

EINZELNE KOGNITIVE TÄUSCHUNGEN UND IHR EINFLUSS AUF DAS URTEIL VON RICHTERN UND PARTEIEN IN GERICHTSVERFAHREN

§ 7 Methode, Auswertung, interne und externe Validität der Studie

I. Methode

- 157 Es wurden zwei briefliche Umfragen durchgeführt; die erste im Oktober 2003, die zweite im September 2004. Die Fragebogen sind in den Anhängen B (2003) und C (2004) wiedergegeben. Beide Fragebogen umfassten fünf Sachverhalte, jeder Sachverhalt begann auf einer neuen Seite. Auf der Titelseite wurden die Richterinnen und Richter gebeten, die Fragen ohne Absprache mit ihren Kolleginnen und Kollegen zu beantworten und darauf hingewiesen, dass die Beantwortung etwa eine Viertelstunde in Anspruch nehmen würde (was leicht untertrieben war). Der persönlich adressierte Begleitbrief mit dem Briefkopf des Europäischen Instituts für Rechtspsychologie erläuterte, dass es sich um ein psychologisches Forschungsprojekt handle, ohne genauere Angaben zu machen. Die Umfrage erfolgte anonym; demografische Angaben wurden nicht erhoben. Jedem Fragebogen lag ein adressiertes und frankiertes Rückantwortcouvert bei, und die Richter wurden aufgefordert, den Fragebogen binnen Wochenfrist zu retournieren. Das empfanden einzelne als Zumutung; Erfahrungen aus dem Direkt-Marketing zeigen aber, dass eine kurze Antwortfrist zu einer höheren Rücklaufquote führt.
- 158 Beide Fragebogen wurden einer Voruntersuchung („Pretest“) unterzogen. Als Versuchspersonen des Pretests dienten meine damaligen Arbeitskollegen bei Meyer Lustenberger Rechtsanwälte, Zürich (N ≈ 30). Die Versuchspersonen der Pretests waren alle Juristinnen und Juristen mit zwischen einem und dreissig Jahren Berufserfahrung. Sie sind mit der Testpopulation nicht direkt vergleichbar, da in der Schweiz das Laienrichtertum nach wie vor weit verbreitet ist und die Versuchspersonen des Pretests sich schwerewichtig mit Wirtschaftsrecht befassen, das an erstinstanzlichen Gerichten (wo der grösste Teil der befragten Richter beschäftigt ist) eine geringe Rolle spielt. Trotzdem zeigte der Pretest, welche Sachverhalte angepasst werden mussten und führte zu einigen Änderungen bei der Formulierung der Sachverhalte (wo diese wesentlich sind, werden sie bei den einzelnen Fragen näher erläutert).
- 159 Für die Umfrage vom Herbst 2003 wurden 401 Fragebögen an die Richterinnen und Richter der Zivil- und Strafgerichte der Kantone Aarau, St. Gallen und Zürich verschickt (inklusive der jeweiligen Ober- resp. Kantonsgerichte). Die Fragebogen folgten dem Aufbau der von GUTHRIE, RACHLINSKI und WISTRICH verwendeten Fragebogen, deren Studie die Grundlage für die erste Umfrage bildete.³⁴¹ Da die Fragebogen nicht nur vom Amerikanischen ins Deutsche übersetzt, sondern auch dem schweizerischen Rechtssystem angepasst werden mussten, war eine exakte Replikation ausgeschlossen. Trotzdem ist es interessant, die Resultate der amerikanischen Studie mit denjenigen der vorliegenden

³⁴¹ GUTHRIE/RACHLINSKI/WISTRICH, FN 20, 777. JEFFREY RACHLINSKI stellte mir freundlicherweise die Original-Fragebogen seiner Umfrage zur Verfügung.

Studie zu vergleichen. Jeder Sachverhalt testete eine kognitive Täuschung – Ankereffekt, Darstellungseffekt, Vernachlässigung der Anfangswahrscheinlichkeit, Rückschaufehler und Selbstüberschätzung. Von den Sachverhaltsschilderungen für Ankereffekt, Darstellungseffekt und Rückschaufehler gab es jeweils verschiedene Versionen, um den Einfluss der manipulierten Variable testen zu können. Die Richter wurden auf diesen Umstand nicht hingewiesen. Ich habe keinen Anlass anzunehmen, dass die Richter die Manipulation bemerkt haben. Keine der zahlreichen schriftlichen Reaktionen wies auf diesen Umstand hin, und auch meine Kontakte zu Bezirksrichtern zürcherischer Bezirksgerichte ergaben, dass niemand eine entsprechende Vermutung hegte.³⁴² Die Sachverhaltsvarianten wurden gemischt (wobei die Reihenfolge der Fragen immer gleich blieb). Da es drei Sachverhaltsvarianten für den Ankereffekt, drei Sachverhaltsvarianten für den Rückschaufehler und zwei Varianten für den Darstellungseffekt gab, wurden insgesamt 18 verschiedene Varianten des Fragebogens versandt.

- 160 181, oder 45 %, der angeschriebenen Richter schickten den Fragebogen zumindest teilweise ausgefüllt zurück. Diese – für eine briefliche Umfrage erfreulich hohe – Rücklaufquote entspricht derjenigen, die andere Forscher mit ähnlichen brieflichen Umfragen unter schweizerischen Richtern erzielen konnten.³⁴³ Selbstverständlich bietet das „Non-response-Problem“ Anlass zu methodischer Kritik.³⁴⁴ Die interne Validität (zum Begriff S. 69 f.) wird aber nur dann in Frage gestellt, wenn zwischen Antwortenden und Verweigernden ein systematischer Unterschied besteht. Im vorliegenden Fall muss ein Kritiker postulieren, dass die antwortenden Richter den kognitiven Täuschungen mehr unterliegen als die Verweigerer. Wenn ich auch nicht beweisen kann, dass diese Annahme nicht zutrifft, glaube ich doch, dass sie nicht besonders plausibel ist. Aufgrund einiger Rückmeldungen glaube ich übrigens, dass die antwortenden Richter die Umfrage ernst genommen haben. Einzelne Fragebogen wurden unausgefüllt zurückgeschickt mit der Bemerkung, man entscheide nur über familienrechtliche Streitigkeiten und sehe sich daher nicht im Stande, die Fragen zu beantworten. Die ausgefüllten Fragebogen stammen daher vermutlich von Richtern, die sich tatsächlich mit ähnlichen Fällen in der Praxis befassen.
- 161 Die zweite Umfrage im Herbst 2004 erfolgte nach derselben Methode, allerdings gab es hier kein Vorbild für den Fragebogen (es gibt natürlich Vorbilder für die Fragestellungen; es handelt sich dabei durchwegs um Fragen, die in der psychologischen Literatur schon wiederholt gestellt und die auf juristische Sachverhalte angepasst wurden). Es wurden bis auf eine Ausnahme (Rückschaufehler) andere kognitive Täuschungen als bei der ersten Umfrage geprüft (der Rückschaufehler wurde mit einem neuen Sachverhalt untersucht). Wiederum gab es von jedem Sachverhalt mindestens zwei bis maximal vier Varianten (zweifaktorieller Versuchsplan bei einzelnen Fragen). Es wurden diesmal 476 Fragebogen an die Richterinnen und Richter der Straf- und Zivilgerichte der Kantone Basel-Land-

³⁴² Der Begleitbrief war von Prof. Dr. Manfred Reh binder unterschrieben, so dass niemand ahnen konnte, wer der eigentliche Urheber der Umfrage war.

³⁴³ ANDRÉ KUHN/PATRICE VILLET AZ/ALINE JOYET/FLORIAN WILLI, *Öffentliche Meinung und Strenge der Richter*, *Crimiscope* Nr. 19, Lausanne 2000, 2, berichten von einer Rücklaufquote von 44 %. Sie legten 290 Richterinnen und Richtern vier Strafrechtsfälle zur Beurteilung vor.

³⁴⁴ RAINER SCHNELL/PAUL B. HILL/ELKE ESSER, *Methoden der empirischen Sozialforschung*, 7. Aufl. München 2005, 306 ff.

Besonderer Teil

schaft, Basel-Stadt, Bern und Graubünden (mit Ausnahme des italienischsprachigen Bezirksgerichts Moesa) verschickt. Titelblatt und Begleitbrief waren weitgehend identisch mit der ersten Umfrage. Anders als beim ersten Mal wurde diesmal nach einer Woche ein persönlich adressiertes Erinnerungsschreiben versandt. Da die Umfrage anonym erfolgte und ich daher nicht wissen konnte, welche Richter den Fragebogen retourniert hatten, haben alle Richter die Erinnerung erhalten; dies ist einzelnen sauer aufgestossen, die den Fragebogen bereits retourniert hatten, liess sich aber nicht vermeiden. 234 Fragebogen wurden zumindest teilweise ausgefüllt zurückgeschickt, was einer Rücklaufquote von 49 % entspricht; die Erinnerung scheint daher zumindest eine gewisse Wirkung gehabt zu haben.

II. Auswertung, oder: was bedeutet „statistisch signifikant“?

- 162 Im Folgenden wird ein beobachteter Unterschied gemäss üblicher Konvention in den Sozialwissenschaften dann als signifikant bezeichnet, wenn der p-Wert kleiner als 0,05 ist.³⁴⁵ Ein p-Wert von weniger als 0,05 wurde auch schon – in durchaus kritischer Absicht – als der „heilige Gral des Hypothesentestens“ bezeichnet.³⁴⁶ Was bedeutet der p-Wert?
- 163 Als erstes muss man klar unterscheiden zwischen „statistisch signifikant“ und dem umgangssprachlichen „wesentlich“ oder „erheblich“. Das Ergebnis eines Versuchs kann statistisch signifikant, aber unwesentlich sein.³⁴⁷ Wenn die Anzahl der untersuchten Fälle sehr gross ist, sind bereits kleine Differenzen statistisch signifikant, aber möglicherweise praktisch bedeutungslos. In dieser Arbeit wird der Ausdruck „signifikant“ ausschliesslich für Ergebnisse verwendet, die im Sinne der herkömmlichen Definition statistisch signifikant sind; hingegen wird ein Effekt als „wesentlich“ oder „erheblich“ beschrieben, wenn er – nach der subjektiven Meinung des Autors – praktisch bedeutsam ist.
- 164 Angenommen, es gibt a priori zwei Hypothesen, um die beobachteten Daten zu erklären. In der Untersuchung zum Darstellungseffekt (S. 95 ff.) beurteilten die Richter beispielsweise den gleichen Sachverhalt, aber eine Gruppe der Richter beurteilte ihn aus Sicht der Klägerin, eine zweite Gruppe aus Sicht der Beklagten. In der Gruppe der Richter, die den Fall aus der Sicht der Klägerin beurteilten, empfahlen 56,6 % (43 von 76), einen Vergleich abzuschliessen. Nur 43 % der Richter (43 von 100), die den Fall aus der Perspektive der Beklagten beurteilten, rieten jedoch der Beklagten zum Vergleich. Die Arbeitshypothese ist, dass die Perspektive einen Einfluss auf die Vergleichsbereitschaft (genauer gesagt die Risikoaversion) hat. Die Nullhypothese ist, dass es keinen Unterschied in der Vergleichsbereitschaft der beiden Gruppen gibt und dass der beobachtete Unterschied auf Zufall zurückzuführen ist.
- 165 Man kann, soviel ist sicher, zwei Fehler machen: die Nullhypothese verwerfen, obwohl sie wahr ist (Fehler 1. Art, auch α -Fehler), oder die Nullhypothese beibehalten, obwohl die

³⁴⁵ SCHNELL/HILL/ESSER, FN 344, 450.

³⁴⁶ PAULA BRAITSTEIN, Give P's a Chance: Understanding p-values in scientific research, Living+ Jan/Feb 2004, 34 (erhältlich unter www.bcpwa.org/articles/issue_28_34_give_ps_a_chance.pdf; besucht am 1. März 2005).

³⁴⁷ SCHNELL/HILL/ESSER, FN 344, 452.

Arbeitshypothese zutrifft (Fehler 2. Art, auch β -Fehler).³⁴⁸ Im Strafverfahren wäre ein Fehler 1. Art, einen Unschuldigen zu verurteilen, während ein Fehler 2. Art darin bestünde, einen Schuldigen freizusprechen.³⁴⁹ Mit diesem Beispiel wird auch klar, dass die Konsequenzen der beiden Fehler durchaus nicht dieselben sein müssen – die meisten Menschen sind sich einig, dass es schwerer wiegt, einen Unschuldigen einzusperren, als einen Verbrecher laufen zu lassen.

- 166 Wie wahrscheinlich ist es nun, dass die im Experiment zum Darstellungseffekt beobachteten Daten auf Zufall zurückzuführen sind? Zum Verständnis kann man folgendes Gedankenexperiment machen:³⁵⁰ angenommen, es gäbe keinen Unterschied der Perspektive, sondern es gäbe einfach Richter, die von ihrer Persönlichkeit her eher zum Vergleich raten und Richter, die eher den Prozess empfehlen. Mit anderen Worten hätten 86 der 176 Richter, die die Frage beantwortet haben, ohnehin einen Vergleich empfohlen, ganz unabhängig davon, ob sie den Fall aus der Perspektive der Klägerin oder der Beklagten beurteilten. Die Wahrscheinlichkeit, dass *alle* 86 Richter, die zum Vergleich rieten, sich *zufällig* in derselben Gruppe befinden, ist offensichtlich sehr klein: Es ist nicht relevant, welcher Gruppe der erste vergleichsfreudige Richter angehört. Der zweite vergleichsfreudige Richter muss aber in der gleichen Gruppe wie der erste Richter sein, die Wahrscheinlichkeit dafür ist 0,5. Der dritte Richter muss wiederum in der gleichen Gruppe sein wie die ersten beiden Richter, die Wahrscheinlichkeit dafür beträgt ebenfalls 0,5. Da die beiden Ereignisse unabhängig sind, ergibt sich die Wahrscheinlichkeit, dass der zweite *und* dritte Richter in der gleichen Gruppe wie der erste Richter sind, aus der Produktregel: $0,5 \cdot 0,5 = 0,25$. Bei 86 vergleichsfreudigen Richtern, die alle in der gleichen Gruppe sein müssen, beträgt die Wahrscheinlichkeit daher $0,5^{86}$.³⁵¹
- 167 Die Wahrscheinlichkeit, dass man alle 86 vergleichsbereiten Richtern in einer Gruppe findet, wenn die Nullhypothese zutrifft und die Verteilung rein zufällig ist, beträgt daher $0,5^{86}$ (eine Null, gefolgt von einem Komma und 25 Nullen, ehe eine 2 folgt). Diese Wahrscheinlichkeit entspricht dem p-Wert, dieser ist nämlich definiert als die Wahrscheinlichkeit, dass sich die beobachteten Daten so wie beobachtet oder extremer realisieren, wenn die Nullhypothese zutrifft.³⁵² Anders formuliert gibt der p-Wert die Wahrscheinlichkeit an, einen Fehler 1. Art zu begehen.
- 168 Offensichtlich wird die Wahrscheinlichkeit, dass sich die Daten wie beobachtet oder extremer realisieren, grösser, je weniger extrem die beobachteten Daten sind. Im Experiment zum Darstellungseffekt haben 43 Richter in der Gruppe „Klägerin“ und 43 in der Gruppe „Beklagte“ zum Vergleich geraten, was auf den ersten Blick nicht gerade auf einen statistisch signifikanten Einfluss der Gruppenzugehörigkeit hindeutet. Allerdings war die

³⁴⁸ LUDWIG FAHRMEIER/RITA KÜNSTLER/IRIS PIGEOT/GERHARD KUNZ, Statistik, 5. Aufl. Berlin etc. 2004, 416. Bekannt ist auch der Fehler 3. Art – die beiden Fehlerarten zu verwechseln.

³⁴⁹ Unter der Annahme, dass „unschuldig“ die Nullhypothese ist, die im ersten Fall (zu Unrecht) verworfen wird.

³⁵⁰ Die Darstellung folgt HANS ZEISEL/DAVID KAYE, Prove it with Figures: Empirical Methods in Law and Litigation, New York 1997, 83 f.

³⁵¹ Da eine Gruppe nur 76 Mitglieder hatte, beträgt die Wahrscheinlichkeit, dass alle Mitglieder dieser Gruppe vergleichsfreudige Richter sind, „nur“ $0,5^{76}$. Die kleinere Gruppe schwächt den Signifikanztest.

³⁵² FAHRMEIER/KÜNSTLER/PIGEOT/KUNZ, FN 348, 420.

Besonderer Teil

Gruppe der Richter, die den Fall aus Sicht der Beklagten beurteilten, grösser (100 verglichen mit 76 in der Gruppe „Klägerin“). Intuitiv ergibt es Sinn, dass die Verteilung in diesem Fall nicht mehr rein zufällig ist – der Anteil der vergleichsbereiten Richter macht in der einen Gruppe 56,6 %, in der anderen 43 % aus; zu erwarten wären 48 % in jeder Gruppe. Wie sehr die Verteilung von der unter der Nullhypothese erwarteten gleichmässigen Verteilung abweicht, kann (bei kategorialen Daten wie im Beispiel) der χ^2 -Test (Chi-Quadrat-Test) beantworten.³⁵³ In diesem Fall beträgt der χ^2 Wert (bei einem Freiheitsgrad) 3,187. Dies entspricht einem p-Wert von 0,074.³⁵⁴ Mit anderen Worten würde man, wenn man den gleichen Versuch 100 Mal wiederholen würde, in etwas mehr als sieben Fällen Daten beobachten, die gleich oder extremer als die beobachteten Daten sind. Der „heilige Gral“ der statistischen Signifikanz wird daher in diesem Fall knapp verfehlt.

- 169 Aus einem Signifikanztest, der zu einem Ergebnis gelangt, das „auf dem 5 % Niveau“³⁵⁵ signifikant ist, darf man unter keinen Umständen schliessen, dass die Arbeitshypothese mit einer Wahrscheinlichkeit von 95 % zutrifft (oder die Nullhypothese mit 95 %-iger Wahrscheinlichkeit falsch ist). Der p-Wert gibt die Wahrscheinlichkeit an, dass die Daten so extrem oder extremer als beobachtet sind *unter der Annahme, dass die Nullhypothese wahr ist*; formal $P(\text{extremere Daten} \mid \text{Nullhypothese})$.³⁵⁶ Die bedingte Wahrscheinlichkeit, dass die Nullhypothese falsch ist, ist $P(\text{Nullhypothese} \mid \text{extremere Daten})$, und diese Wahrscheinlichkeit ist *nicht* $1 - P(\text{extremere Daten} \mid \text{Nullhypothese})$. Dieser Fehler, im strafrechtlichen Zusammenhang als „Trugschluss des Anklägers“ bezeichnet, ist nicht selten (mehr dazu hinten, S. 135 ff.).
- 170 Ob die Nullhypothese falsch ist, hängt nicht nur von den beobachteten Daten ab, sondern auch davon, wie wahrscheinlich die Nullhypothese *a priori* war, d.h. bevor die beobachteten Daten bekannt wurden. Ob die Verwerfung der Nullhypothese bedeutet, dass die Arbeitshypothese (mit hoher Wahrscheinlichkeit) zutrifft, hängt weiterhin davon ab, ob nur die Arbeitshypothese die beobachteten Daten erklären kann oder auch eine andere – möglicherweise nicht bedachte – Hypothese. Wie wahrscheinlich die Nullhypothese *a priori* ist und welche anderen Hypothesen die beobachteten Daten erklären können, lässt sich – anders als der p-Wert – nicht mathematisch berechnen. Das Testen von Hypothesen unter Berücksichtigung der Anfangswahrscheinlichkeit und alternativer Hypothesen (nach dem Bayes-Theorem, näheres hinten, S. 125 ff.) führt daher ein subjektives Element in die anscheinend so objektive Statistik ein. Dies mag mit ein Grund sein, warum in den meisten sozialwissenschaftlichen Veröffentlichungen nur der p-Wert angegeben wird.
- 171 Neben dem χ^2 Test für kategoriale Daten werden in dieser Arbeit der t-Test (für normalverteilte Daten) und der Wilcoxon-Rangsummen-Test (resp. der äquivalente Mann-

³⁵³ FAHRMEIER/KÜNSTLER/PIGEOT/KUNZ, FN 348, 445 ff.

³⁵⁴ Wie ein Blick in die entsprechende Tabelle in einem Statistik-Lehrbuch (z.B. FAHRMEIER/KÜNSTLER/PIGEOT/KUNZ, FN 348, 583 ff.), oder, heutzutage einfacher, das Computerprogramm zeigt.

³⁵⁵ Der p-Wert, ab dem von „Signifikanz“ gesprochen wird, bezeichnet man als „Signifikanzniveau“, daher die Formulierung. Dass dieses Niveau bei 5 % liegt, ist eine mehr oder weniger arbiträre Konvention, an die man sich in den Sozialwissenschaften weitgehend hält.

³⁵⁶ ZEISEL/KAYE, FN 350, 81.

Whitney-U-Test) für ordinale, nicht parametrische Daten verwendet.³⁵⁷ Alle Berechnungen wurden mit SPSS 11.0 für Windows durchgeführt.

- 172 Wo dies möglich ist (für nicht nominale Daten) werden das arithmetische Mittel, der Median und die Standardabweichung angegeben. Das arithmetische Mittel – die Summe der Werte geteilt durch die Anzahl der Werte – wird dem üblichen Sprachgebrauch gemäss als Durchschnitt oder Schnitt bezeichnet. Der Median ist der Wert, der in der Mitte der Reihe liegt, wenn man eine Reihe von Messwerten der Größe nach sortiert.³⁵⁸ Die eine Hälfte der Werte ist größer, die andere Hälfte kleiner als der Median. Im Gegensatz zum Durchschnitt verändert sich der Median durch einzelne Extremwerte kaum; er ist daher geeigneter als der Durchschnitt, die Verteilung der Daten zu beschreiben, wenn einzelne Extremwerte den Durchschnitt verzerren (dies ist beispielsweise bei den Resultaten zum Ankereffekt der Fall, S. 83 ff.). Die Standardabweichung schliesslich ist ein Mass für die Streuung der Daten um das arithmetische Mittel. Je grösser die Standardabweichung, desto weiter weg sind die einzelnen Werte im Schnitt vom arithmetischen Mittel. Da für die Berechnung der Standardabweichung die Abweichungen vom Mittelwert quadriert werden, steigt die Standardabweichung bei einzelnen Ausreissern schnell stark an.³⁵⁹

III. Interne Validität

- 173 Eine Studie ist intern valide, wenn die manipulierte Variable (das Treatment oder die unabhängige Variable) tatsächlich für die Varianz der abhängigen Variablen verantwortlich ist. Werden die Messwerte hingegen durch einen oder mehrere Störfaktoren verändert, dann ist die interne Validität, oder Gültigkeit, verletzt. Sind sowohl Störfaktoren als auch der Stimulus für die beobachteten Effekte verantwortlich, spricht man von einer „Konfundierung“ der Effekte.³⁶⁰
- 174 Wenn der einzige (systematische) Unterschied zwischen Versuchs- und Kontrollgruppe im Treatment besteht, ist die Annahme zulässig, dass ein beobachteter Unterschied der abhängigen Variablen zwischen Versuchs- und Kontrollgruppe auf das Treatment zurückzuführen ist. Durch Zufallszuweisung (Randomisierung) von Untersuchungseinheiten in Versuchs- und Kontrollgruppe kann garantiert werden, dass die Unterschiede zwischen Versuchs- und Kontrollgruppe rein zufällig sind.³⁶¹ Bei der vorliegenden Studie wurden die Richterinnen und Richter zufällig der Versuchs- oder Kontrollgruppe zugewiesen. Zwar wurden keine „echten“ Zufallszahlen verwendet, um die Zuweisung vorzunehmen, aber die Fragebogen wurden gemischt und dann zusammen mit dem Begleitbrief in Couverts verpackt. Es gibt keinen Grund anzunehmen, dass sich die Richter, die eine Version der Frage erhalten haben, systematisch von den Richtern, die eine andere Version erhalten haben, unterscheiden. Wenn ein Unterschied in der abhängigen Variablen besteht – wenn z.B. die

³⁵⁷ Mehr zu diesen Tests bei FAHRMEIER/KÜNSTLER/PIGEOT/KUNZ, FN 348, 437 ff. (t-Test) und 459 ff. (Wilcoxon-Rangsummen-Test).

³⁵⁸ FAHRMEIER/KÜNSTLER/PIGEOT/KUNZ, FN 348, 55.

³⁵⁹ Zu Berechnung der Standardabweichung siehe FAHRMEIER/KÜNSTLER/PIGEOT/KUNZ, FN 348, 69.

³⁶⁰ SCHNELL/HILL/ESSER, FN 344, 219.

³⁶¹ SCHNELL/HILL/ESSER, FN 344, 223.

Besonderer Teil

Richter, bei denen der Kläger eine hohe Genugtuungssumme beantragt, systematisch mehr zusprechen als die Richter, denen kein bezifferter Antrag vorliegt – darf man daher annehmen, dass der Unterschied zwischen den beiden Gruppen auf das Treatment, in diesem Fall den Antrag des Klägers, zurückzuführen ist. M. a.W. gibt es keinen vernünftigen Grund, anzunehmen, dass die vorliegende Studie nicht intern valide ist.

IV. Externe Validität

- 175 Problematischer als die interne ist die externe (oder ökologische) Validität der vorliegenden Studie. Diese ist definiert als Möglichkeit der Generalisierung der experimentellen Resultate auf andere Personen(-gruppen) und Situationen.³⁶² Dass Richter eine höhere Genugtuungssumme zusprechen, wenn sie nach der Lektüre eines Sachverhalts von einer halben Seite mit einem Antrag des Klägers von Fr. 3 Mio. konfrontiert werden, ist zwar interessant. Eigentlich interessiert aber natürlich, ob dieser Effekt auch in einem Gerichtsverfahren wirkt, in dem sehr viel mehr und andere Informationen zur Verfügung stehen als in einem kurzen schriftlichen Sachverhalt.
- 176 Einige Richterinnen und Richter bezweifeln, dass es möglich ist, aufgrund der Studie Aussagen über reale juristische Entscheidungen zu machen. „Ich möchte Ihre Fachkompetenz nicht anzweifeln, habe aber Mühe mit der Vorstellung, dass Antworten auf diese fünf Fragen gültige Aussagen in Bezug auf juristische Entscheidungen in der Realität geben können!“ schrieb eine Richterin. Andere schickten den Fragebogen unausgefüllt zurück mit der Bemerkung, dass „die Fragen weit an den Entscheidungssituationen vorbeigehen, welche mir als Richter begegnen“ oder sie nicht in der Lage seien, aufgrund der spärlichen Unterlagen ein einigermaßen seriöses Urteil zu fällen und daher auch nicht glaubten, mit dem Ausfüllen des Fragebogens einen Beitrag zur Erforschung der richterlichen Entscheidungsfindung leisten zu können. Diese kritische Haltung gegenüber sozialwissenschaftlicher Forschung ist nicht untypisch für Richter.³⁶³
- 177 Kritik an rechtspsychologischer Forschung mit kurzen schriftlichen Fallskizzen (so genannten „Vignetten“) wurde wiederholt geäußert und ist ernst zu nehmen. Am pointiertesten haben wohl KONECNI/EBBESEN die Kritik vorgebracht.³⁶⁴
- 178 Die Gründe, die gegen die externe Validität solcher Studien sprechen, sind wohlbekannt. Die Entscheidungen, die von den Versuchspersonen in Simulationsstudien getroffen werden, haben keine Konsequenzen. Die Fallskizzen reduzieren die Art und den Umfang der zur Verfügung stehenden Informationen in einem Ausmass, die als Karikatur echter Gerichtsfälle – die sich eher durch ein Übermass an Informationen auszeichnen – erscheint. Häufig werden die Stimuli zudem in „zerlegter“ Form dargeboten, um einzelne Faktoren manipulieren zu können. Die Analyse eines komplexen Sachverhalts ist aber in der Praxis häufig der schwierigste Teil der juristischen Entscheidungsfindung. Oft fehlen in Experi-

³⁶² SCHNELL/HILL/ESSER, FN 344, 219.

³⁶³ RICHARD E. REDDING/N. DICKSON REPPUCH, Effects of Lawyers' Socio-political Attitudes on Their Judgments of Social Science in Legal Decision Making, *Law and Human Behavior* 1999, 31-54, 47.

³⁶⁴ VLADIMIR J. KONECNI/EBBE B. EBBESEN, Social Psychology and the Law: The Choice of Research Problems, Settings, and Methodology, in: KONECNI/EBBESEN (Hrsg.), FN 265, 27-44.

menten prozedurale Schritte wie die Beratung in der Gruppe, die einen Einfluss auf das Urteil haben können. Schliesslich entsprechen auch die abhängigen Variablen – beispielsweise die Beurteilung der Schuld auf einer numerischen Skala – nicht den Entscheidungen (schuldig ja/nein), die in Wirklichkeit zu treffen sind.³⁶⁵

- 179 KONECNI/EBBESEN gehen davon aus, dass der Urteilende eine aufgabenspezifische (*task-specific*) Entscheidungsregel für jede konkrete Entscheidung konstruiert.³⁶⁶ Gemäss dieser Auffassung kann es keine Theorie geben, die a priori erklärt, wie ein Urteilender eine bestimmte Aufgabe anpacken wird. KONECNI/EBBESEN stellen daher nicht nur die externe Validität von „*paper and pencil*“ Studien in Frage, sondern auch diejenige realistischerer Studien, die zum Beispiel Videoaufzeichnungen der Aussagen von Parteienanwälten und Zeugen verwenden und Beratung vor der Entscheidung erlauben.³⁶⁷ Sie untermauern diese Aussage empirisch mit drei eigenen Simulationsstudien, die anschliessend mit den Resultaten von Feldstudien verglichen wurden.³⁶⁸ In keinem Fall, so EBBESEN/KONECNI, hätten die Resultate der Simulationsstudien erlaubt, die Resultate der Feldstudien richtig vorauszusagen.³⁶⁹ Sie kommen daher zum Schluss, dass es „gefährlich und nahezu unverantwortlich [ist], auf der Grundlage von Simulationsstudien mit von ihrem lebenswirklichen Kontext losgelösten Effekten Schlussfolgerungen und Empfehlungen für das Rechtssystem zu formulieren“.³⁷⁰
- 180 Für KONECNI/EBBESEN können nur Feldstudien, d.h. Untersuchungen in der Lebenswirklichkeit, zu validen Resultaten kommen. Dabei ziehen sie Archivstudien der Beobachtung von Gerichtsverhandlungen aus verschiedenen methodischen Gründen vor,³⁷¹ wobei sie empfehlen, dass das Rechtssystem differenzierte Verfahren zur Erhebung von Daten zum eigenen Handeln am besten gleich einbaut.³⁷² Die Schlussfolgerungen und Empfehlungen von KONECNI/EBBESEN werden von Vertretern theorielastiger Soziologie oder Psychologie begrüsst; so z.B. von LÖSCHPER in ihrer diskursanalytischen Habilitationsschrift „Bausteine für eine Theorie richterlichen Urteilens“, die Simulationsstudien als „Experimente im sozialen Vakuum“ kritisiert.³⁷³
- 181 Trotz der Kritik von KONECNI/EBBESEN und anderen, bleiben Simulationsstudien unvermindert populär in der rechtspsychologischen Forschung. Häufig werden die beson-

³⁶⁵ KONECNI/EBBESEN, FN 364, 28.

³⁶⁶ VLADIMIR J. KONECNI/EBBE B. EBBESEN, A Critique of Theory and Method in Social-Psychological Approaches to Legal Issues, in: BRUCE DENNIS SALES (Hrsg.), *The Trial Process*, New York 1981, 481-498, 488.

³⁶⁷ KONECNI/EBBESEN, FN 364, 31.

³⁶⁸ EBBE B. EBBESEN/VLADIMIR J. KONECNI, On the External Validity of Decision-Making Research: What Do We Know About Decisions in the Real World? in: THOMAS S. WALLSTEN (Hrsg.), *Cognitive Processes in Choice and Decision Behavior*, Hillsdale 1980, 21-45.

³⁶⁹ EBBESEN/KONECNI, FN 368, 25 ff.

³⁷⁰ VLADIMIR J. KONECNI/EBBE B. EBBESEN, Methodische Probleme in der Forschung über juristische Entscheidungsprozesse – unter besonderer Berücksichtigung experimenteller Simulationen, *Gruppendynamik* 1991, 175-188, 179.

³⁷¹ KONECNI/EBBESEN, FN 366, 493; KONECNI/EBBESEN, FN 364, 35 f.

³⁷² KONECNI/EBBESEN, FN 370, 185.

³⁷³ GABRIELE LÖSCHPER, *Bausteine für eine psychologische Theorie richterlichen Urteilens*, Baden-Baden 1999, 20.

Besonderer Teil

ders unrealistischen „*paper and pencil*“ Studien mit Studierenden als Versuchspersonen durchgeführt. KERR und BRAY haben 1982 die Methoden von 72 Simulationsstudien mit Geschworenen (*mock juries*) untersucht und festgestellt, dass rund 50 % der Studien mit kurzen schriftlichen Fallskizzen und studentischen „Geschworenen“ durchgeführt wurden.³⁷⁴ Der Anteil von Simulationsstudien an den in *Law and Human Behavior* zwischen 1977 und 1996 publizierten Studien zum Verhalten von Geschworenen ist effektiv gestiegen, von rund 50 % auf 70-80 %. Der Anteil der Simulationsstudien mit Studierenden und kurzen Fallskizzen ist ebenfalls gestiegen, zu Lasten aufwändigerer Studien mit Videoaufnahmen und Bürgern, die als Geschworene qualifiziert sind (*jury eligible citizens*).³⁷⁵ Der Trend geht also genau in die entgegengesetzte als die von den Kritikern geforderte Richtung.

- 182 Es gibt zynische, methodologische und pragmatische Gründe für die ungebrochene Popularität von Simulationsstudien. Der zynische Grund ist, dass nur akademische Karriere macht, wer viel publiziert. Studien mit Sachverhaltsskizzen und studentischen Versuchspersonen sind viel schneller und billiger als aufwändige Simulations- oder gar Feldstudien, die in der gleichen Anzahl Publikationen resultieren.³⁷⁶ Solange die Anzahl der Publikationen ein ausschlaggebendes Kriterium für akademische Berufungen ist, werden einfache Simulationsstudien beliebt bleiben.
- 183 Der methodologische Grund für Simulationsstudien liegt darin, dass die experimentelle Kontrolle, die der Versuchsleiter bei Simulationen ausüben kann, grosse Vorteile hat. Sie führt, wie oben dargelegt, in der Regel zu einer hohen internen Validität. Ein Einfluss konfundierender Variablen kann meist ausgeschlossen werden.³⁷⁷ Wollte man beispielsweise den Einfluss des klägerischen Antrags auf die zugesprochene Genugtuungssumme im Feld studieren, so müsste man alle objektiven Faktoren, die die Höhe der Genugtuung beeinflussen (wie Schwere der Verletzung des Opfers und Schwere des Verschuldens des Täters) konstant halten. Tut man dies nicht – wie beispielsweise die beiden in Fn. 422 und 425 erwähnten spanischen Studien – so ist eine beobachtete Korrelation von Antrag des Klägers und zugesprochener Genugtuungssumme vermutlich nicht auf den Ankereffekt, sondern auf die Schwere der Verletzung zurückzuführen – wer schwerer verletzt ist, verlangt mehr und erhält, zu Recht, mehr zugesprochen. Es gibt statistische Techniken, den Einfluss möglicherweise konfundierender Variablen konstant zu halten, z.B. die multiple Regressionsanalyse. Diese setzt aber eine relativ hohe Anzahl Fälle voraus, um zu valablen Resultaten zu gelangen. Da die Zuspreehung von Genugtuung in der Schweiz nicht sonderlich häufig ist, müsste man die Rechtsprechung zahlreicher Kantone während mehrerer Jahre auswerten, um überhaupt genügend Daten für eine valide statistische Auswertung zu erhalten.

³⁷⁴ ROBERT M. BRAY/NORBERT L. KERR, *Methodological Considerations in the Study of the Psychology of the Courtroom*, in: NORBERT L. KERR/ROBERT M. BRAY (Hrsg.), *The Psychology of the Courtroom*, San Diego etc. 1982, 287-323, 294.

³⁷⁵ BRIAN H. BORNSTEIN, *The Ecological Validity of Jury Simulations: Is the Jury Still Out?*, *Law and Human Behavior* 1999, 75-92, 87.

³⁷⁶ KONECNI/EBBESEN, FN 366, 486.

³⁷⁷ BRAY/KERR, FN 374, 296; AMIRAM ELWOK/BRUCE DENNIS SALES/DAVID SUGGS, *The Trial: A Research Review*, in: SALES (Hrsg.), FN 366, 1-59, 51.

- 184 Ein weiterer Vorteil von Simulationsstudien ist, dass sie – anders als Feldstudien – repliziert werden können.³⁷⁸ Während eine einzelne Studie nie genügt, um externe Validität zu generieren, kann die systematische und variierte Replikation einer Studie dazu beitragen, dass man darauf vertrauen darf, dass die Resultate auch in Situationen, die nicht konkret untersucht wurden, Gültigkeit haben.³⁷⁹ Viele der in der vorliegenden Studie gestellten Fragen wurden beispielsweise schon Studierenden gestellt. Zeigen zu können, dass auch Richterinnen und Richter teilweise die gleichen Fehler machen wie studentische Versuchspersonen, ist ein weiterer Puzzlestein, der die externe Validität der den Fragen zu Grunde liegenden psychologischen Theorie stützt. Dabei besteht ein offensichtlicher Trade-off zwischen Realitätsnähe des Stimulus-Materials (also Vignetten statt den gesamten Akten eines Prozesses oder der Videoaufzeichnung mündlicher Plädoyers und Zeugenaussagen) und der Realitätsnähe der Versuchspersonen. Wer sich dafür entscheidet, mit Richtern als Versuchspersonen zu arbeiten, muss sich fast zwangsweise auf kurze Sachverhalte beschränken. Richter sind nicht bereit, sich mehrere Stunden in einen Sachverhalt einzulesen, um an einer sozialwissenschaftlichen Studie teilzunehmen. Die Vermutung sei erlaubt, dass die gleichen Richter, die an der vorliegenden Studie nicht teilgenommen haben, weil sie zu unrealistisch sei, an einer realistischeren Studie nicht teilgenommen hätten, weil sie den dafür notwendige Zeitaufwand als unzumutbar erachten.
- 185 Den Vorteilen der Simulationsstudien stehen Nachteile der Feldstudien gegenüber. Erstens sind gerade die von KONECNI/EBBESSEN favorisierten Archivstudien wegen datenschutzrechtlicher Bestimmungen in Europa häufig schwierig oder gar nicht durchzuführen. Auch werden zahlreiche interessierende Faktoren durch die gerichtlichen Akten nicht erhoben – beispielsweise wird in der Schweiz das letzte Vergleichsangebot der Parteien vor dem Prozess nicht erfasst, so dass es ausgeschlossen ist, das letzte Vergleichsangebot mit der im Urteil zugesprochenen Summe zu vergleichen, wie dies RACHLINSKI getan hat (mehr dazu hinten, S. 92 ff.).³⁸⁰ Die Fallzahlen sind, wie bereits erwähnt, oft zu gering, um quantitative Aussagen machen zu können. Selbst in Deutschland beschränken sich Feldstudien auf häufige Delikte wie den einfachen Diebstahl, da andere Fälle nicht oft genug vorkommen.³⁸¹ In der Schweiz ist dieses Problem noch gravierender als in Deutschland. So ist es beispielsweise ausgeschlossen, mittels einer Feldstudie quantitative Aussagen über den Einfluss einer Gesetzesänderung auf die Anzahl der jährlich ausgesprochenen Verwahrungen zu machen (dazu S. 243 ff.), wenn jährlich im Schnitt nur rund 20 stationäre Massnahmen und Verwahrungen ausgesprochen werden.³⁸² Schliesslich stellen die schriftlichen Zeugnisse, die das Justizsystem produziert, keinen repräsentativen Querschnitt der tatsäch-

³⁷⁸ BRAY/KERR, FN 374, 296.

³⁷⁹ BRAY/KERR, FN 374, 301.

³⁸⁰ RACHLINSKI, FN 219.

³⁸¹ OSWALD, FN 260, 15.

³⁸² Bundesamt für Statistik, Strafurteilsstatistik, Stand der Datenbank 12. August 2004, Verurteilungen nach Art der Massnahme (der Durchschnitt der Verwahrungen nach Art. 42 StGB [Gewohnheitsverbrecher] und der stationären Massnahmen und Verwahrungen nach Art. 41 Ziff. 1 [psychisch abnorme Täter] der letzten fünf Jahre [1999-2003, neuste verfügbare Zahlen] betrug 22,6; die Statistik unterscheidet bei den Massnahmen nach Art. 43 Ziff. 1 leider nicht zwischen stationären Massnahmen und Verwahrungen).

Besonderer Teil

lichen Fälle dar.³⁸³ Die – faktisch überaus zahlreichen – Fälle, die verglichen werden, hinterlassen entweder gar keine (bei aussergerichtlichem Vergleich) oder nur geringe (bei gerichtlichem Vergleich) Spuren in den Gerichtsarchiven, aus denen sich kaum wertvolle Erkenntnisse gewinnen lassen. Wenn man sich daher für den Einfluss psychologischer Faktoren auf die Vergleichsempfehlung durch Richter interessiert (S. 93 ff.), kann man sich nicht auf Archivstudien verlassen.

- 186 Der pragmatische Grund für Simulationsstudien sind die prohibitiv hohen Kosten von Feldstudien. Die konsequente Umsetzung der Forderung von KONECNI/EBBESEN würde dazu führen, dass kaum mehr rechtspsychologische Forschung betrieben würde. Forschungsgelder für gross angelegte Feldstudien sind kaum vorhanden. Wie BRAY/KERR bemerken: „Wir werden nicht viele Erkenntnisse gewinnen, wenn wir jede Forschung verbieten ausser derjenigen, die prohibitiv teuer ist für die meisten Forschenden“.³⁸⁴ Eine Simulationsstudie mit möglicherweise eingeschränkter externer Validität ist dieser Auffassung nach, der ich mich anschliessen möchte, gar keiner Studie immer noch vorzuziehen.³⁸⁵ Selbst KONECNI/EBBESEN können auf Simulationsstudien nicht verzichten: rund ein Drittel der empirischen Studien, die in der von ihnen herausgegebenen Aufsatzsammlung „*The Criminal Justice System*“ erschienen sind, sind Simulationsstudien.
- 187 Bisher wurde dargelegt, warum Simulationsstudien *trotz* ihrer eingeschränkten Validität einer Feldstudie unter Umständen vorzuziehen sind. Wie aber steht es denn nun um die externe Validität von Simulationsstudien? Die Frage, die sich nur empirisch beantworten lässt, ist nur ungenügend erforscht. KONECNI/EBBESEN kommen wie erwähnt zu einem eher pessimistischen Resultat. Andere Forscher haben nicht die Resultate von Feldstudien mit derjenigen von Simulationsstudien verglichen, sondern die Resultate unterschiedlich realitätsnaher Simulationen. Die Wahl der Versuchspersonen, also Studierende (*college undergraduates*) statt als Geschworene wählbare Bürger (*jury eligible citizens*), hat nach den übereinstimmenden Resultaten von mehreren Dutzend Studien keinen Einfluss auf die beobachteten Effekte.³⁸⁶ Erstaunlicherweise hat auch die Verwendung von Videotapes statt kurzer Sachverhaltsskizzen als Stimulus-Material keinen wesentlichen Einfluss auf die Resultate; allerdings ist die Anzahl der Studien, die dies untersuchen, geringer und die Ergebnisse nicht so eindeutig wie bei den Versuchspersonen.³⁸⁷ Wer wie KONECNI/EBBESEN jede Form von Simulation ablehnt, da Entscheidungsregeln aufgabenspezifisch und situativ seien, wird sich dadurch natürlich nicht überzeugen lassen. Wer diesen Standpunkt allerdings ernsthaft vertritt, muss auf jede generelle Theorie zur Erklärung menschlichen Entscheidungsverhaltens unter Unsicherheit, und damit auf jegliche Generalisierbarkeit, verzichten. Für diejenigen, die nicht so weit gehen wollen, ist die Tatsache, dass sich die Ergebnisse von Studien mit kurzen Sachverhaltsskizzen auf aufwändigere

³⁸³ SHARI SEIDMAN DIAMOND, *The Challenges of Socio-Legal Research on Decision Making: Psychological Successes and Failures*, *Journal of Law and Society* 1995, 78-84, 79 f.

³⁸⁴ BRAY/KERR, FN 374, 316: „We do not stand to gain much knowledge by prohibiting all research except that which is restrictingly expensive for most investigators“.

³⁸⁵ Ebenso ELWORK/SALES/SUGGS, FN 377, 1-59, 51

³⁸⁶ BORNSTEIN, FN 375, 78.

³⁸⁷ BORNSTEIN, FN 375, 82.

Studien mit Videoaufnahmen generalisieren lassen, zumindest ein Indiz dafür, dass es um die externe Validität von „*paper and pencil*“ Studien nicht so schlecht stehen kann.

- 188 Im vorliegenden Fall kommt hinzu, dass sich die Fragestellungen auf ein psychologisches Forschungsprogramm stützen, das seit dreissig Jahren mit einer enormen Flut von empirischen Studien erforscht wird. Obwohl die Randbedingungen der einzelnen Urteilsheuristiken weiterhin umstritten sind, handelt es sich bei den hier besprochenen Phänomenen doch um Effekte, die allgemein als robust gelten.³⁸⁸ Insofern die Resultate der Studie bestimmte Denkweisen aufzeigen, ist es wahrscheinlich, dass die Richter auch bei der Beurteilung tatsächlicher Fälle gleich denken, zumal sie sich der Effekte meist nicht bewusst sein dürften. Erhöhte Motivation allein ist kein Allheilmittel. Nur wenn die erhöhte Aufmerksamkeit und die längere Überlegungszeit dazu führen, dass die kognitive Täuschung erkannt und vermieden wird, verbessert sich die Entscheidungsqualität. Dies darf nicht leichthin angenommen werden.³⁸⁹ Der beste Beweis dafür sind Studien, in denen kognitive Illusionen selbst dann die Entscheidungen negativ beeinflussen, wenn es für die Probanden um sehr viel Geld geht. Der Darstellungseffekt beispielsweise wirkt sogar, wenn die (chinesischen) Testpersonen um Einsätze in der Höhe von zwei Monatslöhnen spielen.³⁹⁰
- 189 Kognitive Täuschungen wurden auch in Feldstudien ausserhalb des Labors und/oder in kontrollierten Experimenten mit sehr realistischem Stimulus-Material nachgewiesen.³⁹¹ Kognitive Täuschungen beeinflussen auch das Urteil von Experten ausserhalb des Labors.³⁹² Die Häufigkeit des Vorkommens einer Krankheit wird von Ärzten auch dann nicht genügend beachtet, wenn sie in der Klinik Patienten diagnostizieren.³⁹³ MUSSWEILER und Kollegen schickten einen Schauspieler mit einem echten Opel Kadett bei 60 (uneingeweihten) Automechanikern und –händlern vorbei, um seinen Wert schätzen zu lassen. Die Schätzung wurde erheblich durch den vom Schauspieler manipulierten Ankereffekt beeinflusst.³⁹⁴ Sehr viel realitätsnäher kann eine Studie nicht mehr sein. Diese Generalisierbarkeit der Resultate der Forschung zu den *heuristics and biases* auf viele natürliche Entscheidungssituationen legt es nahe anzunehmen, dass sie sich auch auf Juristen und juristische Entscheidungen generalisieren lassen. Natürlich kann man die Auffassung vertreten, dass Richter anders als alle anderen sind und den Täuschungen, denen fast alle Experten unterliegen, nicht unterliegen, oder dass spezifische Elemente des gerichtlichen Verfahrens verhindern, dass sich kognitive Täuschungen auf juristische Urteile auswirken. Wer diese Behauptung aber aufstellt, trägt in Anbetracht der empirischen Forschung der letzten dreis-

³⁸⁸ Detaillierte Nachweise bei den einzelnen Heuristiken.

³⁸⁹ AMOS TVERSKY/DANIEL KAHNEMAN, FN 325, 274.

³⁹⁰ KACHELMEIER/SHEHATA, FN 304.

³⁹¹ Siehe generell die im Kapitel “Real-World Applications” abgedruckten Studien in: GILOVICH/GRIFFIN/KAHNEMAN, 601-762; CAMERER, FN 303.

³⁹² RACHLINSKI, FN 456, 154; DEREK J. KOEHLER/LYLE BRENNER/DALE GRIFFIN, The Calibration of Expert Judgment: Heuristics and Biases Beyond the Laboratory, in: GILOVICH/GRIFFIN/KAHNEMAN (Hrsg.), 686-715.

³⁹³ DAWES, FN 616, 425.

³⁹⁴ THOMAS MUSSWEILER/FRITZ STRACK/TIM PFEIFFER, Overcoming the Inevitable Anchoring Effect: Considering the Opposite Compensates for Selective Accessibility, Personality and Social Psychology Bulletin 2000, 1142-1150, 1146

Besonderer Teil

sig Jahre die Beweislast dafür.³⁹⁵ Mit der Behauptung alleine ist es nicht getan, solange ihr jegliche empirische Basis fehlt.

- 190 Zusammengefasst gesagt sind Feldstudien Simulationsstudien dort vorzuziehen, wo sie aus ökonomischen Gründen durchführbar und entsprechende Daten vorhanden sind. In den häufigen Fällen, in denen dies nicht der Fall ist, sind Simulationsstudien eine *second best* Lösung mit eigenen Vorteilen, namentlich der grösseren internen Validität von echten Experimenten. Die Resultate aufwändiger Simulationsstudien unterscheiden sich dabei kaum von simplen Studien, die kurze Sachverhaltsskizzen als Stimulus-Material verwenden. Kognitive Täuschungen wurden in einer Reihe von Studien ausserhalb des Labors unter natürlichen Bedingungen und bei Experten beobachtet, so dass man davon ausgehen darf, dass es sich um robuste, nicht nur unter Laborbedingungen auftretende Phänomene handelt. Wer die Auffassung vertritt, dass sie ausgerechnet im Justizsystem keine Rolle spielen, muss in Anbetracht der Erkenntnisse aus dreissig Jahren Forschung den empirischen Nachweis dafür erbringen.

³⁹⁵ Zur „Beweislastumkehr“ beim Vorliegen empirischer rechtssoziologischer Daten MANFRED REHBINDER, Rechtssoziologie, 5. Aufl. München 2003, Rz. 24 und 59.