pedocs
Open Access Erziehungswissenschaften

DIPF

Fiedler, Klaus; Plessner, Henning

Die Stichprobenfalle: Lässt sich eine Sensibilität für metakognitive Probleme beim stochastischen Denken vermitteln?

Unterrichtswissenschaft 32 (2004) 1, S. 23-37

urn:nbn:de:0111-opus-58061

in Kooperation mit / in cooperation with:

BELTZ JUVENTA

http://www.juventa.de

Nutzungsbedingungen / conditions of use

Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Die Nutzung stellt keine Übertragung des Eigentumsrechts an diesem Dokument dar und gilt vorbehaltlich der folgenden Einschränkungen: Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.

We grant a non-exclusive, non-transferable, individual and limited right to using this document.
This document is solely intended for your personal, non-commercial use. Use of this document does not include any transfer of property rights and it is conditional to the following limitations: All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.

Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.
By using this particular document, you accept the above-stated conditions of use.

Kontakt / Contact:

peDOCS
Deutsches Institut für Internationale Pädagogische Forschung (DIPF)
Mitglied der Leibniz-Gemeinschaft
Informationszentrum (IZ) Bildung
Schloßstr. 29, D-60486 Frankfurt am Main
E-Mail: pedocs@dipf.de
Internet: www.pedocs.de

Digitalisiert durch DIPF

Unterrichtswissenschaft

Zeitschrift für Lernforschung
32. Jahrgang / 2004 / Heft 1

Thema
Stochastisches Denken

Verantwortliche Herausgeber
Jürgen Baumert, Gerd Gigerenzer, Laura Martignon

Klaus Fiedler, Henning Plessner

Die Stichprobenfalle: Lässt sich eine Sensibilität für metakognitive Probleme beim stochastischen Denken vermitteln?

The Sampling-Trap: Can Awareness for Metacognitive Problems in Stochastic Thinking be Taught?

Immer wieder zeigt sich im Bereich von diagnostischen Problemen, dass Menschen dazu neigen, die Wahrscheinlichkeit von Ereignissen mit einer niedrigen Basisrate zu überschätzen. Um diesen Fehleinschätzungen vorzubeugen, wird häufig vorgeschlagen, Wahrscheinlichkeitsprobleme in einem Häufigkeitsformat zu präsentieren, da dieses der Natur der menschlichen Informationsverarbeitung eher entsprechen würde. Als eine unter Umständen viel weitergehende Ursache von systematischen Fehleinschätzungen von Wahrscheinlichkeiten hat sich jedoch die mangelnde Sensibilität von Menschen für mögliche Verzerrungen in Beobachtungsstichproben erwiesen, die den erfahrenen Häufigkeiten im Alltag zu Grunde liegen. In einem Experiment mit 40 Trainern aus verschiedenen Mannschaftssportarten wird demonstriert, wie diese mangelnde Sensibilität zur starken Überschätzung der Spielstärke eines Fußballspielers führen kann. Zusätzlich wird gezeigt, dass eine Sensibilisierung der Versuchsteilnehmer für die Stichprobenproblematik zu einer Urteilskorrektur beiträgt. Möglichkeiten für die Schulung einer „Stichprobensensibilität" werden diskutiert.

In the area of diagnostic problem solving, it has repeatedly been demonstrated that people overestimate the conditional probability of low base rate events. In order to prevent these biases it has been proposed that such problems should be presented in a natural frequency format instead of a probability format. This would match the nature of human information processing. Recent studies, however, have found that an even more basic cause for the biased estimation of probabilities may lie in the fact that people are unaware of the biases that can be already inherent in the samples of everyday observations on which they base their inferences. In an experiment with 40 coaches of various team sports, we could show that this lack of awareness can lead to an overestimation of a football player's performance. Additionally, we found that making people aware of the sampling

process can instigate a correction process. Several ideas about how to teach awareness of sampling constraints are discussed.

1. Einleitung

Historisch und entwicklungsgeschichtlich gesehen stellen Stochastik und Statistik recht neue Errungenschaften dar. Während es Medizin, Mathematik, Physik und Philosophie bereits seit Jahrtausenden gibt, taucht statistisches Denken und Wahrscheinlichkeitsrechnung als wissenschaftliche Disziplin erst seit etwa drei Jahrhunderten auf (Bernoulli, 1713). Auch ontogenetisch wird das Denken in Wahrscheinlichkeiten als eine sehr fortgeschrittene Leistung der Intelligenz betrachtet. In Inhelder und Piaget (1958) „Entwicklung der menschlichen Intelligenz" gilt die Fähigkeit, nicht nur über gegebene Ereignisse sondern auch über mögliche oder wahrscheinliche Ereignisse nachzudenken, als Inbegriff der höchsten Stufe der Intelligenzentwicklung, die durch formale Denkoperationen gekennzeichnet ist.

Im Gegensatz zu diesem Ruf des stochastischen Denkens als neuzeitliche und hoch entwickelte Form der Intelligenz, die spontan mit dem Zeitalter von Computern und Datenbanken verknüpft wird, steht jedoch die Einsicht, dass stochastische Aufgaben und Umweltanforderungen sehr alt sein müssen, älter sein müssen als die Menschheit selbst. Denn die Welt, in der wir leben, folgt meist nicht eindeutig vorhersagbaren deterministischen, sondern probabilitischen Gesetzen. Das Wetter und die Naturgewalten, das Verhalten biologischer Feinde, das Auftreten von Krankheiten und die Verbreitung von Erregern, der Ertrag des Jagens und Sammelns, der Ausgang von Kämpfen, das Überleben oder die Zahl der gezeugten Nachkommen hängen allesamt von Prozessen ab, in die Zufallsvariablen eingehen - und die somit die Definition von stochastischen Prozessen erfüllen.

Dies gilt natürlich um so mehr für die neuen Anforderungen der modernen Informationsgesellschaft, die sich unter anderem dadurch auszeichnet, dass stochastische Prozesse so ausführlich wie niemals zuvor erfasst und sichtbar gemacht werden, weil nahezu alle bedeutsamen Daten - über Wetter, Epidemien, Risiken, Unfälle, Leistungen, Bevölkerung, Ökonomie, Ökologie, Sport und Politik - registriert und durch stochastische Modelle beschrieben werden. So gehört zur Gemeinbildung des modernen Menschen der informierte Umgang mit Krankheitsgefahren wie AIDS, SARS oder BSE, die Einschätzung von Risiken, ein Verständnis für die Börse und für komplexe wirtschaftliche Zusammenhänge, die Einschätzung des Arbeitsmarktes und der Bildungschancen der eigenen Kinder, die Gefahr von Kriminalität, das Abschließen von Versicherungen oder das Verstehen und Interpretieren schwacher kausaler Zusammenhänge, wie beispielsweise zwischen politischen Maßnahmen und Wohlergehen, zwischen Ernährung und Gesundheit oder zwischen Erziehungspraktiken und dem Wohlverhalten der Kinder.

Vermutlich spielt für den angepassten Umgang mit diesen alten und neuen Anforderungen der sozialen und physikalischen Umwelt das logisch-deduktive Denken, welches in Schule und Bildung traditionell einen zentralen Stellenwert hat, eine weitaus geringere Rolle als das stochastisch-induktive Denken. Denn die idealisierenden Annahmen logischer Denkmodelle sind für viele reale Anwendungen zu stark. So werden logische Regeln wie Symmetrie ($a > b \Rightarrow b < a$), Transitivität ($a>b, b>c \Rightarrow a>c$) oder Kontextunabhängigkeit ($a > x$ mit $x \in \{x_1,x_2,x_3\} \Rightarrow a > x$ mit $x \in \{x_1,x_2\}$) zum Beispiel bei wirtschaftlichen Entscheidungen häufig verletzt (z.B. Huber & Puto, 1983). Das Erschließen von induktiv beobachteten, stochastischen Regelmäßigkeiten im Alltag scheint indessen der Normalfall der angepassten Intelligenz zu sein. Märkte reagieren sehr sensibel auf kleine statistische Preisschwankungen, Kleinkinder lernen aufgrund von mehr oder weniger unsystematischen Reaktionen der Erziehungspersonen, was sie dürfen und was nicht, und sogar niedere Organismen lernen recht zuverlässig, mit welcher Wahrscheinlichkeit an verschiedenen Plätzen Nahrung zu finden ist. Die neuronale Ausstattung, die erforderlich ist, um die Auftretensrate von Umweltreizen zu unterscheiden, ist relativ bescheiden (Dougherty, Gettys, & Ogden, 1999; Fiedler, 1996), und so verblüfft es nur Laien, aber nicht wirklich Experten der kognitiven Psychologie, wie genau und nahezu mühelos Organismen Häufigkeiten erfassen und unterscheiden können. Menschen können zum Beispiel die Häufigkeit, mit der Stimuli (wie z.B. Wörter einer Sprache) in der Umwelt vorkommen, erstaunlich genau unterscheiden, auch wenn sie niemals bewusst darauf achten (Hasher & Zacks, 1984; Sedlmeier, Hertwig & Gigerenzer, 1998).

Trotz dieser spontanen, fast automatischen Fähigkeit, stochastische Information zu erfassen und zu quantifizieren, belegen auf der anderen Seite psychologische Untersuchungen schwerwiegende Denkfehler im Umgang mit Wahrscheinlichkeitsproblemen (Fiedler, 2000; Rulon, 1941). Ein konkretes und lebensnahes Beispiel aus der medizinischen Diagnostik mag dies verdeutlichen, bezogen auf die Wahrscheinlichkeit, dass jemand, der HIV positiv getestet wird, tatsächlich die Krankheit AIDS hat. Während Laien wie Experten, nach allen verfügbaren Erfahrungen, im Falle eines positiven HIV-Befundes sich subjektiv so gut wie sicher sind, dass AIDS vorliegen muss, erklären Swets, Dawes und Monahan (2000), warum die bedingte Wahrscheinlichkeit p(AIDS/HIV-positiv) tatsächlich viel geringer sein kann als erwartet.

Um das zu verstehen, muss man zunächst wissen, dass es in den USA auf dem medizinischen Markt zwei etablierte HIV-Tests gibt, die im übrigen biochemisch unabhängig funktionieren, das heißt, sie spiegeln nicht dieselben Fehler wider. Mithilfe des Bayes-Theorem der Wahrscheinlichkeitsrechnung kann man folgendes zeigen. Wenn eine zufällig ausgewählte Person - bei der weder Symptome noch irgendein anderer Krankheitsverdacht vorliegen - mit einem der beiden Tests HIV-positiv getestet wird, dann füh-

ren Mediziner in der Regel zunächst zur Kontrolle das andere, unabhängige Verfahren durch. Wenn dann dieselbe Person mit dem zweiten Test erneut positiv getestet wird, so ist die a posteriori Wahrscheinlichkeit p(AIDS/ doppelt positiv HIV getestet) dennoch nicht höher als sage und schreibe etwa 15%. - Wie kann es sein, dass unsere ansonsten so sensible Ausstattung für die Erfassung von Wahrscheinlichkeiten uns bei derartig wichtigen Problemen so in die Irre führt?

Die neuere psychologische Forschung hat verschiedene Lösungsmöglichkeiten für dieses Rätsel zu bieten, warum Wahrscheinlichkeitsurteile einerseits so mühelos und genau erscheinen und andererseits so stark fehlgeleitet sein können. Eine mögliche Antwort könnte lauten, dass die Intelligenz domänenspezifisch ist; das heißt, dass wir stochastisches Denken nur auf vertraute Inhalte anwenden können, während es in unvertrauten, künstlichen Domänen nicht funktioniert. So lernen wir durch wiederholte Erfahrung in unserer vertrauten Umwelt, mit welchen Mitteln man mehr Erfolg hat, wenn man jemanden um einen Gefallen bittet. Die meisten Menschen - einige Virologen ausgenommen - haben jedoch wenig Gelegenheit zum Lernen in der künstlichen technischen Umwelt der AIDS-Diagnostik.

Aber wie schon gesagt zeigen Experten (wie Radiologen oder Virologen) bei solchen Problemen die selben groben Fehleinschätzungen von Wahrscheinlichkeiten, so dass die Idee der Domänspezifität oder Erfahrung mit bestimmten Inhaltsbereichen alleine die verzerrten Urteile kaum erklären kann. Betrachten wir daher eine andere Erklärung, die hauptsächlich von Gigerenzer und Kollegen (z.B., Gigerenzer & Hoffrage, 1995) angeboten wird. Demnach können Menschen (ebenso wie niedere Organismen) auch mit neuartigen Problemen wie AIDS dann angemessen umgehen, wenn die Information frequentistisch präsentiert wird (also in Form von kardinalen Häufigkeiten), nicht aber, wenn die Information probabilistisch präsentiert wird (in Form von Wahrscheinlichkeiten). Denn was Menschen seit der Zeit der Jäger und Sammler effektiv gelernt haben - und lernen mussten, um zu überleben - ist das Abzählen, wie häufig auf bestimmte Signale eine schmerzhafte oder aber eine harmlose Konsequenz folgt. Mit normierten Wahrscheinlichkeiten umzugehen, fällt ihnen hingegen weiterhin schwer. Nehmen wir beispielsweise die folgenden Information über den Zusammenhang zwischen Mammographie und Brustkrebs:

- Die Grundwahrscheinlichkeit (Basisrate) von Brustkrebs in der Population aller Frauen ab vierzig Jahren beträgt $p(B) = 1\%$
- Unter Frauen, die Brustkrebs haben, beträgt die Wahrscheinlichkeit, dass ein Mammogramm positiv ist $p(M+/B) = 80\%$
- Unter Frauen, die nicht Brustkrebs haben, beträgt die Wahrscheinlichkeit, dass ein Mammogramm irrtümlich positiv ist $p(M+/\neg B) = 9.6\%$.

Angesichts dieser Daten schätzen die meisten Untersuchungsteilnehmer - Laien wie Experten - die Wahrscheinlichkeit p(B/M+) sehr hoch ein, also die Wahrscheinlichkeit, dass eine zufällig aus der Bevölkerung ausgewählte Frau über 40 Brustkrebs hat, wenn sie in einer mammografischen Untersuchung positiv getestet wurde. Die Schätzungen liegen typischerweise zwischen 50% und 80%. Kaum jemand erkennt, dass die korrekte Wahrscheinlichkeit gemäß dem Bayes-Theroem nicht höher ist als

$$P(B/M+) = p(B)\cdot p(M+/B) / [p(B)\cdot p(M+/B) + p(\neg B)\cdot p(M+/\neg B)]$$
$$= .01 \cdot 80\% / [.01 \cdot 80\% + .99 \cdot 9.6\%] \quad = 7.8\%$$

Diese fast unglaublich niedrige Lösung des Problems ist jedoch relativ leicht zu verstehen, wenn dieselben Informationen als Häufigkeiten gegeben sind:

- In einer Population von 1000 Frauen von etwa 40 Jahren haben insgesamt 10 Brustkrebs (entspricht der Basisrate von 1%)
- Unter den 10 Frauen mit Brustkrebs haben 8 ein positives Mammogramm (entspricht der Wahrscheinlichkeit von 80%)
- Unter den 990 Frauen ohne Brustkrebs haben 95 ein positives Mammogramm (entspricht der Irrtumsrate von 9.6%).

Bei dieser Darstellung derselben Information in Form von kardinalen Häufigkeiten ist leicht zu erkennen - und das gilt auch für die meisten Teilnehmer an solchen Experimenten - dass unter allen Frauen mit einem positiven Mammogramm (nämlich 8 + 95 = 103) nur eine kleine Minderheit wirklich Brustkrebs hat (nämlich 8/103 ≅ etwas weniger als 8%). Das Beispiel illustriert somit eindringlich, dass Häufigkeiten einfacher zu bearbeiten sein können als Wahrscheinlichkeiten, und zwar nicht nur bei vertrauten, entwicklungsgeschichtlich alten Aufgaben (aus der Zeit der Jäger und Sammler), sondern auch bei modernen Inhalten wie medizinische Diagnostik.

Aus diesen und vielen ähnlichen Befunden (Gigerenzer, Hoffrage, & Ebert, 1998; Hoffrage & Gigerenzer, 1998; Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2000; Sedlmeier, 1999) könnte man - didaktisch und für viele professionelle Entscheidungen in Medizin, Ökonomie und Politik - die Folgerung ableiten, dass ein simples aber hinreichendes Mittel für den Umgang mit alltäglichen Wahrscheinlichkeitsproblemen darin besteht, die Eingangsdaten als natürliche Häufigkeiten statt als missverständliche Wahrscheinlichkeiten zu präsentieren. Entsprechend müsste die Behandlung und Übung stochastischer Aufgaben im Unterricht konzipiert werden (vgl. Hoffrage et al., 2000; Lindsay, Hertwig & Gigerenzer, in press). Unsere eigenen rezenten Experimente (Fiedler, Brinkmann, Betsch & Wild, 2000; Plessner, Hartmann, Hohmann & Zimmermann, 2001) zeigen jedoch, dass mit dieser Empfehlung die Schwierigkeit im Umgang mit vielen realen Wahrscheinlichkeitsproblemen nur zum Teil gelöst werden kann. Auch wenn alle Information in völlig natürlichem Häufigkeitsformat gesammelt

wird, können immer noch schwerwiegende Verzerrungen der Urteile und Entscheidungen auftreten und zwar immer dann, wenn die vorliegenden Beobachtungen hinsichtlich des Urteilskriteriums eine selektive Stichprobe darstellen.

Verdeutlichen wir diese Situation - im Folgenden kurz *kriterium-selektive Stichprobe* genannt - an demselben Beispiel, welches zuvor schon verwendet wurde. In der Untersuchung von Fiedler et al. (2000, Experiment 2 und 3) hatten die Teilnehmer selbst die Möglichkeit, Beobachtungen über den Zusammenhang von Brustkrebs (B) und Mammographie (M+ und M–) zu sammeln, indem sie aus einer Datenbank eine beliebige Anzahl von Fällen ziehen konnten. Die Datenbank war gemäß den oben genannten statistischen Parametern zusammengestellt, das heißt, es gab insgesamt nur eine kleine Basisrate von B-Fällen, der Anteil von M+ war unter den (wenigen) B-Fällen sehr hoch und unter den (häufigen) ¬B-Fällen recht gering aber deutlich höher als 0. Die Teilnehmer konnten so lange Beobachtungen sammeln, bis sie das Gefühl hatten, ein informiertes Urteil über p(B/M+) abgeben zu können. Durch diese selbständige Informationssuche war in allen experimentellen Bedingungen ein natürliches Häufigkeitsformat garantiert; die Teilnehmer mussten niemals mit künstlich normierten Werten im Bereich 0 bis 1 umgehen, sondern lediglich mit der Anzahl der beobachteten Fälle der verschiedenen Art.

Dennoch gab es eine riesige Variation in der Genauigkeit der Urteile zwischen zwei experimentellen Bedingungen. In der einen Bedingung konnten die Teilnehmer die Beobachtungen nach dem *Prädiktor* sammeln, also beliebig viele Frauen mit M+ oder M– aus der Datenbank auswählen, um zu erfahren, ob bei diesen Frauen Brustkrebs vorlag oder nicht. In der anderen Bedingung hingegen war die Informationssuche konditional zum *Urteilskriterium*, das heißt, die Teilnehmer konnten beliebig viele Frauen mit und ohne Brustkrebs betrachten und bekamen eine Rückmeldung darüber, ob das Mammogramm bei diesen Frauen M+ oder M– ergeben hatte. In der ersteren Bedingung waren die Urteile der gesuchten Wahrscheinlichkeit p(B/M+) durchweg sehr genau. Auch wenn nur eine kleine Zahl von M+ Fällen gezogen worden waren, und unabhängig davon, wie viele irrevante M– gezogen worden waren, hatten die Teilnehmer wenig Probleme zu erkennen, dass in der Bezugsmenge nur wenige Brustkrebs-Fälle waren. In der letzteren Bedingung hingegen waren die meisten Urteile von p(B/M+) inflationär überhöht, weil die meisten Teilnehmer eine *kriteriums-selektive Stichprobe* erzeugt hatten. Mit anderen Worten, ein typischer Teilnehmer betrachtet alle Frauen mit Brustkrebs, die in der Datenbank zu finden waren, plus eine etwa ebenso große Vergleichsmenge von Frauen ohne Brustkrebs. Damit ist natürlich die gezogene Beobachtungsstichprobe drastisch nach dem Kriterium (i.e., nach dem Vorkommen von B) verzerrt; anstelle der sehr kleinen Basisrate von B in der gesamten Datenbank befinden sich unter den selektierten Fällen etwa eine Hälfte Fälle mit Brustkrebs. Unsere

experimentellen Ergebnisse zeigen sehr deutlich, dass Urteile den Anteil der B-Fälle in der selektiven Stichprobe recht genau widerspiegeln. Aber da die Häufigkeits-Stichprobe selbst hinsichtlich des Kriteriums drastisch verzerrt ist, fallen diese Urteile viel zu hoch aus, wenn die Beobachtungen nach dem Kriterium selektiert wurden. Wurden sie nach dem Prädiktor selektiert, dann bleibt die tatsächliche Kriteriums-Rate der B-Fälle (bis auf Zufallsfehler) in der Stichprobe erhalten, und die Urteile sind entsprechend genau. In beiden Bedingungen spiegeln die Urteile die Häufigkeiten in der Stichprobe wider; sie sind jedoch nicht sensitiv gegenüber der kritischen Frage, ob die Häufigkeitsstichprobe selbst selektiv verzerrt ist oder nicht. Obwohl die Teilnehmer in allen Bedingungen relativ gut in der Quantifizierung von Häufigkeiten sind, reicht das nicht aus. In der kriteriumsselektiven Bedingung müssten sie eigentlich ihr Urteil nach einer Gewichtungsregel ausrichten. Wenn sie z.B. alle B-Fälle (100%) aus der Population genommen haben, aber nur einen kleinen Teil der ¬B-Fälle (sagen wir 5%), dann müssten sie den Selektionsfaktor von 100% / 5% = 20 bei ihrem Urteil berücksichtigen. Das heißt, statt einfach den B-Anteil in der gezogenen Stichprobe zu schätzen, müssten sie diesen Anteil um den reziproken Faktor 1/20 korrigieren.

Eine Vielzahl von Befunden zeigt (Fiedler, 2000; Fiedler et al., 2000; Fiedler, Freytag, Unkelbach, Schreiber, Wilke, Bayer & Wild, 2003; Hamill, Wilson & Nisbett, 1980), dass vielleicht das schwerwiegendste Problem im stochastischen Denken in dieser Unsensibilität für die Zusammensetzung von Beobachtungsstichproben liegt. Ungeachtet von Bildung und Expertise verlassen sich Urteiler auf die Häufigkeitsverhältnisse in der vorhandenen Stichprobe, die sie recht genau abbilden, aber dabei unkritisch bleiben hinsichtlich möglicher Verzerrungen der Stichprobe. Immer dann, wenn die Stichprobe kriteriums-selektiv verzerrt ist, erweist sich der wirksame Umgang mit Häufigkeiten als wenig hilfreich. Ironischerweise unterstützt das genaue Verarbeiten von Häufigkeitsinformation geradezu die teilweise groben Fehleinschätzungen in solchen Fällen.

Die Schwierigkeit der korrekten Verarbeitung von kriteriums-selektiven Stichproben und das Versäumnis angemessener Korrekturen sind wie gesagt durch viele Befunde belegt, und viele bekannte Befunde können damit (alternativ) erklärt werden, insbesondere auch der oben erwähnte Befund von Gigerenzer und Hoffrage (1995). Betrachtet man noch einmal die Information über Brustkrebs und Mammografie im Wahrscheinlichkeitsformat (siehe oben), dann fällt auf, dass in diese Formulierung der Aufgabe nicht nur Wahrscheinlichkeiten (in %) eingehen, sondern ein hoher Wert von p(M+/B) = 80% im Vergleich zu dem niedrigen Wert von p(M+/¬B) = 9.6%, was dadurch zustande kommt, dass sich die beiden Zahlen auf sehr ungleiche Grundmengen beziehen (i.e., sehr wenige B-Fälle und sehr viele ¬B-Fälle). Die ungleichen Zahlenwerte beziehen sich also implizit auf eine kriteriums-selektive Stichprobe, welche von den Teilnehmern eine unglei-

che Gewichtung bzw. Korrektur der gegebenen Zahlen (entsprechend der gegebenen Selektivität) verlangen würde. Man hätte ja alle Wahrscheinlichkeiten auf dieselbe Grundmenge angeben können. Anstelle der numerisch hohen, irreführenden Angabe, dass p(M+/B) = 80% hätte man die Wahrscheinlichkeiten auf eine nicht-selektive, zufällige Stichprobe aus der gesamten Population beziehen können, etwa p(M+ $\wedge$ B / Zufallsstichprobe) = .008 und p(M+ $\wedge$ $\neg$B Zufallsstichprobe) = .096. Dann hätte man auch im Wahrscheinlichkeitsformat sehen können, dass es weitaus mehr Fälle mit positivem Mammogramm ohne als mit Brustkrebs gibt. Es besteht also die Möglichkeit, dass ein Teil der Befunde, die für einen Vorteil von Häufigkeitsformat sprechen, eher auf das Problem der kriterium-selektiven Stichproben zurückzuführen sind als auf den eigentlichen Unterschied von Häufigkeiten und Wahrscheinlichkeiten (d.h., normierten Häufigkeiten).

In jedem Falle aber stellt sich die Frage, wie man menschliche Urteiler und Entscheider für das Problem kriterium-selektiver Stichproben sensibilisieren kann. Ein Grossteil grob verzerrter Wahrscheinlichkeitsurteile beruht darauf, dass die Verzerrung in der gegebenen Stichprobe von Beobachtungen oder Statistiken nicht erkannt und nicht angemessen korrigiert wird (Fiedler, 2000). Liegt dies daran, dass der menschliche Verstand durch die Überwachung und die Kontrolle verzerrter Stichproben überfordert wird, weil eine logisch oder mathematisch angemessene Korrektur ohnehin viel zu kompliziert wäre? Denkwürdige Befunde zum Teil mit statistischen Experten deuten auf eine fast notorische Blindheit oder Kurzsichtigkeit für verzerrte Stichproben hin (Fiedler et al., 2003). Oder besteht die Möglichkeit, in Unterricht oder spezifischen Trainings menschliche Urteiler für das Problem zu sensibilisieren und zumindest Teilerfolge bei der Beachtung und Korrektur des Problems der kriteriumsverzerrten Stichproben zu erzielen? - Das nachfolgende Experiment, welches bewusst an einem alltäglichen Problem der Bewertung von Fußballleistung orientiert ist, stellt einen Versuch dar, dieser bedeutsamen Frage nachzugehen.

2. *Methode*

2.1 Überblick

Ähnlich wie in der Untersuchung von Plessner et al. (2001) bestand die Aufgabe der Versuchsteilnehmer darin, Beobachtungen über die Leistung einer Fußballmannschaft zu sammeln, wenn ein bestimmter Spieler entweder eingesetzt wird oder nicht, um daraufhin die Wahrscheinlichkeit p(Mannschaft gut/Spieler dabei) zu beurteilen, also die Wahrscheinlichkeit einer guten Mannschaftsleistung, wenn der betreffende Einzelspieler zum Einsatz kommt. Dazu standen ihnen Informationen über die bisherigen Spiele der Mannschaft aus den letzten drei Jahren zur Verfügung. Diese Informationen gaben jeweils an, ob die Mannschaft in einem jeweiligen Spiel gut oder schlecht gespielt hat und ob der Spieler mitgespielt hat oder nicht. Die Informationen lagen in Form von 100 Karteikarten vor, von denen die

Versuchsteilnehmer so viele anschauen konnten, wie sie wollten. Der einen Hälfte der Teilnehmer stand ein Kasten zur Verfügung, der nach dem Prädiktor sortiert war (d.h. danach, ob der Spieler dabei war oder nicht), die andere Hälfte benutzte einen Kasten, der nach dem Kriterium sortiert war (d.h. danach, ob die Mannschaft gut war oder schlecht). Nachdem die Versuchspersonen sich ihrer Ansicht nach ausreichend informiert hatten, wurden sie gebeten, die bedingte Wahrscheinlichkeit p(Mannschaft gut/Spieler dabei) und die allgemeine Leistungsstärke des Spielers zu schätzen. Es wurde erwartet, dass diese Urteile in der Kriteriums-Bedingung, nicht aber in der Prädiktor-Bedingung, wie immer stark verzerrt sein würden, infolge einer kriterium-selektiven Stichprobe. Dieser Befund würde bisherige Erkenntnisse replizieren, allerdings mit einem vertrauten Alltagsproblem anstelle technisch-diagnostischer Probleme. Darüber hinaus sollte jedoch untersucht werden, ob es möglich ist, die Teilnehmer für das Problem der kriteriums-selektiven Stichprobe zu sensibilisieren und dadurch genauere Urteile zu erzielen. Zu diesem Zwecke wurden die Teilnehmer anschließend eine zweites Mal um ihre Urteile gefragt, nachdem sie aufgefordert worden waren, über die unterschiedliche Anzahl der im Karteikasten sichtbaren Karten (entweder bezüglich der beiden Ausprägungen des Prädiktors oder der des Kriteriums) nachzudenken.

2.2 Versuchspersonen

Es nahmen insgesamt 37 Trainer und 3 Trainerinnen aus verschiedenen Mannschaftssportarten als Versuchspersonen an der Untersuchung teil. Sie wurden telefonisch über Vereine aus der Umgebung von Heidelberg für ein Experiment zur Leistungsbewertung im Sport angeworben. Im Durchschnitt waren die Versuchspersonen 31.6 (SD = 11.1) Jahre alt und hatten Erfahrungen als Trainer/innen seit 6.3 (SD = 7.1) Jahren.

2.3 Material und Design

Das Material entsprach vollständig dem in der Untersuchung von Plessner et al. (2001) verwendeten Material. Es bestand aus zwei Karteikästen mit Informationen über einen Fußballspieler. Sie enthielten jeweils 100 Karten. Auf jeder dieser Karten befanden sich Angaben über ein Spiel der bisherigen Mannschaft des Spielers aus den letzten drei Jahren. Diese Angaben bezogen sich zum einen darauf, ob der Spieler mitgespielt hat oder nicht, und zum anderen darauf, ob die Mannschaft gut gespielt hat oder schlecht. Insgesamt gab es 7 Spiele, bei denen der Spieler dabei war und die Mannschaft gut gespielt hat, 19 bei denen er dabei war, die Mannschaft aber schlecht spielte, 2 Spiele bei denen er fehlte und die Mannschaft gut war sowie 72 Spiele, bei denen er fehlte und die Mannschaft schlecht spielte. Die bedingte Wahrscheinlichkeit, dass die Mannschaft gut spielt wenn der Spieler dabei ist, betrug dementsprechend 26.9%; die Kontingenz zwischen dem Mitwirken des Spielers und der Mannschaftsleistung $\Delta = 7/(7+19) - 2/(2+72) = .24$.

Die beiden Karteikästen unterschieden sich nur in der Sortierung der Karteikarten, ansonsten enthielten sie exakt dieselben Informationen. Der „Prädiktor"-Kasten war so sortiert, dass bereits ohne Ziehen der Karteikarten anhand ihrer Farbe sofort erkennbar war, ob der Spieler mitgespielt hat oder nicht und wie häufig das jeweils der Fall war. Um zu erfahren, ob die Mannschaft dann jeweils gut oder schlecht gespielt hat, musste erst eine Karte gezogen werden. Im „Kriterium"-Kasten war es genau umgekehrt. Im Unterschied zu der Untersuchung von Plessner et al. (2001) wurden die Versuchspersonen zufällig der Präditkor- und der Kriteriums-Gruppe zugeteilt.

2.4 Prozedur

Das Experiment wurde in Einzelsitzungen durchgeführt. Zunächst erhielten die Versuchspersonen eine schriftliche Instruktion. Darin wurden sie gebeten, sich in die Lage eines Trainers zu versetzen, der für seine Mannschaft dringend einen neuen Mitspieler benötigt und zu diesem Zweck einen Spielervermittler beauftragt hat, Informationen über einen bestimmten Spieler zu sammeln. Diese Informationen stünden ihnen in Form eines Karteikastens zur Verfügung. Sie sollten sich auf Grundlage der in dem Kasten enthaltenen Informationen einen Eindruck über die Leistungsstärke des Spielers machen. Es folgte eine kurze Instruktion, aus welcher hervorging, dass die Versuchspersonen so viele Karten wie sie wollten, aber mindestens zehn, aus ihrem Karteikasten ziehen und sich deren Inhalt einprägen sollten. Nach dem Ziehen und Lesen einer jeweiligen Karte wurde sie verdeckt beiseite gelegt. Am Ende der Untersuchung wurde von der Versuchsleiterin die genaue Zusammensetzung dieser gezogenen Stichprobe notiert. Wenn die Versuchsteilnehmer angaben, dass sie genügend Karten gelesen hatten, wurde ihnen ein Fragebogen mit verschiedenen Fragen zu den Informationen in den Karteikästen vorgelegt. Die abhängigen Variablen waren die Einschätzungen der bedingten Wahrscheinlichkeit p(Mannschaft gut/Spieler dabei) in Prozent und der Leistungsstärke des Spielers auf einer Ratingskala von 1(sehr schlecht) bis 7 (sehr gut). Nach der Beantwortung dieser Fragen erhielten die Teilnehmer einen neuen Fragebogen, der folgendermaßen überschrieben war: „Bitte denken Sie noch mal darüber nach, dass die blauen Karten (Mannschaft hat gut gespielt bzw. der Spieler war dabei) und die grünen Karten (Mannschaft hat schlecht gespielt bzw. der Spieler war nicht dabei) in unterschiedlicher Zahl in dem Karteikasten vorlagen. Im Hinblick auf diese Tatsache möchten wir Sie nun bitten, die gleichen Fragen noch einmal zu beantworten". Entsprechend wurden den Versuchspersonen die gleichen Fragen zur bedingten Wahrscheinlichkeit und zur Leistungsstärke des Spielers erneut vorgelegt. Nach der anschließenden Erhebung soziodemographischer Variablen wurden die Versuchspersonen über den genauen Zweck der Untersuchung informiert.

3. *Ergebnisse*

3.1 Stichprobenziehen

In der Kriterium-Gruppe wurden im Mittel etwa gleich viele Karten gezogen (M = 15.1, SD = 5.4) gezogen wie in der Prädiktor-Gruppe (M = 14.3, SD = 5.3), $t(38) = 0.48$, $p > .60$. Ein deutlicher Unterschied zeigte sich hingegen wie erwartet bei der Verteilung der gezogenen Karten auf die Bedingungen der Vierfeldertafel bzw. der daraus resultierenden bedingten Wahrscheinlichkeit, dass die Mannschaft gut spielt, wenn der Spieler dabei ist, und der Kontingenz zwischen dem Mitwirken des Spielers und der Mannschaftsleistung. Während in der Prädiktor-Gruppe die tatsächliche Wahrscheinlichkeit in der Gesamtstichprobe von 26.9% ziemlich genau abgebildet wurde (M = 26.3%, SD = 18.2), lag die Wahrscheinlichkeit in der Kriterium-Gruppe deutlich höher (M = 73.1%, SD = 19.2), $t(39) = 8.84$, $p < .001$. In gleicher Weise unterschieden sich die für die Leistungsbewertung wesentlichen Kontingenzen, das heißt, die in der gezogenen Stichprobe erhaltenen Differenzen Δ = p(Mannschaft gut/Spieler dabei) - p(Mannschaft gut/Spieler nicht dabei). Auch hier wurde die in der Gesamtstichprobe enthaltene Kontingenz von $\Delta = .24$ eher in der Prädiktor-Gruppe (M =.23, SD = .19) abgebildet als in der Kriterium-Gruppe (M = .51, SD = .35), $t(39) = 3.44$, $p < .01$.

3.2 Schätzung der bedingten Wahrscheinlichkeit

Die mittlere Einschätzung der bedingten Wahrscheinlichkeit p(Mannschaft gut/Spieler dabei) ist in Abhängigkeit von den vier Versuchsbedingungen in Tabelle 1 zu sehen. Per Augenschein ist dort bereits zu erkennen, dass diese Wahrscheinlichkeit in der Prädiktor-Gruppe relativ genau eingeschätzt wurde während sie in der Kriterium-Gruppe sehr stark überschätzt wurde, solange es keinen spezifischen Hinweis auf die Basisrate gab. Wurde jedoch ein Hinweis gegeben, veränderten die Versuchspersonen der Kriterium-Gruppe ihre Einschätzung in Richtung des korrekten Wertes von 26.9%, während die Versuchspersonen der Prädiktor-Gruppe bei ihrer bereits sehr genauen Einschätzung blieben.

Tab. 1: Einschätzung der bedingten Wahrscheinlichkeit P(Mannschaft gut/Spieler dabei) und der generellen Leistung in Abhängigkeit von Karteikastensortierung und Hinweis

	Prädiktor				Kriterium			
	ohne Hinweis		mit Hinweis		ohne Hinweis		mit Hinweis	
	M	SD	M	SD	M	SD	M	SD
WK	25.2	17.2	24.1	17.3	63.6	27.6	35.0	16.6
Leistung	2.24	1.09	2.34	1.15	3.95	1.57	3.10	0.96

Dieser Eindruck wird unterstützt durch die Ergebnisse einer Varianzanalyse (ANOVA). Die Daten wurden einer 2 x 2 (Kastensortierung x Hinweis) ANOVA unterzogen, mit dem Faktor Hinweis als Messwiederholung. Die-

se Analyse ergibt Haupteffekte für die Faktoren Kastensortierung, $F(1, 38) = 23.4$, $p < .01$, und Hinweis, $F(1, 38) = 16.9$, $p < .01$ sowie einen Interaktionseffekt zwischen beiden Faktoren ($F(1, 38) = 14.3$, $p < .01$). Paarweise t-Tests zeigen entsprechend einen großen Unterschied zwischen den Urteilen mit und ohne Hinweis, wenn den Versuchspersonen der nach dem Kriterium sortierte Kasten zur Verfügung stand, $t(19) = 4.22$, $p < .01$, jedoch keinen vergleichbaren Unterschied, wenn sie ihre Suche nach dem Prädiktor ausgerichtet hatten, $t(19) = 0.44$, $p > .60$.

3.3 Leistungsbewertung

Ergänzend zu der Einschätzung der bedingten Wahrscheinlichkeit waren die Versuchspersonen gebeten worden, ein Gesamturteil über die Leistungsstärke des Spielers (s. untere Zeile in Tabelle 1). Diese Urteile folgen im Wesentlichen den Einschätzungen der bedingten Wahrscheinlichkeiten. Die ANOVA ergibt einen Haupteffekt für die Karteikastensortierung, $F(1, 38) = 15.$, $p < .01$, und eine Interaktion zwischen der Karteikastensortierung und dem Faktor Hinweis, $F(1, 38) = 4.93$, $p < .05$. Auch hier ergaben paarweise t-Tests einen Unterschied zwischen Urteilen mit und ohne Hinweis in der Kriterium-Gruppe, $t(19) = 2.36$, $p < .05$, und keinen vergleichbaren Effekt in der Prädiktor-Gruppe, $t(19) = 0.69$, $p > .40$. Damit zeigt sich auch bei den Leistungsbewertungen, dass die Teilnehmer dieser Untersuchung unter günstigen Umständen dazu in der Lage waren, ihren durch kriteriumsselektive Stichproben verzerrten Eindruck bei entsprechendem Hinweis auf die Basisrate zu korrigieren.

4. Diskussion

Die hier berichtete Untersuchung mit Trainern aus Mannschaftssportarten diente zunächst dazu, das Problem kriteriums-selektiver Stichproben als Ursache grob verzerrter Wahrscheinlichkeitsurteile noch einmal zu illustrieren. Das verwendete Urteilsproblem war nicht aus einer künstlichen, technischen Umwelt genommen (wie HIV-Diagnostik), sondern aus der Alltagswelt. Man darf annehmen, dass die Experimentalsituation durchaus repräsentativ ist für die Art und Weise, wie die Leistung von Sportmannschaften bzw. der Einfluss einzelner Spieler unter realen Bedingungen bewertet wird. Das heißt, die Methode erscheint inhaltsvalide, und man sollte bedenken, dass die Teilnehmer wirkliche Trainer waren, also echte Experten auf dem besagten Gebiet. An Vertrautheit mit dem Inhalt der Aufgabe und an Realitätsnähe kann es also keinesfalls gemangelt haben. Die Trainer wurden im Übrigen nicht mit unnatürlichen Wahrscheinlichkeits-Statistiken gefüttert, sondern konnten auf selbstbestimmte Weise eine Reihe einzelner Beobachtungen sammeln, also genuine Häufigkeitsinformation verwenden. Dennoch legten sie die typischen massiven Fehleinschätzungen der fraglichen Wahrscheinlichkeit - p(Mannschaft gut/Spieler dabei) - an den Tag, aber nur in derjenigen Bedingung, die zu kriteriums-selektiven Stichproben führte. Ähnlich wie kritische seltene Kriteriums-Ereignisse in der Diagnos-

tik (das Vorliegen von AIDS, Brustkrebs etc.) wurde das kritische seltene Ereignis bei diesem Sportproblem (d.h., das insgesamt selten auftretende Ereignis einer guten Mannschaftsleistung) in der selbst gewählten Stichprobe überrepräsentiert, *wenn* die Stichprobenziehung am Kriterium orientiert war (d.h., wenn die Trainer Spiele auswählen konnten, bei denen die Mannschaft gut oder schlecht gespielt hatte). Dieses Herausgreifen der wenigen Spiele mit positivem Ausgang ist nicht sehr verwunderlich, wenn man bedenkt, dass schließlich die Wahrscheinlichkeit dieses Ausgangs (in Abhängigkeit von einem bestimmten Spieler) beurteilt werden sollte. Das Denkproblem besteht indessen darin, dass man den inflationär hohen Anteil positiver Ausgänge in einer so gezogenen Stichprobe nicht direkt als Schätzung für p(Mannschaft gut/Spieler dabei) verwenden darf, sondern erst eine Selektivitätskorrektur durchführen müsste. Genau diese Einsicht liegt jedoch bei den Urteilern nicht spontan vor. In der anderen Bedingung, in der die Beobachtungsstichprobe nach dem Prädiktor gesammelt wird (also Spiele ausgewählt werden, in denen der Spieler anwesend war oder nicht), ist eine Korrektur nicht nötig, und folglich fallen die Schätzungen durchweg sehr genau aus. Die Trainer erkennen, dass auch dann, wenn der kritische Spieler dabei ist, die Rate positiver Gesamtleistungen bescheiden bleibt. So finden wir in derselben Untersuchung beides, einen weiteren Beleg für genaue frequentistische Schätzungen - relativ zu den Häufigkeiten in der Stichprobe - sowie einen Beleg für extrem fehlgeleitete Urteile - trotz Vertrautheit und Häufigkeitsformat - wenn die Stichprobe am Kriterium selektiert und damit von vornherein verzerrt ist. Das eigentliche Denkproblem bei derartigen stochastischen Aufgaben liegt somit darin, den Prozess der Datengewinnung zu durchdringen und zu erkennen, dass nicht jede Stichprobe gleich geeignet ist, um ein genaues Urteil abgeben zu können. Stichproben können - wie nun vielfach illustriert - besonders dann zu drastischen Fehleinschätzungen führen, wenn sie in selektiver Weise das kritische Kriteriumsereignis, dessen Auftretensrate geschätzt werden soll, repräsentieren oder überrepräsentieren.

Ein Blick auf reale Urteilsprobleme zeigt indessen, dass es gute Gründe gibt, warum kriteriums-selektive Stichproben nicht die Ausnahme, sondern meist die Regel sind. Ebenso wie ein Virologe oder Radiologe, der p(Brustkrebs/M+) aufgrund seiner langjährigen Berufserfahrung schätzen soll, aus dem Gedächtnis vermutlich vor allem Fälle mit Brustkrebs abruft, setzen sich im Gedächtnis von Fußballtrainern besonders Spiele mit einem besonderen Ausgang fest, berichten Zeitungen vor allem über seltene Katastrophen, notieren sich Lehrer in ihrem Notizbuch außergewöhnliche Leistungen oder Fehlleistungen. Aus pragmatischen Gründen ist es oft nur schwer möglich, Stichproben nach dem Prädiktor zu sammeln. Um die Wahrscheinlichkeit p(Unfall/Alkohol im Blut) zu schätzen, ist man in der Regel auf Unfall-Statistiken angewiesen, das heißt, auf Stichproben, die an dem Kriterium (Unfall) selektiert sind. Hingegen ist es rechtlich und prag-

matisch kaum möglich, Autofahrer, die Alkohol getrunken haben oder nicht, so lange zu beobachten, bis Unfälle auftreten. Die vorhandenen kriteriums-selektiven Stichproben sind daher denkbar ungeeignet, brauchbare Schätzungen der Zielgröße abzugeben, es sei denn, man kennt den Selektionsfaktor und kann die Schätzungen dementsprechend korrigieren.

Eben weil kriteriums-selektive Stichproben im Alltag vermutlich weit verbreitet sind und die ständige Gefahr von groben Fehlurteilen in sich bergen, ist es unseres Erachtens wichtig, ein Verständnis für dieses Problem im Unterricht zu schärfen und die Möglichkeit einer gezielten Sensibilisierung auszuloten. Die vorläufige Untersuchung stellt einen Versuch dar, Urteilsfehler zu reduzieren, indem man Urteiler für die selektive Zusammensetzung ihrer Erfahrungsstichprobe sensibilisiert. Die Resultate geben zu vorsichtigem Optimismus Anlass, denn nach der Aufforderung zur Reflektion über die selektive Stichprobe konnten die meisten Urteiler der Kriteriums-Bedingung ihre fehlerhaften Urteile erfolgreich korrigieren. Wir müssen jedoch zugeben, dass solche Versuche einer Sensibilisierung in anderen Untersuchungen weniger erfolgreich waren (Fiedler et al., 2003). Die entscheidende Frage ist schließlich nicht, ob man ein einzelnes Urteil korrigieren kann, sondern ob eine derartige Sensibilisierung für das Problem kriteriums-selektiver Stichproben auch Transfereffekte auf neue Aufgaben zeigt, bei denen dieselben Denkfehler von vornherein vermieden werden. Dies zu untersuchen könnte ein prominentes Ziel für die Zusammenarbeit von Kognitiven Psychologen und Unterrichtswissenschaftlern sein.

Literatur

Bernoulli, J. (1713). Ars conjectandi. Basilea: Thurnisius.

Dougherty, M.R.P., Gettys, C.F. & Ogden, E.E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. Psychological Review, 106, 180-209.

Fiedler, K. (1996). Explaining and simulating judgment biases as an aggregation phenomenon in probabilistic, multiple-cue environments. Psychological Review, 103, 193-214.

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

Fiedler, K., Brinkmann, B., Betsch, T. & Wild, B. (2000). A sampling approach to biases in conditional probability judgments: Beyond baserate neglect and statistical format. Journal of Experimental Psychology: General, 129, 399-418.

Fiedler, K., Freytag, P., Unkelbach, C., Schreiber, V., Bayer, M. & Wild, B. (2003). Subjective Validity Judgments of Fictitious Research Findings: A Paradigm for Investigating Sampling Biases in Social Judgments. Submitted for publication.

Gigerenzer, G. & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. Psychological Review, 102, 684-704.

Gigerenzer, G. & Hoffrage, U. & Ebert, A. (1998). AIDS counseling for low-risk clients. Aids Care, 10, 197-211.
Hamill, R., Wilson, T. & Nisbett, R.E. (1980). Insensitivity to sample bias: Generalizing from atypical cases. Journal of Personality and Social Psychology, 39, 578-589.
Hasher, L. & Zacks, R.T. (1984). The automatic processing of fundamental information: The case of frequency of occurrence. American Psychologist, 39, 1372-1388.
Hoffrage, U. & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. Academic Medicine, 73, 538-540.
Hoffrage, U., Lindsey, S., Hertwig, R. & Gigerenzer, G. (2000). Communicating statistical information. Science, 290, 2261-2262.
Huber J. & Puto, C. (1983). Market boundaries and product choice: Illustrating attraction and substitution effects. Journal of Consumer Research, 10, 31-44.
Lindsay, S., Hertwig, R. & Gigerenzer, G. (in press). Communicating statistical evidence. Jurimetrics.
Inhelder, B. & Piaget, J. (1958). The growth of logical thinking from childhood to adolescence. New York: Basic Books.
Plessner, H., Hartmann, C., Hohmann, N. & Zimmermann, I. (2001). Achtung Stichprobe! Der Einfluss der Informationsgewinnung auf die Bewertung sportlicher Leistungen. Psychologie & Sport, 8, 91-100.
Rulon, P.J. (1941). Problems of regression. Harvard Educational Review, 11, 213-223. *
Sedlmeier, P. (1999). Improving statistical reasoning: Theoretical models and practical implications. Mahwah, NJ: Lawrence Erlbaum Associates.
Sedlmeier, P., Hertwig, R. & Gigerenzer, G. (1998). Are judgments of the positional frequencies of letters systematically biased due to availability? Journal of Experimental Psychology: Learning, Memory and Cognition, 24, 754-770.
Swets, J., Dawes, R.M. & Monahan, J. (2000). Psychological science can improve diagnostic decisions. Psychological Science in the Public Interest, 1, Whole No. 1.

Anschrift der Autoren

Prof. Dr. Klaus Fiedler, Psychologisches Institut der Universität Heidelberg
Hauptstr. 47-51, 69117 Heidelberg, Tel.: 06221 547270, Fax: 06221 547745,
E-Mail: klaus.fiedler@psychologie.uni-heidelberg.de

Dr. Henning Plessner, Psychologisches Institut der Universität Heidelberg,
Hauptstr. 47-51, 69117 Heidelberg, Tel.: 06221 547700, Fax: 06221 547745,
E-Mail: henning.plessner@psychologie.uni-heidelberg.de