)

Gerät, Übung	Effekt (p)	η^2	Reihenfolge-Trend	
			EoS	SbS
Ringe, 2 (V2–V)	-------	------	Recency	Primacy
Ringe, 4 (V1-V)	HE Bewertungstyp (0.008)	0.07	Recency	
Reck, 2 (V1–V)	Interaktion (0.037)	0.044	**Primacy**	Recency
	HE Bewertungstyp (0.01)	0.067		
Reck, 4 (V2–V)	Interaktion (0.098)	0.029	Recency	**Primacy**

Nehmen die Kampfrichter für eine Übung mit dem Fehler vorne größere Abzüge vor als bei der Übung mit dem Fehler hinten, zeigt sich eine Primacy-Effekt-Tendenz. Im Gegensatz dazu lässt sich eine Recency-Effekt-Tendenz erkennen, wenn die Kampfrichter für die Übung mit dem Fehler hinten mehr Abzüge vornehmen. Diese Tendenzen werden nach den beiden Bewertungstypen EoS und SbS getrennt aufgeführt.

Bei der Betrachtung der Reihenfolge-Trends zeigt sich keine klar erkennbare Tendenz. Die Ringe-Übung zwei deutet einen Recency-Effekt beim EoS-Prozess an, während der SbS-Prozess einen Primacy-Effekt erkennen lässt. Ringe-Übung vier hingegen zeigt unabhängig vom Bewertungstyp einen Recency-Effekt. Die beiden Reck-Übungen lassen wiederum unterschiedliche Tendenzen erkennen, wobei der Primacy-Effekt durch größere Mittelwertunterschiede gestützt wird.

8.2 Zweitexperiment

8.2.1 Deskriptive Datenauswertung

Wie bereits in den Ergebnissen des Erstexperiments durchgeführt, erfolgt zunächst die Sichtung auf eventuell auftretende *Ausreißerwerte* (7.3). Durch die mögliche Verzerrung der Ergebnisse durch Werte, die stark von den anderen Werten einer Experimentalgruppe abweichen, werden diese mittels Boxplot-Darstellung aufgedeckt. Falls Ausreißerwerte in den einzelnen Untersuchungsbedingungen vorhanden sind, werden sie künftig nicht berücksichtigt, sondern tauchen als fehlende

Werte in den Ergebnisdarstellungen auf. Im Zweitexperiment werden 9 Wertungen als Ausreißer entlarvt, wobei vier am langsamen Gerät Ringe und die restlichen fünf am Reck auffällig werden. Drei dieser extremen Werte zeigen sich bei den Einwertübungen, eine bei der Kontrollübung am Reck und die übrigen verteilen sich auf die Experimentalübungen beider Geräte. Nur die zweite Übung an den Ringen weißt gleich zwei Extremwerte auf.

Diese Daten bilden die Basis für folgende Berechnungen. In den unterschiedlichen Untersuchungsbedingungen zeigen sich die Übungen, trotz kleiner Stichprobengröße und damit anzunehmender Normalverteilung, nach dem Kolmogorov-Smirnov-Test normalverteilt. Auch beim Vergleich der Histogramme mit den eingeblendeten Normalverteilungskurven kann man von annähernd normalverteilten Werten ausgehen.

Die Tabelle 14 und 15 zeigen die Wertungen der Einwert-, Kontroll- und Experimentalübungen an den untersuchten Geräten, Ringe und Reck, für jede Untersuchungsbedingung separat aufgeführt.

Tabelle 14: Deskriptive Statistik der Zweitexperiment-Übungen (Untersuchungsbedingung eins und zwei)

EoS							SbS						
UB 1 - blau	**N**	**M**	**SD**	**Min**	**Max**	**Range**	**UB 2 - gelb**	**N**	**M**	**SD**	**Min**	**Max**	**Range**
Ringe 1, E	22	1,75	0,54	0,65	2,80	2,2	Ringe 1, E	29	1,94	0,67	0,80	3,10	2,3
Ringe 2, V	23	2,01	0,49	0,90	3,00	2,1	Ringe 2, V	27	1,94	0,43	1,20	2,90	1,7
Ringe 3, V	23	**1,25**	0,39	0,50	2,00	1,5	Ringe 3, V	29	**1,63**	0,56	0,70	2,80	2,1
Ringe 4, K	23	0,98	0,34	0,40	1,80	1,4	Ringe 4, K	29	1,08	0,42	0,50	1,90	1,4
Ringe 5, V	23	1,60	0,58	0,60	2,70	2,1	Ringe 5, V	29	1,74	0,52	0,60	2,70	2,1
Ringe 6, V	22	**2,45**	0,45	1,70	3,20	1,5	Ringe 6, V	29	**2,78**	0,68	1,80	4,30	2,5
Reck 1, E	22	1,41	0,35	0,60	2,00	1,4	Reck 1, E	28	1,43	0,35	0,70	2,20	1,5
Reck 2, V	23	2,09	0,48	1,10	3,20	2,1	Reck 2, V	29	2,12	0,64	1,00	3,50	2,5
Reck 3, V	22	2,42	0,51	1,40	3,30	1,9	Reck 3, V	29	2,53	0,62	1,50	3,60	2,1
Reck 4, K	22	**1,40**	0,59	0,70	2,60	1,9	Reck 4, K	29	**1,72**	0,64	0,80	3,10	2,3
Reck 5, V	22	1,83	0,52	1,00	2,70	1,7	Reck 5, V	29	2,03	0,58	1,10	3,20	2,1
Reck 6, V	22	**2,15**	**0,83**	0,80	3,90	**3,1**	Reck 6, V	29	**2,43**	0,77	1,20	3,70	2,5

Die beiden in je zwei Stufen auftretenden manipulierten, unabhängigen Variablen sind die Reihenfolge und der Bewertungstyp. Die Reihenfolge

bezieht sich darauf, in welcher Videohälfte der Hauptfehler in der präsentierten Übung geturnt wird. Unterschieden werden Videosequenzen mit dem Fehler vorne von Sequenzen mit dem Fehler hinten. Der Faktor Bewertungstyp variiert ebenfalls in zwei Stufen. Eine Hälfte der VPn hat die Aufgabe, die Übungen nach den Wertungsvorschriften zu beurteilen und dabei vorgenommene Abzüge in der Ausführung nicht zu notieren (EoS). Die andere Hälfte der VPn bewertet auch nach den gültigen Vorschriften, darf aber die Abzüge notieren (SbS).

Wie im Erstexperiment beschrieben, ist auch im Zweitexperiment die absolute Höhe der gemittelten Abzüge (Tabelle 14 & 15) nicht von Interesse. Unterschiedlich hohe Absolutbewertungen der gezeigten Experimentalübungen machen eine übungsweise Betrachtung der Mittelwertunterschiede in den vier Untersuchungsbedingungen notwendig. Entsprechend erfolgt die Ergebnisdarstellung und -interpretation auf dieser Grundlage.

Tabelle 15: Deskriptive Statistik der Zweitexperiment-Übungen (Untersuchungsbedingung drei und vier)

EoS							SbS						
UB 3 - grün	**N**	**M**	**SD**	**Min**	**Max**	**Range**	**UB 4 - rot**	**N**	**M**	**SD**	**Min**	**Max**	**Range**
Ringe 1, E	27	1,84	0,63	0,40	2,80	2,4	Ringe 1, E	27	1,79	0,65	0,60	3,00	2,4
Ringe 2, H	27	2,04	0,48	1,20	3,10	1,9	Ringe 2, H	27	1,98	0,67	0,80	3,50	2,7
Ringe 3, H	27	**1,49**	0,57	0,50	2,60	2,1	Ringe 3, H	27	1,45	0,65	0,40	2,90	2,5
Ringe 4, K	27	1,07	0,43	0,30	2,00	1,7	Ringe 4, K	27	0,94	0,37	0,40	1,60	1,2
Ringe 5, H	27	1,44	0,67	0,40	2,60	2,2	Ringe 5, H	27	1,60	0,64	0,40	2,70	2,3
Ringe 6, H	27	2,58	0,62	1,40	3,80	2,4	Ringe 6, H	27	2,59	**0,81**	1,00	4,10	**3,1**
Reck 1, E	27	1,21	0,51	0,20	2,00	1,8	Reck 1, E	27	1,33	0,56	0,20	2,30	2,1
Reck 2, H	27	2,02	0,74	1,00	3,90	**2,9**	Reck 2, H	27	2,02	0,76	0,50	3,20	2,7
Reck 3, H	27	**2,26**	0,59	1,50	3,40	1,9	Reck 3, H	27	**2,53**	**0,83**	0,90	4,20	**3,3**
Reck 4, K	27	1,51	**0,76**	0,50	3,00	2,5	Reck 4, K	27	1,59	**0,81**	0,40	3,50	**3,1**
Reck 5, H	27	1,71	0,73	0,60	3,60	**3,0**	Reck 5, H	27	1,90	0,69	0,60	3,00	2,4
Reck 6, H	26	**1,97**	**0,87**	0,90	3,80	**2,9**	Reck 6, H	27	**2,38**	**1,03**	0,70	4,50	**3,8**

Auffällige Werte

Wie bereits im Rahmen des Erstexperiments werden die einzelnen Übungen bezüglich Auffälligkeiten in den berichteten Werten per Sicht-

prüfung untersucht. Die festgelegten Richtwerte des Erstexperiments für auffällige Mittelwertunterschiede (≥ 0,17 Punkten), Standardabweichungen (≥ 0,49 Punkten) und Spannweiten (≥ 1,9 Punkten) werden nicht für das Zweitexperiment verwendet. Sie werden ebenfalls prozentual berechnet, ergeben dabei aber vergleichbar hohe Werte (8.1.2). Eine Begründung hierfür und die spezielle Sichtung des Datenmaterials erfolgt nach der Prüfung auf Auffälligkeiten und der Darstellung der auffälligen Werte in den Kontrollübungen am Ende dieses Kapitels.

Somit ergibt sich für die Prüfung der Ergebnisse dieser Untersuchung ein Richtwert für Mittelwertunterschiede größer gleich 0,24 Punkten[31], eine Standardabweichung größer gleich 0,78 Punkten[32] und eine Spannweite von 2,9 Punkten[33] und mehr. Diese werden als auffällig bezeichnet und in Tabelle 14 und 15 hervorgehoben.

Die Betrachtung zeigt, dass die dritte Ringe-Übung einen Mittelwertunterschied zwischen Untersuchungsbedingung eins (M = 1,25) und zwei (M = 1,63) von 0,38 Punkten. Die Kampfrichter bewerteten hierbei die identischen Videosequenzen, lediglich die Art der Bewertung variierte. Weiterhin wird eine Übung, wenn sie den Fehler hinten zeigt (Untersuchungsbedingung drei; M = 1,49) mit durchschnittlich 0,24 Punkten strenger bewertet, als wenn der Fehler vorne präsentiert wird (Untersuchungsbedingung eins). Dieser Unterschied weist darauf hin, dass die Bewertung der Kampfrichter, wenn sie keine Abzüge notieren, von der präsentierten Reihenfolge der Information abhängig ist.

Die Ringe-Übung sechs weist einen 0,33 Punkte großen Unterschied zwischen dem Mittelwert der Untersuchungsbedingung eins (M = 2,45) und dem der Untersuchungsbedingung zwei (M = 2,78) auf. Dabei ist jeweils der Fehler vorne in der Übung enthalten. Untersuchungsbedingung vier zeigt zusätzlich eine auffällige Standardabweichung von 0,81 Punkten und eine Spannweite von 3,1 Punkten.

Die zweite Übung am Reck wird auffällig durch die erhöhte Standardabweichung von 2,9 Punkten in Untersuchungsbedingung drei. Übung 5 am Reck lässt in Untersuchungsbedingung drei eine Uneindeutigkeit in den Wertungen erkennen, die durch eine Spannweite der Wertungen von 3,0 Punkten kennzeichnet ist.

---

[31] Wie bereits im Erstexperiment ergibt sich dieser Wert aus der Berechnung von 13% des Mittelwerts aller Übungen (aller Untersuchungsbedingungen).

[32] Der Richtwert entspricht 130% der Mittelwerte aller Standardabweichungen.

[33] Dieser Wert stellt 130% aller Spannweitenmittelwerte der Übungen dar.

Reck-Übung drei weist sich durch einen 0,27 Punkte großen Unterschied der Mittelwerte von Untersuchungsbedingung drei (M = 2,26) und vier (M = 2,53; SD = 0,83; Range = 3,3) aus, die den Fehler hinten haben und sich durch die Art der Bewertung unterscheiden.

Die vierte Übung am Reck, die als Kontrollübung eingesetzt und somit im Faktor Reihenfolge nicht variiert wurde, zeigt einen auffälligen Mittelwertunterschied zwischen Untersuchungsbedingung eins (M = 1,40) und zwei (M = 1,72), mit einem Fehler in der ersten Hälfte, der sich auf 0,32 Punkte beläuft. Zusätzlich zeigen sich hohe Standardabweichungen in Untersuchungsbedingung drei (SD = 0,76) und vier (SD = 0,81; Range = 3,1) sowie eine erhöhte Range bei Untersuchungsbedingung drei.

Abschließend ergibt sich für die sechste Reck-Übung ein Unterschied von 0,28 Punkten zwischen Untersuchungsbedingung eins (M = 2,15; SD = 0,83; Range = 3,1) und zwei (M = 2,43). In beiden Untersuchungsbedingungen ist der Fehler in der ersten Hälfte der Übung enthalten. Weiterhin liegt in den Videosequenzen mit dem Fehler in der zweiten Hälfte, Untersuchungsbedingung drei (M = 1,97; Range = 2,9) und vier (M = 2,38), ein Mittelwertunterschied von 0,41 Punkten vor.

Kontrollübungen

Die Kontrollübungen werden auch im Rahmen der Ergebnisdarstellung des Zweitexperiments einer separaten Betrachtung unterzogen. Für das Zweitexperiment wurden Kontrollübungen eingesetzt, die für alle Untersuchungsbedingungen identisch sind und keine Besonderheiten in Form von groben Fehlern oder gar Stürzen beinhalten. Die Übungen unterscheiden sich nicht in der Reihenfolge der präsentierten Informationen, sondern lediglich in der Art der Bewertung.

Wie bereits im Erstexperiment fällt bei der Betrachtung der über die Untersuchungsbedingungen gemittelten Standardabweichungen der Ringe-Kontrollübung (Übung vier) mit einem Wert von SD = 0,39 auf, dass diese geringer als die Kontrollübung am Reck (Übung drei) ist, die einen Wert von SD = 0,7 ergibt. Im Vergleich dazu berechnet sich bei den Experimentalübungen an den Ringen eine Streuung von SD = 0,58 und am Reck ein entsprechend höherer Wert von SD = 0,7. Die Größenordnung je Gerät präsentiert sich annähernd identisch und die Tendenz zur größeren Streuung der Reckwertungen ist wie bereits im Erstexperiment (8.1.2) ausgeprägt.

Für die deskriptiven Ergebnisse des Zweitexperiments fallen zusammenfassend zunächst die erhöhten Abweichungen der Werte in Form von Mittelwerten, Standardabweichungen und Spannweiten auf. Wie bereits im Erstexperiment stellt sich auf den ersten Blick ein etwas uneinheitliches Bild ein. Ein Reihenfolge-Effekt zeigt sich nach visueller Prüfung nur in einer Übung an den Ringen (dritte Übung) und zwar in Kombination mit einem Unterschied in der Art der Bewertung. Der Bewertungstyp erweist sich auch im Zweitexperiment als der Faktor, der einen Unterschied in den Wertungsmittelwerten der Turnübungen hervorzubringen scheint. Die Besonderheit des Erstexperiments, das nur Ringe-Übungen mit dem Fehler hinten und Reck-Übungen mit dem Fehler vorne einen auffälligen Mittelwertunterschied im Bewertungstyp zeigen, ergibt sich für das Zweitexperiment nicht. Wenn ein Fehler vorne gezeigt wird, fallen zwei Ringe-Übungen und zwei Reck-Übungen durch Unterschiede in den Mittelwerten bezüglich der Art der Bewertung auf. Wenn hingegen der Fehler hinten präsentiert wird, ergibt sich nur bei Reck-Übungen ein auffälliger (nach obiger Definition von mehr als 0,24P) Mittelwertunterschied auf dem Faktor Bewertungstyp.

Alle erkennbaren und geschilderten Unterschiede haben gemeinsam, dass die SbS-Bedingung, in der die VPn ihre Abzüge mitschreiben dürfen, auch mehr Abzüge vornehmen, wie ihre Kollegen, die keine Möglichkeit der Mitschrift haben.

Besonders in Untersuchungsbedingung vier, in der die VPn ihre Abzüge mitschreiben durften und der Fehler in der zweiten Hälfte der Übungen präsentiert wurde, gibt es Anzeichen für eine Uneinigkeit zwischen den VPn. Die Übungen am Reck scheinen besonders betroffen zu sein, wenn man sich die Streuungen der Wertungen betrachtet. Ob die Unterschiede nach statistischer Prüfung relevant sind, zeigt folgendes Kapitel.

Festgelegte Richtwerte in unterschiedlicher Höhe

Die festgelegten Richtwerte des Erstexperiments und des Zweitexperiment werden prozentual berechnet und nicht äquivalent verwendet. Die einzelnen Richtwerte werden in einem ersten Schritt berichtet, um in einem zweiten Schritt begründet zu werden und, falls erforderlich, anhand des Datenmaterials belegt zu werden.

- Für das Erstexperiment gelten Mittelwertunterschiede $M \geq 0{,}17$ Punkte als auffällig. Für das Zweitexperiment hingegen stellen Werte von $M \geq 0{,}24$ Punkten einen auffälligen Unterschied dar.

Die Präsentation anderer Übungen im Zweitexperiment (im Vergleich zum Erstexperiment) führt selbstverständlich zu einer veränderten Situation der Mittelwerte der Abzüge. Jede Übung im Gerätturnen hat einen bestimmten Absolutwert an Abzügen in der B-Note und kann daher nicht sinnvoll, nur aufgrund dieser Note, mit anderen Übungen verglichen werden. Somit kann es sein, dass die Ausführungsabzüge der Zweitexperiment-Übungen im Vergleich zu denen des Erstexperiments größer sind. Warum die Streuungsmaße allerdings größer sind, lässt sich damit nicht erklären.

- Bezüglich der Standardabweichungen ergeben sich ebenso Unterschiede zwischen den beiden durchgeführten Untersuchungen. Im Erstexperiment ist SD $\geq$ 0,49 Punkten der Richtwert für auffällige Werte, während im Zweitexperiment SD $\geq$ 0,78 Punkten Gültigkeit findet.
- Die Spannweiten von R $\geq$ 1,9 Punkten im Erstexperiment und R $\geq$ 2,9 Punkten im Zweitexperiment stellen dieselbe Tendenz dar, dass das Zweitexperiment vergleichbar hohe Werte ergibt.

Zur Klärung, erscheint es sinnvoll, sich nur die Standardabweichungen der Kontrollübungen anzuschauen, da diese nicht in der Reihenfolge der präsentierten Informationen variieren, sondern lediglich im Bewertungstyp. Dabei fällt auf, dass die Ringe-Übungen (Erstexperiment – SD = 0,28; Zweitexperiment – SD = 0,39) durchschnittlich geringere Streuungen als die Reck-Übungen (Erstexperiment – SD = 0,4; Zweitexperiment – SD = 0,7) aufweisen.

Eine mögliche Interpretation dieses Ergebnisses ist, dass sich die VPn beim schnellen Gerät Reck uneiniger sind aufgrund der Bewegungsgeschwindigkeit der Elemente. Diese Vermutung deckt sich mit dem Grundgedanken der vierten Hypothese zum Geräte-Effekt, dass Übungen am Reck die Kampfrichter kognitiv höher belasten und diese Belastung durch auffälligere Wertungen zum Ausdruck kommt. Das Urteil der Kampfrichter ist anfälliger für Verzerrungen im Vergleich zum langsamen Gerät Ringe, da die kognitive Verarbeitung nicht so stark beansprucht wird. Damit könnte der Unterschied an den beiden präsentierten Geräten erklärt werden, aber die absolute Höhe der Streuungsmaße bedarf einer weiteren Begründung.

Um eine entsprechend, sinnvolle Begründung für diesen Sachverhalt zu finden, wird das Datenmaterial auf diese Besonderheit hin gesichtet. Vermutet wird, dass die Lizenzhöhe der Kampfrichter den Unterschied zwischen den beiden Untersuchungen bezüglich der Streuung

ausmacht. Grundlage dieser Annahme ist die im Zweitexperiment deutlich höhere Anzahl an Kampfrichtern mit internationaler Lizenz im Vergleich zum Erstexperiment (vgl. 6.1.2). Die anderen Personendaten der beiden Experimente ergeben ein sehr ähnliches Bild und lassen somit keinen Verdacht auf unterschiedliche Bewertungen zu.

Im Erstexperiment haben 20 VPn mit internationaler Lizenz, 31 mit nationaler Lizenzstufe A, 36 mit Lizenzstufe B, 13 mit Lizenzstufe C und 2 Kampfrichter der Lizenzstufe D teilgenommen. Im Vergleich dazu sind im Zweitexperiment Daten von 44 Kampfrichtern mit internationaler Lizenz, 30 mit Lizenzstufe A, 22 mit Lizenzstufe B und 8 mit Lizenzstufe C vorhanden. Damit präsentiert sich eine relativ große Differenz in der Stichprobenanzahl mit internationaler Lizenz. Sollte es zu variierenden Wertungen dieser Kampfrichtergruppe im Vergleich zu anderen Lizenzgruppen kommen, die in die erwartete Richtung zeigen – Kampfrichter mit internationaler Lizenz werten strenger, sehen mehr Fehler und nehmen höhere Abzüge vor – sei dies eine Erklärung für die erhöhten Richtwerte des Zweitexperiments.

Betrachtet man sich zur Klärung dieser Vermutung die 95%-Konfidenzintervalle[34] der Kontrollübungen des Erstexperiments (Abbildung 9) und des Zweitexperiments (Abbildung 10), fallen starke Ähnlichkeiten[35] auf.

Die Kampfrichtergruppe internationaler Lizenzstufe, die in den Abbildungen ganz links abgetragen ist, nimmt in beiden Untersuchungen die größten Abzüge vor (Ausnahme: Ringe-Kontrollübung im Zweitexperiment). Dieses Bild zeigt sich auch bei den Experimentalübungen. Damit könnte sich die größere Anzahl an international lizenzierten Kampfrichtern im Zweitexperiment auf die stärkeren Streuungen der Daten auswirken.

---

[34] Ein Konfidenzintervall schätzt die Grenzen, innerhalb derer sich der wahre Populationswert mit hoher Wahrscheinlichkeit (95%) befindet (Bortz & Döring, 2003, S. 414). Der mittlere Punkt eines Intervalls bildet den Stichprobenmittelwert ab. Die Breite des Intervalls wird bestimmt durch den Sicherheitsgrad der Schätzung, die Stichprobengröße und die tatsächliche Variation der Werte.

[35] Zur besseren Vergleichbarkeit werden die Skalen der Abzüge auf 1,4 Punkte vereinheitlicht. Die beiden VPn des Erstexperiments mit der Lizenzstufe D sind nicht abgebildet. Sie haben uneinheitliche Bewertungen vorgenommen, was zu einem vergleichsweise breiten Konfidenzintervall führt. Dies führt wiederum zu einem schlecht erkennbaren Bild der anderen, dadurch schmalen Konfidenzintervalle.

Abbildung 9: Konfidenzintervalle der Kontrollübungen im Erstexperiment (Ringe-Übung drei – links; Reck-Übung drei – rechts)

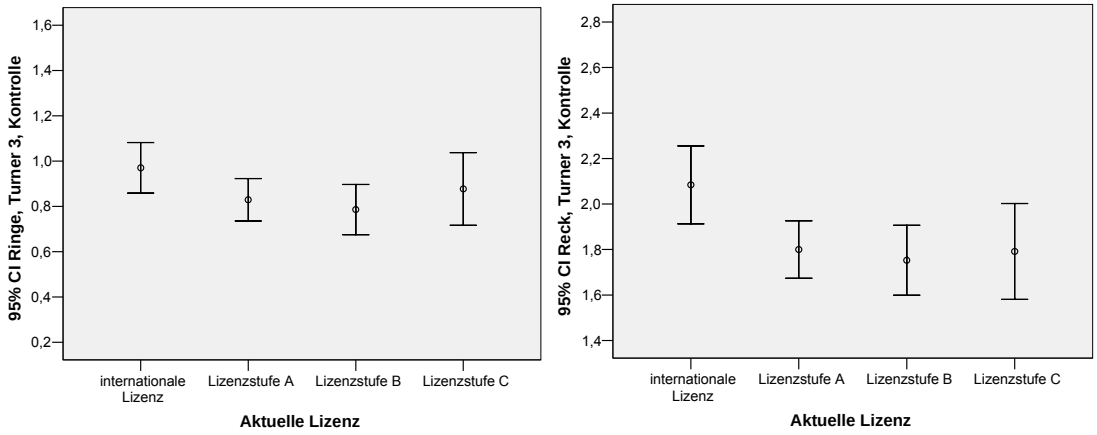
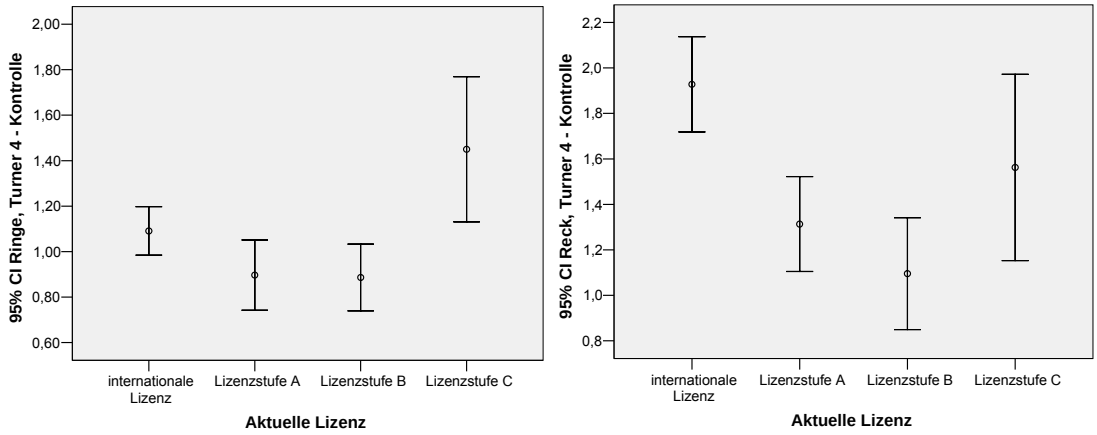


Abbildung 10: Konfidenzintervalle der Kontrollübungen im Zweitexperiment (Ringe-Übung vier – links; Reck-Übung vier – rechts)



8.2.2 Inferenzstatistische Auswertung

Die Auswertung erfolgt wie bereits im Erstexperiment für jede Übung separat. Die erste Übung jedes Gerätes stellt die Einwertübung dar und wird nicht ausgewertet. Die Wertungen der Kontroll- und Experimentalübungen werden geräteweise mittels zweifaktorieller VA (ANOVA) für unabhängige Stichproben mit den Faktoren Reihenfolge und Bewertungstyp statistisch überprüft.

Die a posteriori Teststärkenanalyse ergibt im Zweitexperiment für einen kleinen Effekt, einer Stichprobengröße von N = 106, die auf vier Gruppen aufgeteilt werden, und einem α-Fehlerniveau von 10% eine geringe Teststärke (1-β) von 0,198. Das β -Fehlerniveau von 0,8 und ein kritischer F-Wert von 2,138 können somit angenommen werden.

1. Hypothese: Reihenfolge-Effekt

Ob die Videosequenzen mit dem Fehler in der ersten Hälfte der Übung mit geringeren Abzügen als die Übungen mit dem Fehler in der zweiten Hälfte gewertet werden und es somit zu einem Recency-Effekt kommt, oder ob sich ein Primacy-Effekt zeigt, soll mit Hilfe der univariaten zweifaktoriellen Varianzanalyse ohne Messwiederholung untersucht werden. Dabei stellt die Reihenfolge der präsentierten Informationen den ersten Faktor dar. Sollte ein Reihenfolge-Effekt vorherrschend sein, müsste er sich als Haupteffekt zeigen. Die Kontrollübungen hingegen sollten keine Unterschiede auf dem ersten Faktor erkennen lassen, da keine Manipulation durchgeführt wurde.

Übereinstimmend mit den deskriptiven Ergebnissen des Faktors Reihenfolge kann in keiner Übung ein signifikanter Haupteffekt aufgedeckt werden. Der Reihenfolge-Effekt kann somit im durchgeführten Zweitexperiment nicht nachgewiesen werden und die Reihenfolge der Fehlerpräsentation scheint keinen überzufälligen Einfluss zu haben.

Die Effektstärken liegen im niedrigen Wertebereich (η^2 = 0,001 bis 0,015) und klären nur einen sehr geringen Varianzanteil auf. Die Teststärken nehmen Werte zwischen 5,1% und 36,4% an, was ebenfalls als gering angesehen wird. Die absoluten Unterschiede der Abzüge in den einzelnen UB, berechnet aus der Differenz der gemittelten Wertungen der EoS- und der SbS-Bedingung einer Übung mit dem Fehler vorne und mit dem Fehler hinten, ergeben Werte bis zu 0,153 Punkten an den Ringen und bis zu 0,139 Punkten am Reck.

Auch in Bezug auf den zweiten Datensatz darf die Forschungshypothese eins, es besteht ein Unterschied in den Wertungen der Beurteiler aufgrund der präsentierten Reihenfolge der Informationen, nicht angenommen werden. Da sich der Reihenfolge-Effekt nicht in den Daten zeigt, muss die Nullhypothese beibehalten werden.

2. Hypothese: Effekt des Bewertungstyps

Vermutet wird, dass der Reihenfolge-Effekt unabhängig vom Bewertungstyp entsteht. Es wird davon ausgegangen, dass sich kein Unterschied in den Abzügen ergibt und damit irrelevant ist, ob ein Kampfrichter die Abzüge notiert oder nicht. Kein Haupteffekt des Faktors Bewertungstyp sollte als Ergebnis vorliegen, um die Vermutung anzunehmen.

Alleinig die *sechste Übung am Reck* zeigt sich signifikant bezüglich eines Haupteffekts auf dem Faktor Bewertungstyp (F(0,867) = 3,063; p = 0,05, η^2 = 0,038). Der η^2-Wert schätzt einen kleinen Effekt für diese

Unterschiedsprüfung. Die Reckübung lässt bei beiden Faktoren Unterschiede in den Mittelwerten erkennen.

Der Mittelwert der Abzüge von der SbS-Bedingung (μ = 2,40) liegt konsistent über dem der EoS-Bedingung (μ = 2,05) (Tabelle 16). Das macht einen durchschnittlichen Unterschied der beiden Bewertungstypen von etwa 0,35 Punkten. Daher kann festgehalten werden, dass sowohl bei dem Video mit dem Fehler vorne als auch bei dem mit dem Fehler hinten das Notieren der Abzüge zur schlechteren Wertung führt. In den Interaktionsdiagrammen (Abbildung 11) lassen sich die gleichen Trends erkennen.

Tabelle 16: Abzüge für die Experimentalübung sechs am Reck

Reihenfolge	Bewertungstyp	M	SD	N
Video 1 (Fehler vorne)	EoS	2,150	,8268	22
	SbS	2,428	,7657	29
	Total	2,308	,7967	51
Video 2 (Fehler hinten)	EoS	1,965	,8699	26
	SbS	2,378	1,0345	27
	Total	2,176	,9707	53
Total	EoS	2,050	,8465	48
	SbS	2,404	,8973	56
	Total	2,240	,8878	104

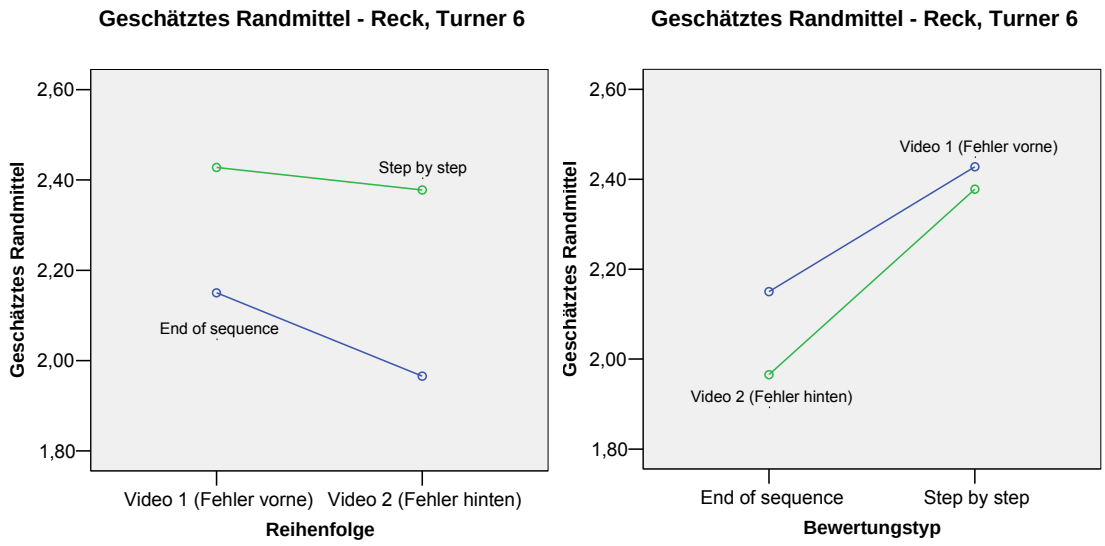


Abbildung 11: Interaktionsdiagramme der sechsten Reck-Übung (Reihenfolge - links und Bewertungstyp - rechts)

Die Effektstärken des Faktors Bewertungstyp belaufen sich bei allen ausgewerteten Übungen auf ein Eta-Quadrat zwischen η^2 = 0,000 und 0,038. Die Teststärke ergibt Werte zwischen 5,1% und 33,9% (Reck, Übung sechs, wurde berichtet und ist nicht inbegriffen). Betrachtet man sich die absoluten Unterschiede der Abzüge, berechnet aus der Differenz der durchschnittlichen EoS-Wertungen und der SbS-Wertungen, ergeben sich Werte bis zu 0,169 Punkten an den Ringen und bis zu 0,203 Punkten am Reck.

Mit Blick auf Forschungshypothese zwei, und im Unterschied zur ersten Untersuchung, kann die Alternativhypothese nicht angenommen werden, da sich die Mittelwertunterschiede der Wertungen der beiden Bewertungstypen nicht statistisch signifikant unterscheiden. Damit kann die Nullhypothese beibehalten werden. Der Bewertungstyp erweist sich im Zweitexperiment lediglich im Zusammenhang mit einer Übung als überzufällig. Alle anderen Übungen weisen keine statistisch bedeutenden Unterschiede auf.

3. Hypothese: Interaktionseffekt ‚Reihenfolge x Bewertungstyp'

Theoretische Überlegungen und die Theorie von Hogarth und Einhorn (1992) (7.1) lassen die Vermutung zu, dass die beiden Faktoren Reihenfolge und Bewertungstyp nicht interagieren.

Das Datenmaterial zeigt nach statistischer Überprüfung bei der *Experimentalübung drei an den Ringen* einen nach Fairnesskriterium (7.2) bedeutenden Interaktionseffekt ‚Reihenfolge x Bewertungstyp' ($F(1,661) = 3,741$, $p = 0.056$, $\eta^2 = 0,035$) ohne weitere Haupteffekte. Für den Interaktionseffekt schätzt der η^2-Wert einen kleinen Effekt.

Die Mittelwerte unterscheiden sich nicht signifikant voneinander, weder bei den Bewertungstypen EoS (μ = 1,38) und SbS (μ = 1,54) noch bei den Videos eins (V) (μ = 1,46) und zwei (H) (μ = 1,47) (Tabelle 17).

Die detaillierte Betrachtung der einzelnen Zellenmittelwerte lässt Folgendes erkennen: Wird ein Fehler zu Beginn der Übung präsentiert, ahnden die Kampfrichter diese mit höheren Abzügen, wenn sie sich Notizen machen (μ = 1,63). Wenn sie keine Notizen machen, bewerten sie die gleiche Übung mit weniger Ausführungsfehlern (μ = 1,25). Bei einem Fehler am Ende der Übung ergibt sich ein sehr geringer Unterschied in gegensätzlicher Richtung. Kampfrichter, die ihr Urteil nach Beendigung des Videos fällen, werten strenger (μ = 1,48) als diejenigen, die dem SbS-Prozess folgen und die Abzüge notieren (μ = 1,45).

Tabelle 17: Abzüge für die Experimentalübung drei an den Ringen

Reihenfolge	Bewertungstyp	M	SD	N
Video 1 (Fehler vorne)	EoS	1,252	,3872	23
	SbS	1,631	,5581	29
	Total	1,464	,5213	52
Video 2 (Fehler hinten)	EoS	1,489	,5713	27
	SbS	1,448	,6524	27
	Total	1,469	,6078	54
Total	EoS	1,380	,5047	50
	SbS	1,543	,6069	56
	Total	1,466	,5643	106

Die dritte Forschungshypothese, es besteht eine Interaktion der beiden Faktoren, kann im Gegensatz zur ersten Untersuchung nicht angenommen werden. An den Ringen zeigt sich bei einer Experimentalübung ein Interaktionseffekt ohne Haupteffekte. Ansonsten weisen die statistischen Berechnungen keine weiteren signifikanten Effekte aus. Die Nullhypothese kann somit beibehalten werden.

4. Hypothese: Geräteeffekt

Angenommen wird, dass jegliche Effekte, im speziellen Fall der Reihenfolge-Effekt oder der Einfluss des Bewertungstyps, falls sie denn vorhanden sind, am schnellen Gerät Reck stärker ausgeprägt sind als am langsamen Gerät Ringe.

Ein Geräte-Effekt lässt sich bezüglich der präsentierten Reihenfolge der Informationen nicht erkennen. Weder an den Ringen noch am Reck lässt sich ein Reihenfolge-Effekt feststellen. Der erste Faktor Reihenfolge scheint daher an keinem Gerät einen überzufälligen Einfluss zu haben und eine explizite Rolle beim Werten von Gerätturnübungen zu spielen. Auch im Zweitexperiment darf die vierte Forschungshypothese nicht angenommen werden und die Nullhypothese ist beizubehalten.

Zusammenfassend erfolgt eine tabellarische Darstellung der Untersuchungsergebnisse des Zweitexperiments (Tabelle 18). Geräteweise werden die signifikanten oder nach Fairnesskriterium auffälligen Unterschiede, die Effektgrößen und die Reihenfolge-Trends dargestellt. Die Reihenfolge-Trends geben Auskunft darüber, ob die Übung mit dem Fehler vorne oder die mit dem Fehler hinten zu größeren Abzügen geführt hat. Zeigt sich für die Übung mit dem Fehler vorne ein größerer Abzug im Vergleich zur Übung mit dem Fehler hinten, spricht das für

eine Tendenz zum Primacy-Effekt. Eine Tendenz zum Recency-Effekt ergibt sich, wenn die Übung mit dem Fehler hinten mehr Abzüge erhält. Diese Tendenzen werden nach den beiden Bewertungstypen EoS und SbS getrennt aufgeführt.

Tabelle 18: Zweitexperiment – Überblick zum inferenzstatistischen Vergleich der Experimentalübungen (Einwert- und Kontrollübungen nicht aufgeführt; HE – Haupteffekt)

Gerät, Übung	Effekt (p)	η^2	Reihenfolge-Trend	
			EoS	SbS
Ringe, 2	------	------	Recency	
Ringe, 3	Interaktion (0.056)	0.035	Recency	Primacy
Ringe, 5	------	------	Primacy	
Ringe, 6	------	------	Recency	Primacy
Reck, 2	------	------	Primacy	
Reck, 3	------	------	Primacy	------
Reck, 5	------	------	Primacy	
Reck, 6	HE Bewertungstyp (0.05)	0.038	Primacy	

Bei der Betrachtung der Reihenfolge-Trends zeichnet sich nur bei der Ringe-Übung zwei, unabhängig vom Bewertungstyp, ein Recency-Effekt ab. Die dritte und die sechste Ringe-Übung deuten beim EoS-Prozess einen Recency-Effekt an, während der SbS-Prozess für einen Primacy-Effekt spricht. Die dritte Reck-Übung weist im EoS-Prozess auf einen Primacy-Effekt hin und in der SbS-Bedingung bewerten die Kampfrichter die Übung mit dem Fehler vorne identisch wie die mit dem Fehler hinten. Daher ergibt sich hierbei kein Trend. Die restlichen Übungen, Ringe-Übung fünf, Reck-Übung zwei, fünf und sechs haben unabhängig vom Bewertungstyp einen Primacy-Effekt gemeinsam. Somit ergibt sich für die Ringe-Übungen ein unklares Bild, während die Reck-Übung klar auf einen Primacy-Effekt hinweisen.

9 Interpretation und Relevanz der Ergebnisse

Zu Beginn der Arbeit wird die Frage aufgeworfen, ob der Reihenfolge-Effekt bei schnellen Bewegungen am Beispiel von Kampfrichterurteilen im Gerätturnen der Männer sichtbar wird und damit für diese Urteilssituation relevant ist. Neben dieser Hauptfragestellung wird in der vorliegenden Arbeit überprüft, ob die Art der Bewertung einen Einfluss auf die Höhe der vorgenommenen Abzüge für die Ausführung der Übung nimmt. Ein letzter interessanter Aspekt, der Beachtung findet, ist die Unterscheidung in schnelle und langsame Geräte. Die vermutlich ebenfalls zu Unterschieden in der Beuteilung von Turnübungen führt.

Zur Prüfung dieser Fragen wurden experimentelle Untersuchungen durchgeführt, in denen unterschiedliche Untersuchungsmaterialien zur Anwendung kamen. Anhand von manipulierten Videosequenzen bekamen die Versuchspersonen (VPn) unterschiedliche Übungen an den Geräten Ringe und Reck zu sehen. Auf vier Untersuchungsbedingungen zufällig zugeordnet, hatte ein Viertel der VPn die Aufgabe, Übungen anhand des EoS-Prozesses zu bewerten und sie durften dabei keinerlei Notizen während der Präsentation der Videosequenzen machen. Sie sichteten, durch für alle Untersuchungsbedingungen identische Kontrollübungen getrennt, Übungen mit dem Fehler in der ersten Hälfte[36]. Eine zweite Gruppe, und damit ein weiteres Viertel der VPn, sichtete dieselben Videosequenzen, allerdings nach dem SbS-Prozess. Dabei notierte die Gruppe alle Fehler in den Übungen anhand von Abzügen. Die dritte Gruppe an VPn bewertete Übungen mit dem Fehler in der zweiten Hälfte der Übung und nach dem Bewertungstyp EoS. Die vierte Gruppe hatte ebenfalls Übungen mit dem Fehler hinten zu beurteilen, sollte aber die Fehler anhand von Abzügen mitschreiben.

Zwei Datensätze lizenzierter, deutschsprachiger Kampfrichter wurden erhoben. Die beiden dafür durchgeführten Untersuchungen unterscheiden sich hinsichtlich der präsentierten Anzahl an Übungen, dem in Nuancen variierendem Untersuchungsdesign, in einer optimierten Vorbereitung der Videosequenzen und Verbesserungen der Untersuchungsdurchführung beim Zweitexperiment im Vergleich zum Erstexperiment (vgl. 6.3).

[36] Im Erstexperiment sind auch Übungen mit dem Fehler hinten enthalten. Diese Besonderheit wird in der Interpretation und Diskussion der Ergebnisse berücksichtigt.

Keine signifikanten Ergebnisse bezüglich der Hauptfragestellung

Beide Untersuchungen ergeben bezüglich der Hauptfragestellung, ob es einen Reihenfolge-Effekt gibt, statistisch nicht signifikante Ergebnisse. Die Nullhypothese wird in beiden Fällen beibehalten. Somit kann festgehalten werden, dass die Ergebnisse nicht den Erwartungen entsprechen.

Der *Reihenfolge-Effekt, Hypothese 1,* konnte bei keiner Übung eindeutig nachgewiesen werden. Dieses Ergebnis führt zu unterschiedlichen Vermutungen, die versuchen, einen Erklärungsansatz zu liefern und Ideen für weitere Forschungsarbeiten zu generieren (10). Dabei entsteht die Vermutung, dass sich die Reihenfolge-Effekte gegenseitig aufheben und daher nicht nachgewiesen werden können. Dies würde bedeuten, dass der von Hogarth und Einhorn (1992) vorhergesagte Recency-Effekt für komplexe Aufgaben durch einen Primacy-Effekt aufgehoben wird und infolgedessen nicht sichtbar ist. Ein Primacy-Effekt (vgl. 3.3) könnte aufgrund folgender Besonderheiten der gewählten Urteilssituation vorherrschend sein:

1. Wird die *Leistungsfähigkeit* als ein stabiles Gebilde betrachtet (Jones et al., 1968), könnte dies dazu führen, dass nach der Bildung eines ersten Eindrucks des Kampfrichters nur noch Urteilsanpassungen vorgenommen werden. Durch die Stabilität des Fähigkeitsgebildes ist der Primacy-Effekt im Kontext Sport allgemein recht dominant (Greenless et al. 2007). Diese Art der Begründung ist für die gewählte Urteilssituation im Gerätturnen denkbar und kann für den speziellen Sachverhalt wie folgt erläutert werden: Wenn ein Fehler in der ersten Übungshälfte wahrgenommen wird, entsteht beim Kampfrichter möglicherweise der erste Eindruck, dass diese Übung bzw. der Athlet an diesem Gerät nicht gut ist. Damit werden darauffolgende Elemente kritischer betrachtet und führen eher zur Wahrnehmung bzw. negativeren Einschätzung von Fehlern. Daraus resultiert eine schlechtere Bewertung dieser Übung, die sich in der B-Note in Form von höheren Abzügen bemerkbar macht. Zeigt sich hingegen in der ersten Hälfte der Übung kein gravierender Fehler, entsteht beim Kampfrichter der Eindruck einer guten Übung bzw. eines guten Turners am entsprechenden Gerät. Dieser Eindruck bewirkt die Abwertung inkonsistenter Informationen, also beispielsweise mittlere Fehler in der Ausführung werden als kleine angesehen, und führt letztendlich dazu, dass Fehler in der zweiten Hälfte nicht so streng geahndet werden und damit eine relativ gute Bewertung abgegeben wird.

2. Die für die Untersuchungen ausgewählten Kampfrichter konnten im Rahmen verschiedener Veranstaltungen in denen sie ihrer Kampfrichtertätigkeit nachgegangen sind untersucht werden. Dabei wurden sie in größeren Pausen oder nach Beendigung der Veranstaltung gebeten, an der Studie teilzunehmen. Vermutet werden könnte, dass sich zu diesem Zeitpunkt bereits eine gewisse *Müdigkeit* aufgrund der vorhergehenden Kampfrichtertätigkeit eingestellt hat, die dazu führte, dass ein Konzentrationsabfall im Laufe der Untersuchung stattfand und damit die anfänglichen Informationen stärker in das Urteil einbezogen wurden. Dabei wird angenommen, dass es sich nicht um einen motivational gesteuerten, bewussten Vorgang, sondern um einen natürlichen Ermüdungsprozess handelt. Diese Einschätzung resultiert aus den Reaktionen der VPn, die im persönlichen Gespräch nach dem Experiment schilderten, dass sie sich auf die Studie gefreut haben und gespannt waren, was sie erwarten würde.
3. Als weiterer Faktor, der auf die unter 2. geschilderte Untersuchungssituation zurückzuführen ist, stellt den möglicherweise entstandenen *Zeitdruck* der VPn dar. Dieser wurde nicht aufgrund der Untersuchung an sich hervorgerufen, da explizit keine zeitliche Vorgabe der Wertungsaufgabe festgelegt wurde (6.2), sondern aufgrund des eigenen Ehrgeizes der teilnehmenden Kampfrichter. Einerseits könnte die Zeitspanne, die von der Pause abgezogen wird bzw. vom bevorstehenden Feierabend trennt, dazu geführt haben. Diese Erklärung stellt einen eher motivational orientierten Ansatz dar, der eher für die Kampfrichter gültig sein könnte, die weniger motiviert an der Untersuchung teilgenommen haben. Andererseits könnten sich die Kampfrichter zeitlich unter Druck gesetzt gefühlt haben, da sie oftmals gleichzeitig mit ihren Kollegen am Experiment teilnahmen und so möglicherweise eine leichte Konkurrenzsituation ausgelöst wurde. Zeitdruck führt zur Tendenz, die ersten Informationen stärker in der Urteilsbildung zu berücksichtigen (Kruglanski & Webster, 1996), was dafür spricht, dass sich ein Primacy-Effekt eingestellt haben könnte.
4. Eine weitere Begründung für das Nichtauftreten eines deutlichen Reihenfolge-Effekts könnte sein, dass die Bewegungen an beiden Geräten in *zu schneller Folge* aufeinander folgen und damit die Zeitspanne zwischen den Elementen zu kurz ist, um bei den VPn als separate Informationen wahrgenommen zu werden. Somit nimmt der Kampfrichter eine Übung auch als eine Information wahr und

verarbeitet diese entsprechend. Die VPn sieht die Informationen nicht als unterschiedliche, sequenziell auftretende an, die entweder aufsteigend (bzw. in der Reihenfolge negativ-positiv) oder absteigend (bzw. in der Reihenfolge positiv-negativ) verlaufen. Somit ergibt sich für den Moment der Urteilsbildung das Bild, dass ein Fehler erkannt wurde und nicht, dass dieser am Ende der Übung stattgefunden hat.

5. Als weitere Erklärung könnte eine statistische Herangehensweise hilfreich sein: Wenn man sich mittels *Teststärkenanalyse* die ermittelten Werte (8.1.3 & 8.2.2) genauer anschaut, sieht man, dass es mit der erhobenen Stichprobengröße nicht sehr wahrscheinlich ist, einen Effekt zu belegen. Bei einem mittleren Effekt (und in diesem Fall geht man eher von einem kleinen bis mittleren Effekt aus), einem α-Fehler von 0,05 und einer Gesamtstichprobe von 100 Personen (3 Freiheitsgrade, 4 Gruppen) ergibt sich eine Teststärke (1-β) von 0,518. Das bedeutet, dass die Wahrscheinlichkeit einen mittleren Effekt zu entdecken, wenn es einen gibt, mit 95%iger Irrtumswahrscheinlichkeit, bei etwa 50% liegt. Sobald allerdings ein kleiner Effekt gewählt wird, verringert sich diese Wahrscheinlichkeit noch einmal dramatisch (im Beispiel auf 0,003). Somit kann man argumentieren, dass es in der gegebenen Untersuchung, mit eingeschränkter und nur schwer steigerbarer Stichprobengröße und kleinem zu erwartenden Effekt gar nicht zu einem signifikanten Ergebnis kommen muss, um den Effekt dennoch beobachten zu können. Aufgrund dessen werden die Reihenfolge-Trends (Tabelle 13 & Tabelle 18) im Folgenden vorsichtig interpretiert.

Reihenfolge-Trend-Interpretation

Bei der Betrachtung der Reihenfolge-Trends des Erstexperiments und des Zweitexperiments lässt sich vorsichtig die Tendenz deuten, dass die präsentierte Reihenfolge der Informationen am langsamen Gerät Ringe nicht in eine klare Richtung zeigt, während am schnellen Gerät Reck ein Primacy-Effekt erkennbar ist. Die absoluten Abzüge, bezüglich präsentierter Fehlerposition, belaufen sich an den getesteten Geräten Ringe und Reck sowohl im Erstexperiment als auch im Zweitexperiment, durchschnittlich auf über 0,1 Punkte. Diese gerätespezifische Herangehensweise deckt sich mit dem vermuteten *Geräte-Effekt der 4. Hypothese,* der in den Untersuchungen statistisch nicht nachgewiesen wird und zur Beibehaltung der Nullhypothese führt. Aus genannten Gründen wird eine gerätespezifische Interpretation vorgenommen.

Das *langsame Gerät Ringe*, an dem mit statischen Übungselementen geturnt wird, zeigt sowohl im Rahmen des Erstexperiments als auch im Zweitexperiment einen tendenziellen *Recency-Effekt*. Dieser Trend zeigt sich allerdings nur im Zusammenhang mit dem Bewertungstyp *EoS*, also wenn die VPn ihre Abzüge nicht notieren dürfen. Erklärt werden könnte dieses Ergebnis, das sich nicht mit den Aussagen von Hogarth und Einhorn (1992) deckt, durch (1) Vergessensunterschiede und die (2) zusätzliche kognitive Aufgabe, die diese VPn bearbeiten müssen.

(1) Die zuletzt präsentierten Informationen, also im speziellen Fall Turnelemente, sind besser aus dem Gedächtnis abrufbar als frühere Informationen (Ebbinghaus, 1885; Miller & Campbell, 1959). Man kann sich vorstellen, wie schwierig es ist, sich nach einer Reihe von aufeinander folgenden Turnelementen alle Fehler zu erinnern, wenn man keine Notizen machen darf.

(2) Als zusätzliche kognitive Aufgabe könnte man das Zusammenrechnen der Abzüge im Laufe der Übung interpretieren. Im Gerätturnen werden die Kampfrichter daraufhin ausgebildet und beraten, die Wertungen mithilfe von Notizen als eine Art Gedächtnisstütze zu ermitteln. Dabei folgen sie dem SbS-Prozess, den sie bei regelmäßiger Ausübung als die Art und Weise der Urteilbildung heranziehen. Somit ist ein Großteil der Kampfrichter daran gewöhnt, die einzelnen Abzüge zu extrahieren und zu vermerken. Wenn sie die Möglichkeit der Notiz nicht haben, kann gemutmaßt werden, dass sie automatisch alle Abzüge im Gedächtnis zusammenzählen, um ihr Urteil zu fällen. Diese zusätzlich ausgeführte Rechenaufgabe begünstigt den Recency-Effekt (Luchins, 1957; Greenless et al., 2007).

Wenn die Kampfrichter ihre Abzüge am Gerät Ringe mitschreiben dürfen, kann man vorsichtig von einer Tendenz zum *Primacy-Effekt* sprechen. Dieses Ergebnis entspricht den vermuteten Begründungen bezüglich des Nichtauftretens, des von der Theorie abgeleiteten Recency-Effekts.

Das als schnell bezeichnete Gerät *Reck*, dass keine statischen Übungselemente enthält und für das menschliche Auge durch die sehr schnell ablaufenden Bewegungen vermutlich zu einer höheren kognitiven Belastung führt, zeigt hingegen in beiden durchgeführten Experimenten eine Tendenz zum Primacy-Effekt. Der Primacy-Effekt könnte, wie bereits formuliert, aufgrund der Besonderheiten der Untersuchungen erklärt werden.

Der *Einfluss des Bewertungstyps, Hypothese 2,* führt in den beiden Experimenten zu unterschiedlichen Ergebnissen. Entsprechend den Vorhersagen der zu prüfenden Modellvorhersagen (Hogarth & Einhorn, 1992), dass einen Reihenfolge-Effekt bei komplexen Aufgaben unabhängig vom Bewertungstyp auftritt, konnte im Erstexperiment nicht gezeigt werden, wurde aber im Zweitexperiment bestätigt.

Eine Erklärung für diese unterschiedlichen Ergebnisse kann wiederum nur vermutet werden. Denkbar ist, dass das Zweitexperiment aufgrund des höheren Anteils an international lizenzierten Kampfrichtern (8.2.1), die die am besten ausgebildeten VPn darstellen, keine Unterschiede hervorbringt. Gut ausgebildete Kampfrichter beschäftigen sich im Vergleich zu den weniger gut Ausgebildeten intensiver mit der Kampfrichterei und haben insgesamt gesehen eine bessere Wahrnehmungsfähigkeit erlangt. Die kognitive Belastung ist aufgrund der routinierteren Bearbeitung des Wertens nicht mehr so hoch wie bei den weniger Geübten. Damit könnte erklärt werden, dass die zusätzliche Aufgabe des Zusammenrechnens von Abzügen im EoS-Prozess, aufgrund der geringeren Belastung des Wertens, besser tolerierbar ist und daher geringere Vergessensunterschiede entstehen. Dadurch gleichen die Wertungen des EoS-Prozesses eher dem SbS-Prozess und es werden keine Unterschiede im Bewertungstyp sichtbar.

Anders verhält es sich im Erstexperiment, in dem anteilig vergleichsweise mehr national lizenzierte Kampfrichter teilgenommen haben. Diese VPn konnten die zusätzliche kognitive Belastung des Zusammenzählens weniger gut kompensieren und daher konnte der Bewertungstyp unterschiedliche Urteile für ein und dieselbe Übung unterschiedlichen Bewertungstyps zeigen.

Die Art der Bewertung scheint somit, vor allem für weniger routinierte Kampfrichter, nicht unbedeutend für die Sportart Gerätturnen zu sein. Die absoluten Abzüge, bezüglich verwendetem Bewertungstyp, belaufen sich an den getesteten Geräten Ringe, auf durchschnittlich 0,16 Punkte, und am Reck auf 0,2 Punkte. Dieses Ergebnis überrascht jedoch nicht, da das Notieren von Abzügen ein gewolltes und bei hochklassigen Wettkämpfen gefordertes Hilfsmittel bei der Bewertung von Turnübungen ist. Die Bewertungspraxis und die durchgeführten Untersuchungen, in denen sich die VPn, die mitschreiben sollten, dies nicht taten und anmerkten, sie würden das nie tun, zeigt jedoch, dass es durchaus Kampfrichter gibt, die ihre Abzüge nicht mitschreiben. Diese Form der Bewertung ist somit keine komplett unübliche, sondern eine weniger oft praktizierte.

Kampfrichter-Fortbildung

Zusammenfassend und mit Blick auf die Kampfrichterausbildung und -fortbildung lassen sich aus den durchgeführten Untersuchungen sowohl bezüglich des Reihenfolge-Effekts als auch des Einflusses des Bewertungstyps und des Geräte-Effekts folgende Ratschläge ableiten:

- In der Kampfrichteraus- und -fortbildung sollte eine ausreichende Aufklärung über die verschiedenen Einflüsse, denen der Kampfrichter im Wettkampf ausgesetzt ist, durchgeführt werden. Viele Einflüsse lassen sich durch die Warnung auf ein Mindestmaß reduzieren oder gar komplett verhindern. Hierauf deutet eine im Sport durchgeführte Untersuchung (Greenless et al., 2009) hin, die zu dem Schluss kommt, dass sich der Reihenfolge-Effekt durch eine Warnung verhindern lässt (3.3).
- Der eher ‚stiefmütterliche' Umgang mit Kontrollen bezüglich der vorgenommenen und notierten Abzüge in der Kampfrichterausbildung und auch in Wettkämpfen, sollte vor allen Dingen bei Kampfrichtern forciert werden, die keine internationalen Wettkämpfe werten. In internationalen Wettkämpfen ist das Mitschreiben der geturnten Elemente in Kurzschrift und den dazugehörigen Abzügen in der B-Note, um eine spätere Zuordnung der Fehler auf konkrete Elemente vornehmen zu können, eher vorzufinden als in weniger hochklassigen Wettkämpfen. Wie bereits beschrieben, weisen Unterschiede zwischen dem EoS- und SbS-Bewertungstyp darauf hin, dass das Notieren der Abzüge zu Wertungsunterschieden führt. Im EoS-Prozess werden höhere kognitive Anforderungen abverlangt, die sich auf die Wertungen auswirken. Somit sollte in der Aus- und Fortbildung der Kampfrichter und in den Wettkämpfen vermehrt darauf geachtet werden, dass diese ihre Abzüge für die B-Note korrekt mitschreiben.
- Besonders die schnellen Geräte, die ohne statische Elemente geturnt werden, beanspruchen einen hohen Anteil der beim wertenden Kampfrichter vorhandenen Informationsverarbeitungskapazität. Um diese Kapazität nicht zu überschreiten, ist die häufige Sichtung von verschiedenartigen Elementen ratsam. Dies könnte beispielsweise im Rahmen von Übungsmaterial ermöglicht werden. Anhand von speziell bearbeiteten Videosequenzen, die als eine Art Computerspiel mit Lernstufen organisiert sind, könnte den Kampfrichtern die häufigere Sichtung von sehr schnellen, neue eingeführten, aber

auch seltenen Elementen zugänglich gemacht werden, um ihre visuelle Leistungsfähigkeit verstärkt zu schulen.

Die ermittelten Ergebnisse der Untersuchungen lassen sich nur auf die Grundgesamtheit der Kampfrichter im Gerätturnen übertragen. Vermutet werden kann, dass die ermittelten Erkenntnisse dieser Untersuchungen auch bei Kampfrichtertätigkeiten anderer ästhetischer Sportarten, wie beispielsweise dem Eiskunstlauf, Gültigkeit haben.

10 Ausblick

Bezug nehmend auf die ermittelten Ergebnisse der durchgeführten Untersuchungen zum Reihenfolge-Effekt und die vorgenommen Interpretationen, erfolgt eine Einschätzung der zukünftig denkbaren Forschungsarbeiten auf dem Gebiet der Urteilsbildung des Kontextes Sport.

Experimentelle Studien sind für den eher praxisbezogenen Bereich komplexer Urteilsbildung recht spärlich durchgeführt worden. Daher ergibt sich die Forderung nach weiteren experimentellen Untersuchungen unterschiedlicher Sportarten im Labor, aber auch im Feld (Greenless et al., 2007). Hauptsächlich könnte durch weitere Forschung in den technisch-kompositorischen Sportarten, die zumeist wertvolle Ratschläge für die Sportpraxis ableiten lassen, dazu beitragen, die häufig stark angezweifelte Objektivität der Leistungsurteile der Kampfrichter zu erhöhen.

Untersuchungen, die die Einschätzung der Unparteiischen über die *Stabilität der Leistungsfähigkeit* mit erheben, könnten Hinweise darüber geben, ob der Unparteiische eher Primacy-Effekt-anfällig oder Recency-Effekt-anfällig ist (Greenless et al., 2007). Je nachdem welches Leistungskonzept über die Fähigkeit beim Kampfrichter verankert ist, fällt die Tendenz in eine spezifische Reihenfolge-Effekt Richtung. Geht der Kampfrichter davon aus, dass die Leistungsfähigkeit ein recht stabiles Gebilde ist, wird er eher zum Primacy-Effekt neigen, während ein Kampfrichter mit der Annahme, die Leistungsfähigkeit sei rigide, eher einen Recency-Effekt zeigt.

Etwas allgemeiner angesetzt, könnte der Einfluss unterschiedlicher Persönlichkeitseigenschaften der Kampfrichter, wie beispielsweise die Extrovertiertheit, aber auch bezüglich der Einstellung der Kampfrichter gegenüber dem Zusammenhang bestimmter *Persönlichkeitseigenschaften* der Athleten und deren Leistungsfähigkeit, ein interessantes Forschungsfeld darstellen.

Der Faktor *Zeitdruck* könnte in experimentellen Untersuchungen eingebaut werden, um zu überprüfen, ob der Reihenfolge-Effekt eher auftritt, wenn der Unparteiische wenig Zeit für die Urteilsbildung hat.

Aber auch *weitere Einflüsse*, wie beispielsweise die Konformität, sollten in zukünftigen Studien unter kontrollierten Bedingungen miteinbezogen werden, um dem Realsetting noch ähnlicher zu sein und die kognitive Beanspruchung weiter zu steigern (Findlay & Ste-Marie, 2004). Möglich wäre, dass der Reihenfolge-Effekt eher in Situationen vorherrscht, in

denen die Kampfrichter weniger ‚unter Beobachtung' stehen und damit der Prozess der Urteilsbildung ‚natürlicher' vonstattengeht.

Die im Rahmen der Hauptfragestellung aufgestellten Vermutung, dass die präsentierten Turnelemente nicht als *separate Informationen*, sondern als eine Information wahrgenommen und verarbeitet wird, könnte durch folgende Änderung im Untersuchungsdesign überprüft werden. Eine forcierte, stärkere Verzögerung zwischen den präsentierten Informationen könnte zeigen, ob der Recency-Effekt begünstigt wird (Luchins, 1957; Miller & Campbell, 1959). Umgesetzt werden kann diese Forderung anhand von Videomanipulationen, die nicht nur einen Fehler, sondern einen Sturz in der Übung zeigen. Damit würde, zusätzlich zum wahrgenommen Fehler, eine Zeitspanne von etwa 30 Sekunden (die maximal mögliche Dauer einer Übungsunterbrechung vor einer Übungsfortführung) entstehen, die diese Verzögerung der Informationsverarbeitung darstellen würde. Über eine Variation des Sturzes zu Beginn bzw. am Ende der Übung könnte man mögliche Wertungsunterschiede aufgrund der Position des Sturzes überprüfen.

Weitere Forschungsarbeiten auf dem Gebiet der Reihenfolge-Effekt-Thematik in Urteilssituationen könnten im Bereich Gerätturnen vermehrt mit *unterschiedlich lizenzierten Kampfrichtern* durchgeführt werden. Ein Interpretationsansatz dieser Arbeit verdeutlicht, dass die unterschiedliche Zusammensetzung bzw. der hohe Anteil an international lizenzierten Kampfrichtern im Zweitexperiment zu keinem Reihenfolge-Effekt führt. Die logische Folge daraus fordert Untersuchungen mit einer Mehrzahl an Kampfrichtern unterschiedlicher Lizenzstufen, um einen Lizenzvergleich aussagekräftig durchführen zu können.

Der möglichen ‚*Vermischung' von Effekten* müsste experimentell weiter nachgegangen werden. Allerdings ist diese Forderung relativ schwer zu realisieren, bevor die einzelnen Effekte separat nicht hinreichend untersucht worden sind. So, wie zu vermuten ist, dass sich die beiden Reihenfolge-Effekte, Recency- und Primacy-Effekt, gegenseitig aufheben, könnten auch andere Urteilsverzerrungen Wechselwirkungen aufweisen. Es ist denkbar, dass sich einige Urteilsverzerrungen ausgleichen und somit aufheben, während sich andere aufsummieren und daraus unterschiedliche Fehlurteile resultieren.

Als letzte Anmerkung zur weiterführenden Forschung auf dem Gebiet der Urteilsbildung ist die empirische Überprüfung der vermuteten Effekte unterschiedlicher Autoren (u.a. Chaplan, 1990; Calkin, 1979) denkbar, die bis zum aktuellen Zeitpunkt nur auf Beobachtungen basieren.

Literaturverzeichnis

Adelman, L., Tolcott, M.A. & Bresnick, T.A. (1993). Examining the effect of information order on expert judgement. *Organizational Behavior and Human Decision Processes, 56*, 348 - 369.

Alfermann, D. & Würth, S. (2009). Gruppenprozesse und Intergruppenbeziehungen. In W. Schlicht & B. Strauß. (Hrsg.), *Enzyklopädie für Psychologie: Grundlagen der Sportpsychologie* (S. 765 – 769). Göttingen: Hogrefe.

Anderson, N. H. (1959). Test of a Model for Opinion Change. *Journal of Abnormal and Social Psychology, 59 (3),* 371 - 381.

Anderson, N.H. (1965). Primacy effects in personality impression formation using a generalized order effect paradigm. *Journal of Personality and Social Psychology, 2,* 1 - 9.

Anderson, N.H. (1974). Cognitive algebra: Integration theory applied to social attribution. In L. Berkowitz (Ed.), *Advances in experimental social psychology, 7,* 1 - 101. San Diego, CA: Academic Press.

Anderson, N.H. & Hubert, S. (1963). Effects of concomitant verbal recall on order effects in personality impression formation. *Journal of Verbal Learning and Verbal Behavior, 2,* 379 - 391.

Ansorge, C.J. & Scheer, J.K. (1988). International bias detected in judging gymnastic competition at the 1984 Olympic Games. *Research Quarterly for Exercise and Sport, 59,* 103 - 107.

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.

Aronson, E., Wilson, T.D. & Akert, R.M. (2004). *Sozialpsychologie* (4., aktualisierte Aufl.). München: Pearson Studium.

Asch, S.E. (1946). Forming impression of personality. *Journal of Abnormal and Social Psychology, 41,* 258 - 290.

Ashton, A.H. & Ashton, R.H. (1988). Sequential belief revision in auditing. *The Accounting Review*, 623 - 641.

Atkinson, R.C. & Shiffrin, R.M. (1968). Human memory. A proposed system and its control processes. In K. Spence & J. Spence (Hrsg.), *The psychology of learning and motivation* (Bd. 2). New York: Academic Press.

Balmer, N.J., Nevill, A.M. & Williams, A.M. (2001). Home advantage in the Winter Olympics (1908-1998). *Journal of Sports Sciences, 19,* 129 - 139.

Bard, C., Fleury, M., Carrière, L. & Hallé, M. (1980). Analysis of gymnastics judges visual search. *Research Quarterly for Exercise and Sport, 51,* 267 - 273.

Baumeister, R.F. & Steinhilber, A. (1984). Paradoxical effects of supportive audiences on performance under pressure: The home field disadvantage in sports championships. *Journal of Personality and Social Psychology, 47*, 85 – 93.

Bergholz, P.A. (2003). *Bewegungsfertigkeiten im Sportunterricht. Theoretische Überlegungen, Analysen und empirische Befunde zum*

fertigkeitsspezifischen Leistungsspektrum bei Schulanfängern. Dissertation. Zugriff am 18. August 2008 unter http://w210.ub.uni-tuebingen.de/volltexte/2003/1017/pdf/complete.pdf

Bergus, G.R., Levin, I.P. & Einstein, A.S. (2002). Presenting risks and benefits to patients, the effect of information order on decision making. *Journal of Internal Medicine, 17,* 612 - 617.

Bless, H. & Keller, J. (2006). Urteilsheuristiken. In H.-W. Bierhoff & D. Frey (Hrsg.), *Handbuch der Psychologie: Sozialpsychologie und Kommunikationspsychologie* (S. 294 - 300). Göttingen: Hogrefe.

Boen, F., Vanden Auweele, Y., Claes, E., Feys, J. & De Cuyper, B. (2006). The impact of open feedback on conformity among judges in rope skipping. *Psychology of Sport and Exercise, 7,* 577 - 590.

Borman, W.C. (1975). Effects of Instructions to Avoid Error on Reliability and Validity of Performance Evaluation Ratings. *Journal of Applied Psychology, 60,* 556 – 560.

Bortz, J. & Döring, N. (2003). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (3. überarbeitete Aufl.). Heidelberg: Springer Verlag.

Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler* (6. überarbeitete Aufl.). Heidelberg: Springer Verlag.

Brand, R. & Ness, W. (2004). Regelanwendung und Game-Management. Qualifizierende Merkmale von Schiedsrichtern in Sportspielen. *Zeitschrift für Sportpsychologie, 11 (4),* 127 - 136.

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

Bruine de Bruin, W. (2005). Save the last dance for me: Unwanted serial position effects on jury evaluations. *Acta Psychologica, 118,* 245 - 260.

Bruine de Bruin, W. (2006). Save the last dance II: Unwanted serial position effects in figure skating judgments. *Acta Psychologica, 123,* 299 - 311.

Bühl, A. & Zöfel, P. (2005). *SPSS 12. Einführung in die moderne Datenanalyse unter Windows* (9. überarbeitete und erweiterte Aufl.). München: Pearson Studium.

Calkin, G.F. (1979). Judging effects in men´s collegiate judging. *International Gymnast, 21 (1),* 55.

Chaplan, M. (1990). What really moves gymnastics judges? *International Gymnast, 32,* 43.

Clark-Carter, D. (1997). *Doing quantitative psychological research: From design to report.* UK: Psychology Press.

Costabile, K.A. & Klein, S.B. (2005). Finishing strong: Recency effects in juror judgements. *Basic and Applied Social Psychology, 27,* 47 - 58.

Coulomb-Cabagno, G., Rascle, O. & Souchon, N. (2005). Players' gender and male referees' decisions about aggression in French soccer : A preliminary study. *Sex Roles, 52 (7/8),* 547 - 553.

Courneya, K.S. & Carron, A.V. (1992). The home advantage in sport competitions: A literature review. J*ournal of Sport and Exercise Psychology, 14,* 13 - 27.

Craik, F.I.M. & Lockhart, R.S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11,* 671 - 684.

Cromwell, H. (1950). The relative effect on audience attitude of the first versus the second argumentative speech of a series. *Speech Monogram, 17,* 105 - 122.

Damisch, L. (2004). *Daumen drücken für den Vorgänger? Der Einfluss von Vergleichen auf das Kampfrichterurteil im Gerätturnen.* Unveröffentlichte Diplomarbeit. Bayerische Julius-Maximilians-Universität Würzburg.

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.

Deutsche Presse Agentur (dpa) (24. August 2008). „Ausraster beim Taekwondo“. *Sonntag Aktuell* (am 24. August, S. 11).

Deutscher Turnerbund (Hrsg.) (2009). *Satzung Deutscher Turnerbund – Verband für Turnen und Gymnastik – Leistungssport, Freizeit- und Gesundheitssport*. Kassel.

Ebbinghaus, H. (1885). *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie (*Unveränderte und ungekürzte Ausgabe 1992). Darmstadt: Wissenschaftliche Buchgesellschaft.

Ebersberger, H., Malka, J. & Pohler, R. (1996). *Schiedsrichter im Fußball. Ein Lehrbuch für Schiedsrichter, Trainer und Spieler.* Wiesbaden: Limpert.

Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2007). GPower 3: A flexible statistical power analysis program for the social, behavioural, and biomedical sciences. *Behaviour Research Methods, 39,* 175 - 191.

Faulkner, J. & Loken, N. (1962). A further comment on gymnastic scores. *Modern Gymnast,* 25.

Fédération Internationale de Gymnastique (Ed.) (2006). *Code de pointage - gymnastique artistique masculine*. Suisse.

Fédération Internationale de Gymnastique (Internationaler Turnerbund) (2006). *Technisches Komitee Männer – Wertungsvorschriften*. Ausgabe März.

Fenwick, I. & Chatterjee, S. (1981). Perception, Preference, and Patriotism: An Exploratory Analysis of the 1980 Winter Olympics. *The American Statistician, 35,* 170 - 173.

Fiedler, K. & Bless, H. (2002). Soziale Kognition. In: W. Stroebe, K. Jonas & M. Hewstone (Hrsg.), *Sozialpsychologie* (S. 125 – 163). Heidelberg: Springer.

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review, 107,* 659 - 676.

Findlay, L.C. & Ste-Marie, D. (2004). A reputation bias in figure skating. *Journal of Sport & Exercise Psychology, 26,* 154 - 166.

Fischer, P., Greitemeyer, T. & Frey, D. (2006). Rationalität bei Entscheidungen. In H. W. Bierhoff & D. Frey (Hrsg.), *Handbuch der Psychologie: Sozialpsychologie und Kommunikationspsychologie* (S. 273 – 279). Göttingen: Hogrefe.

Ford, G.G., Gallagher, S.H., Lacy, B.A., Bridwell, A.M. & Goodwin, F. (1997). Repositioning the home plate umpire to provide enhanced perceptual cues and more accurate ball-strike judgments. *Journal of Sport Behavior, 22,* 28 - 44.

Frank, M.G. & Gilovich, T. (1988). The dark side of self- and social perception: Black uniforms and aggression in professional sports. *Journal of Personality and Social Psychology, 54,* 74 - 85.

Gilovich, T., Vallone, R. & Tversky, A. (1985). The hot hand in Basketball. On the misperception of random sequences. *Cognitive Psychology, 17,* 295 - 314.

Göhner, U. (1992). Einführung in die Bewegungslehre des Sports – Teil 1: Die sportlichen Bewegungen. Schorndorf: Hofmann.

Greenlees, I., Buscombe, R., Thelwell, R., Holder, T. & Rimmer, M. (2005). Impact of opponents' clothing and body language on impression formation and outcome expectations. *Journal of Sport & Exercise Psychology, 27,* 39-52.

Greenlees, I.A., Hall, B., Filby, W.C.D., Thelwell, R.C., Buscombe, R. & Smith, M.J. (2009). Warnings given to observers can eliminate order effects. *Psychology of Sport & Exercise, 10(2),* 300 - 303.

Greenless, I., Dicks, M., Holder, T. & Thelwell, R. (2007). Order effects in sport: Examining the impact of order of information presentation on attributions of ability. *Psychology of Sport and Exercise, 8,* 477 - 489.

Grieve, P.G. & Hogg, M.A. (1999). Subjective uncertainty and intergroup discrimination in the minimal group situation. *Personality and Social Psychology Bulletin, 25,* 926 - 940.

Güldenpfennig, S. (1996). Philosophie der sportlichen Leistung. In Haag, H. (Hrsg.), *Handbuch Sportphilosophie* (S. 173 - 208). Schorndorf: Hofmann.

Haase, H. (1972). Die Objektivität der Bewertung komplexer sportlicher Leistungen. *Leistungssport, 2,* 346 - 351.

Hagemann, N., Strauß, B. & Leißing, J. (2008). When the referee sees red… *Psychological Science, 19,* 769 - 771.

Highhouse, S. & Gallo, A. (1997). Order effects in personnel decision making. *Human Performance, 10,* 31 - 46.

Hill, R.A. & Barton, R.A. (2005). Red enhances human performance in contests. *Nature Publishing Group, 435,* 293.

Hogarth, R.M. & Einhorn, H.J. (1992). Order effect in belief updating: The belief–adjustment model. *Cognitive Psychology, 24,* 1 - 55.

Jendrusch, G. (2002). Probleme bei der Bewegungsbeobachtung und -beurteilung durch Kampf-, Schieds- und Linienrichter [Judges', referees', and linesmen's difficulties in the perception and evaluation of movements]. *Psychologie & Sport, 9,* 133 - 144.

Jendrusch, G., Schmidt, O., Wilke, G. & de Marées, H. (1993). Zur visuellen Leistungsfähigkeit von Handball-Schiedsrichtern. In H.-F. Vaigt (Red.), *An der RUB – Sportpraxis nachgedacht, Bd 1: Bewegungen lesen und antworten* (S. 73 - 87). Ahrensburg: Czwalina.

Johnson, M. (1971). Objectivity of Judging at the National Collegiate Athletic Association Gymnastic Meet: A twenty-year follow-up study. *Research Quarterly, 42*, 454 - 455.

Jones, E.E. & Goethals, G.R. (1972). Order effects in impression formation: Attribution context and the nature of the entity. In E.E. Jones, D.E. Kanouse, H.H. Kelley, R.E. Nisbett, S. Valins & B. Weiner (Eds.), *Attribution: Perceiving the causes of behaviour* (pp. 27 - 46). Morristown, NJ: General Learning Press.

Jones, E.E. & Berglas, S. (1976). A recency effect in attitude attribution. *Journal of Personality, 44,* 433 - 448.

Jones, E.E., Rock, L., Shaver, K.G., Goethals, G.R. & Ward, L.M. (1968). Pattern of performance and ability attribution: An unexpected primacy effect. *Journal of Personality and Social Psychology, 10 (4),* 317 - 340.

Jones, M.V., Paull, G.C. & Erskine, J. (2002). The Impact of team´s aggressive reputation on the decisions of association football referees. *Journal of Sports Sciences, 20,* 991 - 1000.

Jones, R.G. & Welsh, J.B. (1971). Ability attribution and impression formation. in a strategic game: A limiting case of the primacy effect. *Journal of Personality and Social Psychology*, *20*, 166 - 175.

Jürgens, E. (2005). *Leistung und Beurteilung in der Schule. Eine Einführung in Leistungs- und Bewertungsfragen aus pädagogischer Sicht* (6. aktualisierte und stark erweiterte Aufl.). Sankt Augustin: Academia.

Kahneman, D. & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*, 430 - 454.

Kanning, U.P. (1999). *Die Psychologie der Personenbeurteilung*. Göttingen: Hogrefe.

Kruglanski, A.W. & Webster, D.M. (1996). Motivated closing of the mind: "Seizing" and "freezing". *Psychological Review, 103,* 263 - 283.

Kurz, D. (1983). Freude am Sport – sich erproben und vergleichen. In: H. Digel (Hrsg.), *Lehren im Sport. Ein Handbuch für Sportlehrer, Sportstudierende und Übungsleiter* (S. 63 – 75). Reinbek: Rowohlt.

Landers, D.M. (1970). A Review of Research on Gymnastic Judging. *Research Bulletin, September,* 85 - 88.

Lehman, D.R. & Reifman, A. (1987). Spectator influence on basketball officiating. *The Journal of Social Psychology, 127,* 673 - 675.

Luchins, A. S. (1957). Primacy-recency in impression formation. In C.I. Hovland (Ed.), The order of presentation in persuasion (pp. 33 - 61). New Haven, CT: Yale University Press.

Lund, F.H. (1925). The Psychology of Belief. The Law of Primacy in Persuasion. *Journal of Abnormal and Social Psychology, 20,* 183 - 191.

Lutter, H. (1982). Messen und Bewerten der sportlichen Leistung. In Röthig, P. & Größing, S. (Hrsg.) (1982). *Bewegungslehre*. *Kursbuch 3 für die Sporttheorie in der Schule* (1. Aufl.) (S. 93 - 120). Wiesbaden.

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 football refereeing. *International Journal of Sport Psychology, 37,* 99 - 120.

McAndrew, F.T. (1981). Pattern of performance and attributions of ability and gender. *Personality and Social Psychology Bulletin*, 7, 583 - 587.

Messner, C. & Schmid, B. (2007). Über die Schwierigkeit unparteiische Entscheidungen zu fällen: Schiedsrichter bevorzugen Fußballteams ihrer Kultur. *Zeitschrift für Sozialpsychologie, 38 (2),* 105 - 110.

Miller, N. & Campbell, D.T. (1959). Recency and primacy in persuasion as a function of the timing of speeches and measurements. *Journal of Abnormal and Social Psychology, 59,* 1 - 9.

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

Moormann, P.P. (1994). *Figure skating performance – A psychological study.* Dissertation. Leiden University.

Neumaier, A. (1988). *Bewegungsbeobachtung und Bewegungsbeurteilung im Sport* (Schriften der Deutschen Sporthochschule Köln, 21). Sankt Augustin: Academia.

Nevill, A.M., Balmer, N.J. & Williams, A.M. (2002). The influence of crowd noise and experince upon refereeing decisions in football. *Psychology of Sport and Exercise, 3,* 261 - 272.

Nieschlag, R., Dichtl, D. & Hörschgen, H. (1988). *Marketing*. Berlin: Duncker & Humbolt.

Nisbett, R. & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, N.J.: Prentice Hall.

O´Brien, K. (1991). Bias in the judging of international elite gymnasts. In J. Standeven, K. Hardman & D. Fisher (Eds.), *Sport for all: Into the 90s* (pp. 148 - 153). Aachen: Meyer & Meyer.

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

Pflughoeft, M. (1984). Gymnastics Judging. Fatique, frustration, confusion. *International Gymnast, 26,* 40.

Pieters, R.G.M. & Bijmolt, T.H.A. (1997). Consumer memory for television advertising: A field study of duration, serial position, and competition effects. *Journal of Consumer Research, 23,* 362 – 372.

Plessner, H. & Raab, M. (1999). Kampf- und Schiedsrichterurteile als Produkte sozialer Informationsverarbeitung. *Psychologie & Sport, 6,* 130 - 145.

Plessner, H. & Schallies, E. (2005). Judging the cross on rings: A matter of achieving shape constancy. *Applied Cognitive Psychology, 19,* 1145 - 1156.

Plessner, H. (1997). *Urteilsverzerrungen bei Kampfrichtern im Gerätturnen – Der Einfluß von Erwartungen*. Aachen: Shaker.

Plessner, H. (1999). Expectation biases in gymnastics judging. *Journal of Sport and Exercise Psychology, 21,* 131 - 144.

Plessner, H. (2001a). Empfehlungen für die Praxis. *Leon, 5,* 30 - 32.

Plessner, H. (2001b). Ist wirklich wahr, was Juroren als wahr wahrnehmen? *Leon, 4,* 30 - 32.

Plessner, H. (2004). Irren ist menschlich! *Der Handball-Schiedsrichter*, *2*, 2 - 7.

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

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

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

Prohl, P. (2003a). Leistung. In P. Röthig & R. Prohl (Hrsg.), *Sportwissenschaftliches Lexikon* (S. 332 - 337). Schorndorf: Hofmann.

Prohl, P. (2003b). Sportliche Leistungsbewertung. In P. Röthig & R. Prohl (Hrsg.), *Sportwissenschaftliches Lexikon* (S. 337). Schorndorf: Hofmann.

Puhl, J. (1980). Use of video replay in judging gymnastics vaults. *Perceptual and Motor Skills, 51,* 51 - 54.

Rheinberg, F. (2001). Bezugsnormen und schulische Leistungsbeurteilung. In: F. E. Weinert (Hrsg.): *Leistungsmessung in Schulen* (S. 59 - 71). Weinheim: Beltz.

Richter, L. & Kruglanski, A.W. (1998). Seizing on the Latest: Motivationally Driven Recency Effects in Impression Formation. *Journal of Experimental Social Psychology, 34,* 313 – 329.

Salmela, J.H. (1978). Gymnastics judging: A complex information processing task, or (who's putting one over on who?) Part 1 & 2. *International Gymnast, 20,* 54 - 56 & 62 - 63.

Scheer, J.K. & Ansorge, C.J. (1975). Effects of naturally induced judges' expectations on the ratings of physical performances. *Research Quarterly, 46,* 463 - 470.

Scheer, J.K. & Ansorge, C.J. (1979). Influence due to expectations of judges: A function of internal-external locus of control. *Journal of Sport Psychology, 1,* 53 - 58.

Scheer, J.K. & Ansorge, C.J. (1980). Expectations in judging. *International Gymnast Technical Supplement, 1,* 1 - 2.

Scheer, J.K. (1973). Effect of placement in the order of competition on scores of Nebraska high school students. *The Research Quarterly, 44,* 79 - 85.

Scheer, J.K., Ansorge, C.J. & Howard, J. (1983). Judging bias induced by viewing contrived videotapes: A function of selected psychological variables. *Journal of Sport Psychology, 5,* 427-437.

Schlicht, W. & Strauß, B. (2003). *Sozialpsychologie des Sports - ein Lehrbuch*. Göttingen: Hogrefe.

Sedlmeier, P. (1996). Jenseits des Signifikanztest-Rituals: Ergänzungen und Alternativen. *Methods of Psychological Research Online, 1(4),* 41 - 63.

Seltzer, R. & Glass, W. (1991). International politics and judging in Olympic skating events. *Journal of Sport Behavior*, 14, 189 - 200.

Slovic, P., Fischhoff, B. & Lichtenstein, S. (1977). Behavioral decision theory. *Annual Review of Psychology, 28,* 1 - 39.

Souchon N., Coulomb-Cabagno G., Traclet A. & Rascle O. (2004). Referees' decision-making in handball and transgressive behaviours: influence of stereotypes about gender of players? *Sex Ro-l*es, *51,* 445 - 453.

Stahel, W. (2007). *Statistische Datenanalyse. Eine Einführung für Naturwissenschaftler* (5. überarbeitete Aufl.). Braunschweig: Vieweg.

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

Ste-Marie, D. & Valiquette, S.M. (1996). Enduring memory-influenced biases in gymnastic judging. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1498 - 1502.

Ste-Marie, D. (1996). International bias in gymnastic judging: Conscious or unconscious influences? *Perceptual and Motor Skills, 83,* 963 - 975.

Ste-Marie, D. (1999). Expert-novice differences in gymnastic judging: An information processing perspective. *Applied Cognitive Psychology, 13,* 269 - 281.

Ste-Marie, D. (2000). Expertise in women's gymnastic judging: An observational approach. *Perceptual and Motor Skills, 90,* 543 - 546.

Ste-Marie, D. (2003). Memory biases in gymnastic judging: Differential effects of surface feature changes. *Applied Cognitive Psychology, 17,* 733 - 751.

Ste-Marie, D., Valiquette, G. & Taylor, G. (2001). Memory-influenced biases in gymnastic judging occur across different prior processing conditions. *Research Quarterly for Exercise and Sport, 72,* 420 - 426.

Stevens, J. (1999). *Intermediate Statistics. A Modern Approach.* London: Erlbaum.

Strauss, B. (1999). Wenn Fans ihre Mannschaft zur Niederlage klatschen: Zuschauer und sportliche Leistungen. *Sportwissenschaft, 29,* 393 – 411.

Sutter, M. & Kocher, M.G. (2004). Favoritism of agents - The case of referees' home bias. *Journal of Economic Psychology, 25,* 461 - 469.

Taha, A., Osman, M. & Ehlerz, M. (1991). Technikbeurteilung zur Objektivierbarkeit der Wertung im Gerätturnen. In R. Daugs (Hrsg.), *Sportmotorisches Lernen und Techniktraining* (S. 266 - 269). Schorndorf: Hoffmann.

Taylor, S.E. & Fiske, S.T. (1981). Getting inside the head: Methodologies for process analysis in attribution and social cognition. In J.H. Harvey, W. Ickes, & R.F. Kidd (Eds.), *New directions in attribution research* (Vol. 3, pp. 459 - 524). Hillsdale, N.J.: Lawrence Erlbaum.

Teipel, D. (2003a). Kampfrichter. In P. Röthig & R. Prohl (Hrsg.), *Sportwissenschaftliches Lexikon* (S. 287 - 288). Schorndorf: Hofmann.

Teipel, D. (2003b). Schiedsrichter. In P. Röthig & R. Prohl (Hrsg.), *Sportwissenschaftliches Lexikon* (S. 460 - 461). Schorndorf: Hofmann.

Thieß, G. & Tschiene, P. (1999). *Handbuch zur Wettkampfl*ehre. Aachen: Meyer und Meyer.

Thomas, A. (1978). *Einführung in die Sportpsychologie*. Göttingen: Hogrefe.

Thorndike, E.L. (1920). A Constant Error in Psychological Ratings. *Journal of Applied Psychology, 4,* 25 - 29.

Tversky, A. & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 207 - 232.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124 - 1131.

Vanden Auweele, Y., Boen, F., De Geest, A. & Feys, J. (2004). Judging bias synchronized swimming: Open feedback leads to nonperformance-based conformity. *Journal of Sport & Exercise Psychology, 26,* 561 - 571.

Wanderer, J.J. (1987). Social factors in judges' rankings of competitors in figure skating championships. *Journal of Sport Behavior, 10,* 93 - 102.

Webster, D.M, Richter, L. & Kruglanski, A.W. (1996). On leaping to conclusions when feeling tired: Mental fatigue effects on impressional primacy. *Journal of Experimental Social Psychology, 32,* 181 - 195.

Weigelt, M. (2006). Nature Neuroscience 10/2005 – Die eigene Bewegungsbibliothek bildet die Grundlage, Handlungen anderer Menschen besser zu verstehen. Sportpsychologie-Digest. *Zeitschrift für Sportpsychologie, 13 (3),* 115 - 117.

Wells, F.L. (1907). A Statistical Study of Literary Merit. *Archives of Psychology, 1,* 1 - 30.

Wilson, V.E. (1976a). Judging gymnastic judging. In J.H. Salmela (Ed.), *The advanced study of gymnastics*. Chapter 10 (pp. 151 – 166). Springfield, IL: Charles C. Thomas Publisher.

Wilson, V.E. (1976b). Objectivity, validity and reliability of gymnastic judging. *Research Quarterly, 47,* 169 - 174.

Wilson, V.E. (1977). Objectivity and effect of order of appearance in judging synchronized swimming meets. *Perceptual and Motor Skills, 44,* 295 - 298.

Wyer, R.S. (1973). Effects of information inconsistency and grammatical context on evaluation of persons. *Journal of Personality and Social Psychology, 25,* 45 - 49.

Zimbardo, P.G. (1995). *Psychologie* (6. Aufl.). Berlin: Springer.

Anhang

A: Verwendete Fragebögen der Untersuchung

1. Bewertungsbogen – Version „end of sequence“ (des Erstexperiments)

Kampfrichter-Studie Bewertungs-Papierbogen 1

Gerät/Übung	B-Note
Reck/Turner 1	
Reck/Turner 2	
Reck/Turner 3	
Reck/Turner 4	
Ringe/Turner 1	
Ringe/Turner 2	
Ringe/Turner 3	
Ringe/Turner 4	

Proband Nr.:

Wird vom Versuchsleiter ausgefüllt!

2. Bewertungsbogen – Version „step by step“ (des Erstexperiments)

Kampfrichter-Studie Bewertungs-Papierbogen 1

Gerät/Übung	Notierte Abzüge	B-Note
Reck/Turner 1		
Reck/Turner 2		
Reck/Turner 3		
Reck/Turner 4		
Ringe/Turner 1		
Ringe/Turner 2		
Ringe/Turner 3		
Ringe/Turner 4		

Proband Nr.:

Wird vom Versuchsleiter ausgefüllt!

3. A. PERSONENFRAGEBOGEN – SEITE 1

Kampfrichter-Studie Personeninformationen 1

Um die Daten optimal auswerten zu können, möchte ich Sie bitten noch einige Angaben zu Ihrer Person zu machen. Wie bereits erwähnt, werden sämtliche Informationen aus dieser Untersuchung anonym behandelt.

Proband Nr.: ☐

Wird vom Versuchsleiter ausgefüllt!

Bitte beantworten Sie nun einige kurze Fragen zu Ihrer Person:

☞ **Frage 1:**

Ihr Geschlecht:

weiblich	☐
männlich	☐

☞ **Frage 2:**

Wie alt sind Sie?
Bitte die Zahl ins Kästchen schreiben. ☐

☞ **Frage 3:**

Welchen Beruf üben Sie aus?
Bitte ins Kästchen schreiben. ☐

☞ **Frage 4:**

Wie lange sind Sie schon als Kampfrichter tätig?
Bitte die Zahl in Jahren ins Kästchen schreiben. ☐

☞ **Frage 5:**

Für welche Nation sind Sie als Kampfrichter tätig?
Bitte das intern. Nationenkürzel ins Kästchen schreiben. ☐

☞ **Frage 6:**

Welche Kampfrichter-Lizenz haben Sie aktuell? *(nur 1 Kreuz)*

Internationale Lizenz	☐
Lizenzstufe A	☐
Lizenzstufe B	☐
Lizenzstufe C	☐
Lizenzstufe D	☐

3. B. PERSONENFRAGEBOGEN – SEITE 2

Kampfrichter-Studie Personeninformationen 2

☞ **Frage 7a:**

Waren oder sind Sie selbst als Turner aktiv? Ja ☐ Nein ☐

☞ **Frage 7b:**

Wenn ja, wie lange?
Bitte die Zahl in Jahren ins Kästchen schreiben.

☞ **Frage 8a:**

Waren oder sind Sie selbst als Trainer tätig? Ja ☐ Nein ☐

☞ **Frage 8b:**

Wenn ja, wie lange?
Bitte die Zahl in Jahren ins Kästchen schreiben.

☞ **Frage 9:**

Bei wie vielen Wettkämpfen waren Sie im vergangenen Jahr als Kampfrichter tätig?
Bitte die Zahl ins Kästchen schreiben.

☞ **Frage10:**

War Ihnen einer oder gar mehrere Turner aus den Videosequenzen bekannt? Ja ☐ Nein ☐

☞ **Frage 11:**

Haben Sie irgendwelche Anmerkungen zu der soeben durchgeführten Untersuchung? Wenn ja, welche?

Oder: Ist Ihnen irgend etwas Besonderes aufgefallen? Wenn ja, was?

Herzlichen Dank für Ihre Mitarbeit!

4. Tabellarische Darstellungen im Anhang

Tabelle 19: Darstellung der varianzanalytischen Kennwerte des Erstexperiments (Einwertübungen nicht aufgeführt)

Quelle	df	F	Sig.	eta2	Teststärke
Ringe, Turner 2					
Korrigiertes Modell	3	0,862	0,463	0,026	0,232
Reihenfolge	1	0,693	0,407	0,007	0,131
Bewertung	1	0,942	0,334	0,010	0,161
Reihenfolge * Bewertung	1	1,307	0,256	0,013	0,205
Ringe, Turner 3					
Korrigiertes Modell	3	0,930	0,429	0,028	0,248
Reihenfolge	1	0,643	0,424	0,007	0,125
Bewertung	1	1,946	0,166	0,020	0,282
Reihenfolge * Bewertung	1	0,291	0,591	0,003	0,083
Ringe, Turner 4					
Korrigiertes Modell	3	2,458	0,067	0,071	0,595
Reihenfolge	1	0,206	0,651	0,002	0,073
Bewertung	1	7,282	0,008	0,070	0,762
Reihenfolge * Bewertung	1	0,017	0,896	0,000	0,052
Reck, Turner 2					
Korrigiertes Modell	3	3,930	0,011	0,108	0,816
Reihenfolge	1	0,092	0,762	0,001	0,060
Bewertung	1	6,991	0,010	0,067	0,745
Reihenfolge * Bewertung	1	4,462	0,037	0,044	0,552
Reck, Turner 3					
Korrigiertes Modell	3	1,001	0,396	0,031	0,265
Reihenfolge	1	2,335	0,130	0,024	0,328
Bewertung	1	0,951	0,332	0,010	0,162
Reihenfolge * Bewertung	1	0,125	0,725	0,001	0,064
Reck, Turner 4					
Korrigiertes Modell	3	1,096	0,355	0,033	0,288
Reihenfolge	1	0,318	0,574	0,003	0,086
Bewertung	1	0,243	0,623	0,003	0,078
Reihenfolge * Bewertung	1	2,789	0,098	0,029	0,380

Tabelle 20: Darstellung der varianzanalytischen Kennwerte des Zweitexperiments (Einwertübungen nicht aufgeführt)

Quelle	df	F	Sig.	eta2	Teststärke
Ringe, Turner 2					
Korrigiertes Modell	3	0,202	0,895	0,006	0,087
Reihenfolge	1	0,133	0,716	0,001	0,065
Bewertung	1	0,448	0,505	0,004	0,102
Reihenfolge * Bewertung	1	0,004	0,950	0,000	0,050
Ringe, Turner 3					
Korrigiertes Modell	3	2,007	0,118	0,056	0,503
Reihenfolge	1	0,062	0,805	0,001	0,057
Bewertung	1	2,429	0,122	0,023	0,339
Reihenfolge * Bewertung	1	3,741	0,056	0,035	0,483
Ringe, Turner 4					
Korrigiertes Modell	3	0,778	0,509	0,022	0,212
Reihenfolge	1	0,126	0,724	0,001	0,064
Bewertung	1	0,055	0,814	0,001	0,056
Reihenfolge * Bewertung	1	2,079	0,152	0,020	0,298
Ringe, Turner 5					
Korrigiertes Modell	3	1,102	0,352	0,031	0,290
Reihenfolge	1	1,509	0,222	0,015	0,229
Bewertung	1	1,601	0,209	0,015	0,241
Reihenfolge * Bewertung	1	0,003	0,955	0,000	0,050
Ringe, Turner 6					
Korrigiertes Modell	3	1,120	0,345	0,032	0,294
Reihenfolge	1	0,065	0,799	0,001	0,057
Bewertung	1	1,756	0,188	0,017	0,259
Reihenfolge * Bewertung	1	1,537	0,218	0,015	0,233
Reck, Turner 2					
Korrigiertes Modell	3	0,167	0,918	0,005	0,080
Reihenfolge	1	0,447	0,505	0,004	0,102
Bewertung	1	0,012	0,911	0,000	0,051
Reihenfolge * Bewertung	1	0,020	0,889	0,000	0,052
Reck, Turner 3					
Korrigiertes Modell	3	1,052	0,373	0,030	0,278
Reihenfolge	1	0,379	0,540	0,004	0,094
Bewertung	1	2,186	0,142	0,021	0,310
Reihenfolge * Bewertung	1	0,436	0,511	0,004	0,100

Quelle	df	F	Sig.	eta2	Teststärke
Reck, Turner 4					
Korrigiertes Modell	3	0,958	0,416	0,028	0,255
Reihenfolge	1	0,010	0,920	0,000	0,051
Bewertung	1	2,134	0,147	0,021	0,304
Reihenfolge * Bewertung	1	0,764	0,384	0,008	0,139
Reck, Turner 5					
Korrigiertes Modell	3	1,211	0,310	0,035	0,316
Reihenfolge	1	0,997	0,321	0,010	0,167
Bewertung	1	2,393	0,125	0,023	0,335
Reihenfolge * Bewertung	1	0,002	0,964	0,000	0,050
Reck, Turner 6					
Korrigiertes Modell	3	1,579	0,199	0,045	0,405
Reihenfolge	1	0,456	0,501	0,005	0,103
Bewertung	1	3,951	0,050	0,038	0,503
Reihenfolge * Bewertung	1	0,151	0,699	0,002	0,067