Die empirische Herangehensweise im Zivilrecht

Lebensnähe und Methodenehrlichkeit für die juristische Analytik?

von Wiss. Mit. Dr. Dr. *Hanjo Hamann*
und Wiss. Mit. *Leonard Hoeft*, Bonn*

Inhaltsübersicht

I. Einleitung ... 311
II. Vorklärung: Was bedeutet „empirische Herangehensweise"? ... 313
III. Empirie-Rezeption als genuin juristische Herangehensweise ... 317
1. Formulierung der Fragestellung ... 318
2. Auffinden früherer Forschung ... 321
3. Würdigung der Erkenntnisse ... 323
4. Verarbeiten der Feststellungen ... 328
IV. Praktische Instrumente einer „empirischen Herangehensweise" ... 328
1. Sprachwissenschaft (Verständlichkeitstests) ... 329
2. Umfrageforschung (Demoskopie) ... 330
3. Informelle Befragung (Indizienverfahren) ... 331
4. Experimentelle Befragung (Vignettenstudien) ... 334
V. Fazit: Perspektiven der Empirie im Zivilrecht ... 335

I. Einleitung

Dass ein statistisches Grundverständnis zu den „analytischen Methoden für Juristen" gehöre,[1] wird auch in Deutschland vermehrt angenommen.[2] Entgegen der ebenfalls verbreiteten Lesart des Sprichworts vom Richter, der nicht

* Die Autoren danken Christoph Engel, Andreas Engert, Jens Frankenreiter, Meirav Furth, Fabian Iwanczik, Peter McColgan, Alexander Morell und Rima-Maria Rahal für Unterstützung und hilfreiche Verbesserungsvorschläge.

[1] *Jackson/Kaplow/Shavell/Viscusi/Cope*, Analytical Methods for Lawyers, 2003, Kap. 8, 9.

[2] Vgl. die Lehrveranstaltungen von *Eidenmüller/Engel*, Analytische Methoden für Juristen § 13 Statistik, Universität München WiSe 2014/15, Veranstaltungsnr. 03023 (mit Folien unter www.horst-eidenmueller.de/upload/folien_methodenlehre_wise_1415.pdf, 335-378); *Kuntz*, Analytical Methods for Lawyers C. Statistics and legal analysis, Universität Bremen WiSe 2016/17, Kursnr. 06-027-7-728; *Frankenreiter*, Analytische Methoden für Juristen, Humboldt-Universität Berlin SoSe 2017, Veranstaltungsnr. 10520; dazu zuletzt *Hamann*, JURA 2017, 759, 762 bei Fn. 27.

AcP 217, 311–336 – DOI: 10.1628/000389917X15002739282653
ISSN 0003-8997 – © Mohr Siebeck 2017

rechnet, stehen Rechtsmethodik und Statistik „gerade nicht ‚auf Kriegsfuß'", sondern ergänzen einander geradezu als Ausprägungen der „Fähigkeit zu analytischem Denken und logischer Schlussfolgerung".[3]

Dennoch sind praktische Beispiele für das Potential und die Grenzen statistischer (d.h. quantitativ-empirischer) Erhebungen in zivilrechtlichen Archivzeitschriften bislang rar.[4] Nun gibt ein neuer Vorschlag willkommenen Anlass zu weiterführenden Überlegungen: *Alexander Stöhr* plädiert am Beispiel der Transparenzkontrolle im Arbeitsrecht „für eine empirische Herangehensweise".[5] Dazu verwendet er zwei Klauseln aus Arbeitsverträgen, die das Bundesarbeitsgericht 2007 bzw. 2011 am Maßstab des § 307 Abs. 1 S. 2 BGB zu beurteilen hatte, und befragt knapp dreißigtausend Angehörige seiner Universität per E-Mail, wie sie die vermeintlich streitentscheidende Frage des Falls jeweils entschieden hätten. Aus den fast eintausend Antworten, die mehrheitlich von der des BAG abwichen, folgert Stöhr, dass der vom Bundesarbeitsgericht entwickelte „Transparenzmaßstab an der Realität vorbeigeht" (560), und schlägt deshalb vor, „die Paradigmen der Transparenzkontrolle grundlegend zu überdenken" (571). Erörterungen zu verschiedenen empirischen Instrumenten runden die Darstellung ab und geben wertvolle Impulse für weiterführende Überlegungen zur Zukunft einer empirischen Rechtsforschung.

Diese Impulse kommen genau zur rechten Zeit, denn jüngst wurde auch in den USA eine „dringende" und „radikale" methodische Erneuerung der „Vertragsauslegung durch Umfragen und Experimente" vorgeschlagen.[6] Sie soll insbesondere zur Beurteilung überraschender Klauseln in Verbraucherverträgen geeignet und zulässig sein,[7] aber auch im größeren Maßstab die richterliche Auslegung durch empirische Erhebungen anreichern oder gar erset-

---

[3] *W. Hamann*, in: FS Assenmacher 2012, 307, 308.

[4] Immerhin etwa *Nußbaum* AcP 154 (1955), 453 und in den letzten Jahren z.B. *Nowak/Rott/Mahr* ZGR 2005, 252 (Ereignisstudie); *Bayer/Hoffmann/Weinmann* ZGR 2007, 457 (Ereignisstudie); *Eidenmüller* ZGR 2007, 168 (Rechtstatsachen); *Woywode/Keese/Tänzler* ZGR 2012, 418 (Rechtstatsachen); *Braun/Eidenmüller/Engert/Hornuf* ZHR 177 (2013), 131 (Rechtstatsachen); *Morell* AcP 214 (2014), 387 (Metastudien).

[5] *Stöhr* AcP 216 (2016), 558; eingeklammerte Seitenzahlen im Haupttext beziehen sich auf diesen Text; bisweilen wird auch eine neuere Rezension desselben Autors zitiert (*Stöhr* AcP 217 [2017], 144), dies aber durchweg in Fußnoten.

[6] *Ben-Shahar/Strahilevitz*, Interpreting Contracts via Surveys and Experiments (University of Chicago Coase-Sandor Institute for Law & Economics Research Paper No. 791), SSRN 25.1.2017, online unter www.ssrn.com/abstract=2905873.

[7] *Ben-Shahar/Strahilevitz* (Fn. 6), 18: "The benchmark case for the application of the survey interpretation method is a consumer contract."; 47: "under the doctrine of unconscionability courts have to determine, among other things, whether a term is surprising and unexpected to the typical consumer. Under the American Law Institute's proposed Restatement of Consumer Contracts, this can be shown by survey evidence".

zen – in der Stoßrichtung identisch mit Stöhr, in Mitteln und Methodik jedoch grundlegend anders.

Die damit eröffnete Diskussion sei nun aufgegriffen (II.) und durch eine methodenkritische Auseinandersetzung mit dem Plädoyer Stöhrs fortgesetzt (III.). Daran schließen sich einige Reflexionen an, die in eigene Umsetzungsvorschläge münden (IV.).

II. Vorklärung: *Was bedeutet „empirische Herangehensweise"*?

Stöhr beginnt seine Ausführungen (559) unter Berufung auf die Sein-Sollen-Dichotomie, die nach Ansicht vieler Juristen das Reich der Rechtsdogmatik durch einen Limes der Epistemologie schützt: Hier die normative Wissenschaft, dort die deskriptiven. Die Rechtswissenschaft, so Stöhr, untersuche eben „nicht das Sein, sondern das Sollen" (559) – sei also „als normative Wissenschaft an sich nicht auf empirische Untersuchungen" angelegt (582).

Gleichwohl öffnet Stöhr einen Durchgang durch den schützenden Grenzwall, indem er aufzeigt, dass Juristen für bestimmte Fragen dennoch eine „empirische Herangehensweise" benötigen (573 ff.), soweit Gesetzesbegriffe auf tatsächliche Gegebenheiten verweisen. Sein zunächst sehr offener Begriff der „empirischen Herangehensweise" gewinnt schärfere Konturen, wenn man sich die von Stöhr vorgeschlagenen Methoden und seine eigene Studie näher anschaut. Sie passen nahtlos in jene Tradition, die 1914 als „Rechtstatsachenforschung"[8] begründet wurde: Dazu zählen „Beobachtungsstudien mit beschreibendem Erkenntnisinteresse", die das Zivilrecht empirisch anreichern, wobei sie „fast nur beschreibende Statistik verwenden".[9] Zu den zahlreichen Studien, die in über einhundert Jahren dieser Forschung vorgelegt wurden, gehörten schon früh solche zum Recht der Allgemeinen Geschäftsbedingungen[10] und zum Arbeits-

---

[8] *Nussbaum*, Die Rechtstatsachenforschung. Ihre Bedeutung für Wissenschaft und Unterricht, 1914; *ders.* (Fn. 4).

[9] Näher *Hamann*, Evidenzbasierte Jurisprudenz. Methoden empirischer Forschung und ihr Erkenntniswert für das Recht am Beispiel des Gesellschaftsrechts, 2014, 40 f. Fn. 258.

[10] Bspw. *Engling*, Geschäftsbedingungen und Handelsbräuche im deutschen Weinhandel. Eine rechtstatsächliche und rechtspolitische Untersuchung, 1936; *Schaudwet*, Bankenkontokorrent und Allgemeine Geschäftsbedingungen, 1967; *Keller*, Allgemeine Geschäftsbedingungen. Eine Rechtstatsachenuntersuchung in einigen Zweigen der schweizerischen Wirtschaft, 1970; *Rehbinder*, Das Kaufrecht in den Allgemeinen Geschäftsbedingungen der deutschen Wirtschaft, 2. Aufl., 1979; *Loßmann*, in: Chiotellis/Fikentscher, Rechtstatsachenforschung. Methodische Probleme und Beispiele aus dem Schuld- und Wirtschaftsrecht, 1985, 503–536.

vertragsrecht.[11] Deshalb verstand es sich für Rechtstatsachenforscher schon vor über dreißig Jahren „von selbst", dass „die Einstellung und das Verhalten" der durch Geschäftsbedingungen betroffenen Verkehrskreise „rechtserhebliche" Untersuchungsgegenstände seien.[12] Aus dieser Traditionslinie stammt die Methodik, die Stöhr nun erstmals auf die Transparenzkontrolle im AGB-Recht (§ 307 Abs. 1 Satz 2 BGB) anwendet.[13]

Stöhr hält es ganz grundsätzlich für „unzulässig", „das eigene Verständnis des Rechtsanwenders ohne nähere Erläuterung auf den objektiven Durchschnittsarbeitnehmer zu verallgemeinern" (571), und will durch empirische Ansätze zur „Methodenehrlichkeit" (559) beitragen. Das verdient für sich genommen große Zustimmung,[14] und entspricht insbesondere den bisherigen Postulaten rechtslinguistischer Forschung (nicht nur) zum Arbeitsrecht.[15] Ins Positive gewendet, bedeutet Stöhrs Ansinnen wohl dasselbe, was in einem Lehrbuch über „Juristisches Denken" so formuliert wurde:

> „Der Jurist benötigt die Erkenntnismethoden der Sozialwissenschaften, um das Problem zu erkennen, das zu lösen ist. Dafür setzt er sowohl die deskriptiven wie die normativen Methoden der Sozialwissenschaften ein […] Insoweit ist der Jurist Empirist im Sinne Webers."[16]

In diesem Zusammenhang stellt Stöhr dar, dass für eine „empirische Herangehensweise" sowohl teleologische[17] und verfassungsrechtliche[18] als auch ökonomische[19] Gründe sprächen. Da seine geraffte Darstellung indessen Missver-

---

[11] Vgl. *Lotmar*, Der Arbeitsvertrag, 2 Bde. 1902/08, jetzt 2. Aufl., 2001, 56 ff.

[12] *Schweizer*, in: Schweizer/Quitt, Rechtstatsachenermittlung durch Befragen Bd. 1, 1985, 35 ff., 40 (Sonderdr. aus *Chiotellis/Fikentscher* [Fn. 10 a.E.]).

[13] Stöhrs Methodik ähnelt zudem der „empirischen Sprachgebrauchsanalyse" im Strafrecht nach *Lorenz/Pietzcker/Pietzcker* NStZ 2005, 429.

[14] Zu Empirie und Methodenehrlichkeit in der Arbeitsrechtsprechung schon *Hamann*, in: Vogel (Hrsg.), Zugänge zur Rechtssemantik. Interdisziplinäre Ansätze im Zeitalter neuer Medien, Berlin 2015, 184, 203.

[15] Vgl. *Vogel/Pötters/Christensen*, Richterrecht der Arbeit – empirisch untersucht, Berlin 2015, 68 f.

[16] *Mastronardi*, Juristisches Denken. Eine Einführung, 2. Aufl., 2003, Rn. 334; in Rn. 335 mit der Gegenüberstellung zum „Normativist im Sinne Kelsens".

[17] Dazu bereits *Hamann* (Fn. 9), 35–36 m.w.N.

[18] *Stöhr* (576) will empirische Erkenntnisse in die Erforderlichkeitsprüfung bei der Rechtfertigung von Grundrechtseingriffen einfließen lassen; ebenfalls in Betracht kommen Geeignetheit (*Eidenmüller* JZ 1999, 53, 54: „Realwissenschaft kann [dem Gesetzgeber] unter Umständen zeigen, dass sich ein bestimmtes Ziel mit dem geplanten Mittel überhaupt nicht oder nur unter Inkaufnahme von unerwünschten Nebeneffekten erreichen lässt.") und Angemessenheit i.e.S. (*Petersen* Der Staat 2010, 435, 442 ff.: „Bei der Grundrechtsprüfung im Verfassungsrecht spielen empirische Erkenntnisse […] eine zentrale Rolle […] bei der Abwägung im Rahmen der Angemessenheit.").

[19] So zuletzt schon *Morell* (Fn. 4), 395 ff.

ständnisse hervorrufen könnte, sei zumindest eine Ergänzung und Klarstellung versucht:

Stöhr setzt durchweg „normativ" mit „subjektiv" und „empirisch" mit „objektiv" gleich und meint: „Empirische Erkenntnisse beruhen auf Daten, nicht auf bloßen Vermutungen, rechtspolitischen Vorstellungen oder Ideologien" (574).[20] Das legt aus rechtstheoretischer Warte, wo man um „den prekären Status der wissenschaftlichen Objektivität" weiß,[21] den Vorwurf einer „dem Denken des 19. Jahrhunderts verpflichteten" wissenschaftspositivistischen Perspektive nahe.[22] Denn obwohl Stöhr ‚normativ-subjektive' und ‚empirisch-objektive' Erkenntnisse unterscheiden will, erfordert schon jede Daten*erhebung* wertende Vorentscheidungen (dazu unten III.3), ebenso wie auch ihre Interpretation normativer Setzungen bedarf.[23] Bei einer Anreicherung der Auslegung durch empirische Methoden kann es also nicht darum gehen, normative Erwägungen zu ersetzen, sondern nur darum, sie besser zu untermauern. Der normative Bewertungsakt ist unverzichtbar, er lässt sich lediglich transparent machen und dadurch zur Diskussion stellen. Deshalb wurde die „empirische Herangehensweise" andernorts als „Evidenzbasierung" bezeichnet, weil sie „normative Bewertungen nicht ersetzen, sondern empirisch unterfüttern soll – eben ‚auf Evidenz basieren'."[24]

Denkt man Stöhrs Anregung konsequent weiter, wäre allerdings zu fragen, welche Berechtigung der erkenntnistheoretische Grenzwall zwischen Sein und Sollen überhaupt (noch) haben kann. Denn auch aus rechtlichen Werten allein lässt sich ohne eine bestimmte Vorstellung vom Sein keine Handlungsempfehlung ableiten:[25] Der syllogistische Schluss vom Sollen-im-Prinzip auf das Sollen-im-Konkreten erfordert einen Untersatz, der empirisch gehaltvoll sein *muss*. Denn er überführt einen rechtsethisch begründeten, damit apriorisch-empiriefreien, Obersatz (Sollen-im-Prinzip) in eine handlungsleitende, also pragmatisch-realweltliche, Schlussfolgerung (Sollen-im-Konkreten). Erkenntnisse über das Sein schlagen mithin die Brücke zwischen Sollen und Sol-

---

[20] Wortgleich *Stöhr* 2017 (Fn. 5), 145.

[21] *Augsberg* Rechtstheorie 46 (2015), 71, 77 f. m.w.N.; schon *Augsberg* Der Staat 2012, 117, 120 erläuterte, dass „allüberall" die „alten Vorstellungen von ‚wissenschaftlicher Objektivität' zusammengekehrt werden".

[22] So *Augsberg*, in: Buchner/Ladeur (Hrsg.): Wissensgenerierung und -verarbeitung im Gesundheits- und Sozialrecht, 2016, 73, 80 Fn. 34 schon gegen *Hamann* (Fn. 9), 1 – dort aber ohne Hinweis auf *Hamann* (Fn. 9), 79–81 („Schließende Statistik als angewandte Rhetorik"), und 107–109 (s. nächste Fn.).

[23] So *Hamann* (Fn. 9), 107–109 (Rezeptionsgrundsatz I.): „Alle empirische Forschung ist implizit normativ."

[24] *Hamann* (Fn. 9), 14 f.

[25] Dazu *Chang/Wang*, The Empirical Foundation of Normative Arguments in Legal Reasoning, SSRN 2016, 5 – online unter www.ssrn.com/abstract=2733781.

len: Der ohne Empirie versuchte Sprung vom Rechtsprinzip zur Handlungsempfehlung beinhaltet ebenso einen „logischen Fehlschluss" wie die Ableitung eines Sollens direkt aus dem Sein.[26] Dass dies von Juristen selten erkannt und fast nie thematisiert wird, schwingt in *Claus-Wilhelm Canaris'* emphatischer These mit, dass die gesamte Rechtswissenschaft „auf einem naturalistischen Fehlschluss" beruhe.[27] Nur ein von „jenseits" des juristischen Limes geholter Untersatz ermöglicht „diesseits" den Übergang von der Theorie zur Praxis des Rechts.

Es überrascht deshalb nicht, dass die Sein/Sollen-Dichotomie auch von (Rechts-)Philosophen in Frage gestellt wird. Eingeräumt wird zwar, dass es zum „Kern der juristischen Denkschulung" gehöre, zwischen „Beschreiben (deskriptiv) und Werten (normativ)" unterscheiden zu können.[28] Aber ebenso wenig wie das sachenrechtliche Trennungsprinzip zwingend ein Abstraktionsprinzip nach sich zieht, folgt aus der *Unterscheidung* zwischen Sein und Sollen bereits ihre erkenntnistheoretische *Verabsolutierung*:

> „Rechtswissenschaft ist zwar eine normative Wissenschaft. Aber nicht nur: sie muss auch die soziale Wirklichkeit erkennen. Sie ist damit Geisteswissenschaft und Sozialwissenschaft in einem, ein Verfahren der gegenseitigen Übersetzung zwischen Norm und Faktum. […] Rechtswissenschaft muss daher zugleich *Seinswissenschaft* und *Normwissenschaft* sein."[29]

Auch juristische Methodenlehrbücher schlagen deshalb inzwischen vor, die „Sachfaktoren" von Rechtsnormen „empirisch zu ermitteln" und eine eigene juristische Methode der „Normbereichsanalyse als wesentlichen Faktor juristischer Entscheidung" zu entwickeln.[30] Diesen Schritt geht Stöhr einst-

---

[26] *Chang/Wang* (Fn. 25): "without an empirical premise that performing a certain action will achieve a given goal or value, it is also a logical fallacy to jump from the goal or value to a normative claim that this action ought to be done".

[27] Zit. in *Martens* AcP 214 (2014), 93, 96: „Die Rechtswissenschaftler machten sich heute häufig nicht mehr klar, dass ihr ganzes Fach auf einem naturalistischen Fehlschluss beruht."

[28] *Mastronardi* (Fn. 16), Rn. 246.

[29] *Mastronardi* (Fn. 16), Rn. 287, 333; ähnl. zuvor Rn. 245; zur Gewichtung Rn. 54, 720: „Juristisches Denken wird seiner interdisziplinären Aufgabe nur gerecht, wenn es Norm und Realität, Logik und Empirie zugleich umfasst. Es kann auf keine der beteiligten Rationalitäten verzichten. Es darf aber auch keiner Denkweise generell eine Priorität vor anderen zuerkennen."; Rn. 328: „Die Trennung von Sein und Sollen ist nur unter folgendem doppelten Paradigma zwingend: a) der deduktiven Logik als alleiniger Denkform des Erkennens b) der cartesianischen Trennung von Subjekt und Objekt."

[30] *Müller/Christensen*, Juristische Methodik Bd. 1: Grundlegung für die Arbeitsmethoden der Rechtspraxis, Berlin, 11. Aufl., 2013, 527 f. (Rn. 482 f.); nun auch *Reimer*, Juristische Methodenlehre, 2016, 5 („Für die Rechtsanwendung heißt das: Sie ist mehr denn je Arbeit an Tatfragen. Vor die Interpretation der Normen (des Prüfungs*maß-*

weilen noch nicht, doch seine Initiative könnte das Vertrauen auf die Beständigkeit des juristischen Limes mittelfristig durchaus erschüttern. Um die epistemologischen Konsequenzen des von ihm Vorgetragenen abzuschätzen, bedarf es nun allerdings einer näheren Auseinandersetzung.

III. Empirie-Rezeption als genuin juristische Herangehensweise

Der wesentliche Beitrag Stöhrs liegt darin, eigene Befunde zu einer bislang in Deutschland kaum untersuchten Frage beizusteuern. Dadurch betätigt sich Stöhr als „Produzent" empirischer Forschung, tritt gewissermaßen aus den derzeitigen Grenzen des juristischen Binnendiskurses hinaus – und den an der Nutzung seiner Erkenntnisse interessierten „Rezipienten" gegenüber.[31] Diesen Schritt kennzeichnete er selbst als „ein gewisses Wagnis", dessen Beurteilung „anderen" zu überlassen sei,[32] womit wohl die juristischen Fachkollegen gemeint sein dürften. Schließlich ist es eine „genuin rechtswissenschaftliche Aufgabe",[33] festzustellen, „welche außerrechtlichen Fakten für eine bestimmte normative Fragestellung herangezogen werden sollen" und ob sich empirische Studien dafür „als fruchtbar erweisen."[34] Dementsprechend ist Stöhr beizupflichten, dass in der Rechtsliteratur eine „Rezeption von empirischen Erkenntnissen deutlich verstärkt werden" sollte.[35]

Sind wir nun als Rezipienten mit der von Stöhr produzierten Studie konfrontiert, so benötigen wir eine „systematisch ansetzende Methodik", um die erhobene Empirie „mit den Elementen der Normtextauslegung rational […] in Beziehung zu bringen".[36] Eine solche Methodik kann sich an das Vorgehen in sog. „evidenzbasierten" Disziplinen anlehnen,[37] deren „wissenschaftliche Anforderungen" laut Stöhr zwar auf „Juristen durchaus abschreckend" oder „frustrierend" wirken könnten,[38] aber aus Gründen der auch von Stöhr (559)

---

stabs) schiebt sich als erhebliche Herausforderung die Interpretation des Prüfungs*gegenstands*."), 52 ff. (Rn. 67 ff.), 177 ff. (Rn. 368 ff.).

[31] Begrifflichkeiten nach *Hamann* (Fn. 9), 25 ff.

[32] *Stöhr* 2017 (Fn. 5), 148, 149.

[33] *Langenbucher* ZGR 2012, 314, 315; ähnl. schon *Schön*, in: Engel/Schön (Hrsg.), Das Proprium der Rechtswissenschaft, 2007, 313, 318: „genuine Aufgabe der Jurisprudenz".

[34] *Schön* (Fn. 33).

[35] *Stöhr* 2017 (Fn. 5), 149.

[36] *Müller/Christensen* (Fn. 30), 527 f. (Rn. 482 f.).

[37] So ausf. *Hamann* (Fn. 9), 25 ff.

[38] *Stöhr* 2017 (Fn. 5), 148.

befürworteten Methodenehrlichkeit wohl unumgänglich sind.[39] Nur wenn empirische Studien systematisch kritikfähig gemacht und entsprechend hohe Maßstäbe an die Produzenten empirischer Befunde angelegt werden, lässt sich gewährleisten, dass die „empirische Herangehensweise" nicht unversehens zum Danaergeschenk wird und die von Stöhr gerade kritisierten „künstlichen" bzw. „subjektiven" Wertungen des Rechtsanwenders[40] (566) mit der erhöhten Glaubwürdigkeit empirischer Argumentformen in die Rechtsdogmatik rollt.

Zur praktischen Umsetzung einer solchermaßen disziplinierten Rezeption ergibt die „evidenzbasierte" Methodik im Wesentlichen vier Arbeitsschritte: Die Formulierung einer Fragestellung, das Auffinden früherer Forschung, die Würdigung dieser Forschung und die Verarbeitung der Feststellungen.[41] Entlang dieses Erkenntnisprogramms ist nun auch Stöhrs empirische Studie kritisch zu würdigen:

1. Formulierung der Fragestellung

Für die Formulierung einer empirisch untersuchbaren Fragestellung „muss der juristische Rezipient durch eine gründliche dogmatische Auseinandersetzung diejenigen empirischen Annahmen oder Bezüge herausarbeiten, die der Rechtsfrage […] zugrundeliegen."[42] Rechtsfrage in diesem Sinne ist bei Stöhr die Frage, ob eine bestimmte arbeitsvertragliche Klausel „klar und verständlich" ist im Sinne des § 307 Abs. 1 Satz 2 BGB. Da Stöhr zwei unterschiedliche Sachverhalte untersucht, lässt sich auch von zwei Fällen mit je eigener Rechtsfrage sprechen.[43]

Im ersten Fall zitiert Stöhr eine 2007 vor dem Bundesarbeitsgericht streitentscheidende Arbeitsvertragsklausel mit den Worten:

---

[39] Gegen die Ansicht *Arndts*, von Juristen könne „nicht empirische Arbeit derselben Art und Güte wie von Forschern anderer Disziplinen erwartet werden" (Empirie in den Rechtswissenschaften – Fluch oder Segen?, Diplomarbeit Augsburg 2008, urn:nbn:de:bvb:384-opus4-10243, 84), vgl. schon *Hamann* (Fn. 9), 30.

[40] Ohne die auch *Stöhr* nicht auskommt – etwa wenn er ohne empirischen Beleg annimmt, dass die Rechtsprechung die Kompetenz von Kfz-Händlern unter- und von Versicherungsnehmern überschätze (575), oder dass „eine überstrenge Transparenzkontrolle" Arbeitgeber „davon abhalte, Sonderzahlungen zu leisten" (577).

[41] *Hamann* (Fn. 9), 31 ff.

[42] *Hamann* (Fn. 9), 31.

[43] Die „Rechtsfrage" würde damit ähnlich abgegrenzt wie der zivilprozessuale „Streitgegenstand" nach der herrschenden zweigliedrigen Lehre: Dazu gehören Klageantrag *und* Klagegrund. Beide Begriffe ähnlich zu verwenden liegt auch deshalb nahe, weil Stöhr sich gerade um Urteilskritik bemüht.

„Darüber hinaus erhalten Sie einen gewinn- und leistungsabhängigen Bonus […]. Die Zahlung des Bonus erfolgt in jedem Falle freiwillig und begründet keinen Rechtsanspruch für die Zukunft." (565)

Stöhr meint, das Bundesarbeitsgericht habe einen Widerspruch der beiden Sätze festgestellt, der die Klausel intransparent mache (569), und will deshalb ihre Verständlichkeit empirisch überprüfen, indem er sie („der Einfachheit halber" abgewandelt und ohne Auslassungszeichen[44]) seinen Befragungsteilnehmern vorlegt.

Diese sog. „Operationalisierung" der juristischen Fragestellung verdeutlicht eine Herausforderung, der empirische Studien stets begegnen: Rechtsprobleme sind stets zu komplex, um sie in Gänze empirisch zu untersuchen, daher sind Auswahl- und Kürzungsentscheidungen unumgänglich.[45] Dieser Gedanke leitete Stöhr wohl dabei, den in seinem Zitat mit Auslassungszeichen gekennzeichneten Teil der Klausel nicht in den Befragungstext aufzunehmen.

Wiewohl daran nichts Grundsätzliches auszusetzen ist, ist der hier von Stöhr ausgelassene Text doppelt so lang wie der zitierte und umfasst praktisch alle vom Bundesarbeitsgericht als entscheidungserheblich erkannten Passagen. Im Originalfall nämlich lagen dem Gericht drei separate Absätze der Vergütungsregelung vor:

„[Abs. 3] Darüber hinaus erhalten Sie einen gewinn- und leistungsabhängigen Bonus, der im ersten Jahr Ihrer Betriebszugehörigkeit EUR 7.700,– nicht unterschreiten wird und im Frühjahr des Folgejahres zur Auszahlung kommt. Danach nehmen Sie an dem in unserem Hause üblichen Bonussystem teil.

[Abs. 4] Die Zahlung des Bonus erfolgt in jedem Falle freiwillig und begründet keinen Rechtsanspruch für die Zukunft.

[Abs. 5] Der Anspruch auf Zahlung eines Bonus entfällt, wenn Sie am 1.4. des Auszahlungsjahres nicht mehr in einem ungekündigten Arbeitsverhältnis mit unserem Hause stehen."[46]

Diese Klauseln hielt das BAG deshalb für intransparent, weil „die Regelung in Nr. 3 IV des Arbeitsvertrags […] im Widerspruch zu den in

---

[44] Die in *Stöhrs* Befragung verwendete Formulierung lautet: „Der Arbeitgeber zahlt ein Weihnachtsgeld. Hierbei handelt es sich um eine freiwillige Leistung, auf die kein Anspruch für die Zukunft besteht" (567).

[45] *Stöhr* meint insoweit (eher beiläufig), dass es „nur darum gehe, den konkreten Lebenssachverhalt möglichst genau abzubilden, bei der Transparenzkontrolle also die in Rede stehende Vertragsklausel. Die Formulierung dürfte somit keine nennenswerten Schwierigkeiten bereiten." (581); er lässt offen, warum Weihnachtsgeld ein „möglichst genaues" Abbild des „konkreten" Leistungsbonus sein soll, obwohl beide Zulagen schon aufgrund ihrer Bezeichnungen für unterschiedlich verbindlich gehalten werden dürften.

[46] BAG NZA 2008, 40.

Nr. 3 III und in Nr. 3 V des Arbeitsvertrags getroffenen Vereinbarungen" gestanden habe.[47] Dabei setzte sich das Gericht ausführlich mit drei Passagen in den beiden letztgenannten Absätzen auseinander,[48] von denen bei Stöhr zwei gänzlich fehlen und von der ersten gerade jene zwei Drittel ihres Textes – über Höhe und Fälligkeit der Bonuszahlungen – die das Gericht ausdrücklich zitiert hatte. Gerade die weggelassenen Passagen dürften maßgeblich zu der vom BAG wahrgenommenen Verbindlichkeit des dritten Absatzes beigetragen haben, weshalb ihre Auslassung den vom BAG erkannten Widerspruch denknotwendig aus der empirischen Untersuchung ausklammerte.

Damit kann der von Stöhr entworfene Befragungstext, der die entscheidungserheblichen Formulierungen auslässt, zwar interessante Einblicke in das Verständnis einer entsprechend vereinfachten (Weihnachtsgeld-)Klausel ermöglichen, verfehlt aber sein Ziel einer konkreten Methodenkritik gerade an der genannten BAG-Entscheidung. Ob Erkenntnisse über eine solcherart veränderte Klausel auf die in der Praxis anzutreffenden noch übertragbar sein könnten, bedürfte weiterer Reflexionen, die hier kaum nachzuholen sind.

Die durch die verkürzte Wiedergabe abgeschnittenen methodischen Anschlussfragen stellen sich dafür umso deutlicher im zweiten von Stöhr untersuchten Fall. Dort hatte das Bundesarbeitsgericht 2011 über folgende Klausel zu entscheiden (565):

> „Sonstige, in diesem Vertrag nicht vereinbarte Leistungen des Arbeitgebers an den Arbeitnehmer sind freiwillig und jederzeit widerruflich. Auch wenn der Arbeitgeber sie mehrmals und regelmäßig erbringen sollte, erwirbt der Arbeitnehmer dadurch keinen Rechtsanspruch für die Zukunft."

Stöhr legt nun dar, dass § 307 Abs. 1 Satz 2 BGB wegen der „Informationsfunktion" des Transparenzgebots (561) auf einen faktischen (und damit empirisch ermittelbaren) Verständnishorizont verweist, und fragt deshalb, ob „Arbeitnehmer den feinsinnigen Unterschied zwischen Freiwilligkeits- und Widerrufsvorbehalt kennen und daher deren Kombination als unverständlich und widersprüchlich identifizieren" könnten (566).

---

[47] BAG NZA 2008, 41 f. (Rn. 16), ebenso Rn. 18: „Die Regelung in Nr. 3 IV des Arbeitsvertrags ist jedoch deshalb nicht klar und verständlich i.S. von § 307 I 2 BGB, weil sie zu den in Nr. 3 III und in Nr. 3 V des Arbeitsvertrags getroffenen Vereinbarungen in Widerspruch steht."

[48] BAG NZA 2008, 41 (Rn. 19): Abs. 3 Satz 1 sei „typisch für die Begründung eines Entgeltanspruchs", Abs. 3 Satz 2 lasse sich „vom Wortlaut her nur dahingehend verstehen, dass dem Kl. eine Bonuszahlung zusteht, wenn die […] Voraussetzungen vorliegen" und Abs. 5 setze „das Entstehen eines Anspruchs auf die Bonuszahlung voraus. Nur ein entstandener Anspruch kann ‚entfallen'.".

2. Auffinden früherer Forschung

Ebenso wie dogmatische Arbeiten müssen auch empirische Studien den bisherigen Erkenntnisstand umfassend berichten, um Redundanzen zu minimieren, anschlussfähig zu werden und ihre Kontextualisierung zu ermöglichen. Für die von Stöhr aufgeworfene Frage nach der Erkenntnisfähigkeit von Arbeitnehmern hätte eine Anknüpfung an frühere empirische Forschung nahegelegen, wie sie sich aus Studienberichten in sozialwissenschaftlichen Datenbanken ergibt.[49] Für den deutschsprachigen Raum stellt Stöhr zwar fest, dass „im Rahmen des Transparenzgebots [.] eine empirische Herangehensweise – soweit ersichtlich – bislang noch nicht in Erwägung gezogen" wird (559). Dieses Forschungsdefizit scheint er jedoch nur für Deutschland attestieren zu wollen, da ihm durchaus die „stärkere empirische Fokussierung […] der US-amerikanischen Rechtswissenschaft" (559) vor Augen steht.[50]

Die bisherige US-amerikanische und internationale Forschungsliteratur ergibt unterschiedliche Anknüpfungspunkte: Zum einen wird die seit knapp vierzig Jahren betriebene[51] empirische Forschung an Formularverträgen (*standard-form contracts*, *boilerplate*, *form-adhesive contracts*) ausdrücklich auch mit ihrer Verwendung in Arbeitsverhältnissen begründet.[52] Diese Forschung belegt bisher nahezu durchweg, dass weder Verbraucher noch Arbeitnehmer ihre Vertragsbedingungen überhaupt verständig lesen.[53] Wer also aus der *ex-ante*-Perspektive Wert auf die Allgemeinverständlichkeit von Formularverträgen legt, kann sich nur entweder auf ein kontrafaktisches Autonomieverständnis berufen, das mehr Schaden als Nutzen stiften könnte,[54] oder

---

[49] Vgl. zu diesem zweiten Arbeitsschritt *Hamann* (Fn. 9), 31.

[50] Diese wird freilich mit Quellen von 1994 und 2002 belegt (559 Fn. 2), wobei die wichtigsten Entwicklungen auf diesem Gebiet gerade seit 2004 stattgefunden haben – etwa die Gründung einer eigenen Zeitschrift für Rechtsempirie (Journal of Empirical Legal Studies, JELS, 2004), einer jährlichen Konferenzreihe (CELS) der gleichnamigen Gelehrtengesellschaft (SELS) seit 2006, und ihrer Erweiterung nach Europa (CELSE Amsterdam) 2016 und Asien (CELSA Taipei) 2017.

[51] Frühes Beispiel zu kaufrechtlichen Gewährleistungsklauseln: *Wisdom* Stan. L. Rev. 31 (1979), 1117; ausf. Auseinandersetzung mit neueren Studien bspw. in den Beiträgen zum Book Symposium der Jerusalem Review of Legal Studies 12 (2015), 105–182.

[52] *Eigen* J. Inst. Theor. Econ. 168 (2012), 124, 125: "Employers require employees to sign such contracts restricting their rights.".

[53] *Johnston* Mich. L. Rev. 104 (2006), 857, 864: "most if not all individual consumers and employees will never read, let alone understand, all the terms in a firm's standard-form contract."; speziell für Schlichtungsklausel im Arbeitsvertrag *Eigen* Conn. L. Rev. 41 (2008), 381, 409 ff.

[54] So *Ben-Shahar* Eur. Rev. Contract L. 5 (2009), 2, 7: "Real people don't read standard form contracts. […] the opportunity to read […] might even hurt transactors" und *Wilkinson-Ryan* Cornell L. Rev. 103 (2017), Entwurfsfassung verfügbar unter www.

auf das (inzwischen ebenfalls zweifelhafte) ökonomische Argument, dass bereits eine AGB-lesende Minderheit (*informed minority*) Anreize zur verbraucherfreundlichen AGB-Gestaltung setzen kann.[55] Erst diese Aporie lenkt den Blick auf die *ex-post*-Perspektive, in der die (Un-)Verständlichkeit einer Vertragsklausel auch die effektive Rechtsdurchsetzung im späteren Streitfall beeinflusst, indem „die Gefahr besteht, dass der Vertragspartner seine Rechte nicht wahrnehmen […] kann" (561) und von der „Durchsetzung bestehender Rechte abgehalten" werden könnten (560).[56]

Jedenfalls befasst sich eine zweite Linie der empirischen Forschung mit der sprachlichen Vereinfachung von Vertragsformularen – etwa im Bank-,[57] Verbraucherkredit-,[58] Darlehenshypotheken-, Grundstückskauf- und Mietvertragsrecht[59] – und mit der statistischen Untersuchung der Lesbarkeit von Rechtstexten,[60] teils auch mit experimentellen Methoden.[61] Schließlich lassen sich auch Untersuchungen zum Einfluss der allgemeinen Lesekompetenz auf die Vertragsgestaltung einbeziehen.[62]

Eine nähere Sichtung dieser Literatur wäre sicher gewinnbringend gewesen, denn bei aller Kulturkontingenz empirischer Forschung im Allgemeinen und der Sprachabhängigkeit von Leseverhalten und Vertragsgestaltung im Besonderen erlaubt der Anschluss an bereits bestehende Literatur eine bessere Begründung der eigenen Methodenwahl und erleichtert die Interpretation und Einordnung der eigenen Ergebnisse.

---

ssrn.com/abstract=2738567; *dies.*, The Behavioral Paradox of Boilerplate, USC Research Paper CLASS17–12, verfügbar unter www.ssrn.com/abstract=2938011: "This Article makes the empirical case that unread […] fine print inhibits reasonable challenges to unfair deals.".

[55] Skeptisch insoweit allerdings *Becher/Unger-Aviram* DePaul Bus. & Comm. L. J. 8 (2010), 199, deren Umfrageergebnisse "do not support the assumption found in some literature that a substantial minority of consumers read their contracts and thus might discipline sellers."; ebenso *Bakos/Marotta-Wurgler/Trossen* J. L. Stud. 43 (2014), 1, 5: "we show that the informed-minority hypothesis, the most widely applied argument for the efficiency of standard-form contract terms, does not seem compelling".

[56] Ähnlich *Ben-Shahar/Strahilevitz* (Fn. 6), 44 f.

[57] *Campbell* J. Bus. Comm. 36 (1999), 335, 341 ff.

[58] *Davis* Va. L. Rev. 63 (1977), 841, 856 ff.

[59] *Masson/Waldron* Appl. Cogn. Psy. 8 (1994), 67, 71 ff.

[60] *Fry* J. Reading 30 (1987), 338; *Friman* Loy. Cons. L. Rep. 7 (1994), 103; *Marotta-Wurgler/Taylor* N.Y.U. Law Rev. 88 (2013), 240.

[61] *Van Boom/Desmet/Van Dam* J. Cons. Pol. 39 (2016), 187.

[62] *White/Mansfield* Stan. L. & Pol'y Rev. 13 (2002), 233.

3. Würdigung der Erkenntnisse

Im dritten Arbeitsschritt der empirischen Rezeption „muss der Rezipient Möglichkeiten und Grenzen der herangezogenen Studien reflektieren".[63] Da hier nur Stöhrs eigene Studie zu würdigen ist, wird zunächst seine Methodenwahl als solche betrachtet (a), sodann Fragen der Stichprobenauswahl (b) und der Messmethodik (c).

a) Die Methode seiner Wahl beschreibt Stöhr als „Laborexperiment mittels experimenteller Befragung" (567),[64] wenngleich seine Untersuchung weder in einem Labor stattfand noch das von ihm selbst genannte Definitionsmerkmal eines Experiments erfüllt, dass es „Störfaktoren ausschaltet bzw. kontrolliert" (566 f.). Nur eine solche experimentelle Kontrolle ermöglicht nämlich den Vergleich verschiedener Einflussfaktoren unter ansonsten gleichen Rahmenbedingungen (*ceteris paribus*) und lässt überhaupt empirische Kausalschlüsse zu. Stöhr dagegen setzt den Begriff „Experiment" wohl eher mit jeglicher Form von „Datenerhebung" gleich, da er beide Worte als Oberbegriffe für die „Befragung" und die davon zu unterscheidende „Beobachtung" verwendet (567, 580). Deshalb will er sogar einen E-Mail-Verteiler als Laborumgebung verstehen, obwohl er sieht, dass dort „nicht wie in einem klassischen Experiment" die Fragebögen „unter kontrollierten Bedingungen wie in einem Labor ausgefüllt werden" (570).

Um diese begrifflichen Unschärfen zu vermeiden, sollte wohl auf die bisherige sozialwissenschaftliche Nomenklatur zurückgegriffen werden, die Befragungen als eine *Variante*, Experimente dagegen als *Antipoden* der Beobachtungsstudie verstehen.[65] Danach wäre Stöhrs Studie angesichts der fehlenden Umgebungskontrolle und der unterbliebenen Neutralisierung von Störfaktoren wohl kaum als Experiment, sondern vielmehr als *beobachtende Feldstudie durch Befragung* zu kategorisieren. Diese Nomenklatur erlaubt eine präzisere Würdigung der Erkenntnisse, denn sie weist augenblicklich auf konzeptionelle Stärken und Schwächen hin – wie etwa die fundamentale Unzulässigkeit von Ursachenschlüssen. Eine alternative Erhebungsmethode, die eher als „experimentelle" Befragung zu bezeichnen wäre, weil sie Störvariablen neutralisiert, wird noch darzustellen sein (s.u. IV.4). Auch Stöhr erkennt zwar die fehlende Kontrolle von Störvariablen als Problem (570),[66] meint aber ange-

---

[63] *Hamann* (Fn. 9), 31.

[64] Ähnlich *Stöhr* 2017 (Fn. 5), 145 Fn. 10: „experimentelle Befragung" (trotz Kenntnis des in der folgenden Fußnote nachgewiesenen abweichenden Sprachgebrauchs).

[65] Ausf. *Hamann* (Fn. 9), 137 ff.; zu Befragungen ebd. 176 ff.

[66] Und ergänzt: „Zudem wurden keine Reihenfolgen-Effekte kontrolliert." – anders nun *Ben-Shahar/Strahilevitz* (Fn. 6), 23 ("The order in which respondents saw these vignettes was randomized.")

sichts seiner „gravierenden" Ergebnisse, diese Methodenfragen hintanstellen zu dürfen. Dagegen spricht jedoch, dass sich gravierende Verständnisunterschiede *gerade* aus unkontrollierten Störvariablen erklären können,[67] so dass die Eignung einer beobachtenden Befragungsstudie für den von Stöhr verfolgten Zweck von vornherein zweifelhaft erscheinen musste, ja womöglich „gar nichts beweist".[68]

b) Jedenfalls stellen sich auch Fragen nach der Stichprobenziehung. Hier ist eine transparente und detaillierte Erläuterung besonders wichtig, um die Replizierbarkeit der Studie zu gewährleisten, also anderen Forschern jene Wiederholung der Studie unter gleichen Bedingungen zu ermöglichen, die schon konzeptionell unabdingbares Begriffselement jeder statistischen Auswertung ist.[69] Dazu berichtet Stöhr, er habe „sämtliche Studierende und Mitarbeiter der Philipps-Universität Marburg […] mit den entsprechenden Email-Verteilern" eingeladen, „wodurch insgesamt 29.404 Personen erreicht wurden" (567). Nicht berichtet war, welche E-Mail-Verteiler verwendet wurden, wie sie bestückt werden und wie (bzw. wie lange) den Befragten die Teilnahme ermöglicht wurde. Deshalb lässt sich insbesondere nicht abschätzen, wie viele der 29.404 „erreichten" E-Mail-Adressen tatsächlich noch aktiv sind und Empfängern gehören, die ihre E-Mails im Studienzeitraum abgerufen haben, den Befragungstext einwandfrei angezeigt bekamen, des Deutschen mächtig sind und auch alle sonst unterstellten Voraussetzungen für eine Teilnahme erfüllten. Künftigen Studien sollte deshalb wenigstens ein methodischer Anhang (und sei es nur im Internet als sog. *online supplement*) beigegeben werden, der die verwendeten Prozeduren und Materialien eingehend und nachvollziehbar schildert.

Der von Stöhr berichtete Rücklauf von 797 bzw. 799 Antworten (568 f.) ergibt jedenfalls eine Rücklaufquote von 2,7 %, also umgekehrt einen Ausfall von 97,3 %. Stöhr ist mit dem Begriff der „Nonresponse-Rate" zwar vertraut (580), reflektiert sie aber nicht in Bezug auf seine eigene Studie, was durchaus aufschlussreich gewesen wäre. Denn allgemein sollte der Rücklauf „bei einer wissenschaftlichen Studie stets dokumentiert, mehr oder minder umfassend ausgewertet und in Publikationen auch berichtet werden in Form von

---

[67] Fragte man etwa in einer Umfrage nach rechtswidrigem Vorverhalten der Befragungsteilnehmer, könnte kein noch so deutliches „Nein"-Ergebnis die Rechtstreue aller Menschen belegen, sondern lediglich die Unerwünschtheit von „Ja"-Bekenntnissen zur Rechtswidrigkeit.

[68] So jüngst ausdr. *Ben-Shahar/Strahilevitz* (Fn. 6), 22: "In designing the interpretation surveys, we needed to do more than throw facts of cases in front of respondents and ask them to vote. This would have proven nothing".

[69] Dazu *Hamann* (Fn. 9), 95.

Rücklaufkurve, Rücklaufquote und Rücklaufstatistik."[70] In Unkenntnis der Rohdaten lässt sich dies nicht nachholen, sondern nur feststellen, dass bei Befragungen eine Rücklaufquote im einstelligen Prozentbereich zwar nicht völlig unüblich ist,[71] dass aber „oft" Rücklauf bis zu 40 % beobachtet[72] und der hälftige Rücklauf sogar zum „Maßstab" (*benchmark*) der heutigen Umfrageforschung erkoren wird.[73] Selbst die „sehr geringe" Rücklaufquote von etwa 5 % in postalischen Umfragen[74] liegt noch fast doppelt so hoch wie die von Stöhr berichtete, während Rücklauf von unter 3 % in Fachzeitschriften praktisch nie vorkommt.[75] Ein so geringer Rücklauf überrascht umso mehr angesichts der von Stöhr stark vereinfachten und kaum zeitaufwändigen Fragestellung sowie des minimalen Rücksendeaufwands. Da erheblicher Ausfall in Befragungsstudien stets einen sog. Selektionseffekt hat – also den Kreis besonders engagierter Umfrageteilnehmer übermäßig hoch gewichtet –, lassen sich Stöhrs Ergebnisse vielleicht schon dadurch erklären, dass die besonders engagierten Teilnehmer der Umfrage auch mehr Erfahrung mit Texten im Allgemeinen und Verträgen im Besonderen haben und deshalb eine konsistentere Interpretation entwickeln können. Umgekehrt könnten aber auch diejenigen E-Mail-Empfänger, die sich über die Bedeutung der Vertragsklauseln unsicher waren (sie also für unklar gehalten haben würden), gerade deshalb von der Teilnahme zurückgeschreckt und nicht in die Analyse eingegangen sein.

Zudem erwähnt Stöhr selbst das Problem, „dass die Probanden überwiegend aus dem akademischen Umfeld kommen und die Befragung daher nicht für alle Arbeitnehmer repräsentativ ist" (570). Bedauerlicherweise lässt sich der Auswertung nicht einmal entnehmen, welche Antworten von den 6.551

---

[70] *Döring/Bortz*, Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften, 5. Aufl., 2016, 412.

[71] *Hamann* (Fn. 9), 179 mit anekdotischen Nachweisen, allerdings auch dem Hinweis (Fn. 256), dass im Zeitraum 1992–2003 über 231 Primärstudien hinweg ein durchschnittlicher Rücklauf von 34 % festgestellt worden war – und das schon unter Führungskräften in Unternehmen.

[72] So *Döring/Bortz* (Fn. 70); unter 638 Befragungsstudien ermittelten *Baruch* Hum. Rel. 52 (1999), 421, 430 und *Baruch/Holtom* Hum. Rel. 61 (2008), 1139, 1149 sogar einen Rücklauf von durchschnittlich gut 48 % für die Jahre 1995, 2000 und 2005.

[73] *Whelan*, Response Rates in 21st Century Organizational Survey Research, Diss. Raleigh 2015, 34: "convergent results [51–57 %] accomplished the noteworthy aim of providing a benchmark as to the average response rate for survey data collection efforts in 21st century organizational research."; ebenso Dawn M. Chutkow (Schriftleiterin des Journal of Empirical Legal Studies), CELSA-Workshop in Taipei, 16.6.2017: "If your response rate is below 50 % or even 60 %, you will have to explain why that is not a problem. [...] We begin to get nervous when we see response rates below 50 %".

[74] *Döring/Bortz* (Fn. 70), 414.

[75] *Baruch/Holtom* (Fn. 72) finden unter 463 Studien der Jahre 2000 und 2005 keine einzige mit weniger als 3,0 % Rücklauf.

befragten Beschäftigten der Universität stammten, die ja als Arbeitnehmer zwangsläufig arbeitsvertraglichen Regelungen ausgesetzt sind, und welche von den 22.853 befragten Studierenden, die vielleicht noch nie mit Arbeitsverträgen zu tun hatten. Eine differenziertere Auswertung hätte hier wertvolle Rückschlüsse ermöglicht. Ungeachtet dessen ist Stöhr zuzustimmen, dass keiner seiner Befragten Erfahrung mit denjenigen Formularverträgen haben dürfte, um die es in den BAG-Entscheidungen ging.[76] Einerseits könnten sie deshalb mehr Unklarheiten wahrgenommen haben als erfahrene privatwirtschaftliche Arbeitnehmer in derselben Situation, andererseits aufgrund ihres höheren Bildungsniveaus vielleicht Unklarheiten aufgelöst haben, die bei nicht-akademischen Arbeitnehmern größere Verwirrung stiften. Welcher der beiden gegenläufigen Einflüsse überwiegt, lässt sich mangels Störvariablenkontrolle nicht mehr feststellen. Auch andere Anforderungen an die Stichprobenziehung, die speziell für Online-Befragungen zu beachten sind,[77] finden im Studienbericht keinen Niederschlag. Dabei zeigt die neuere Literatur, dass dank des Internets inzwischen sogar repräsentative Stichproben aus dem Kreis der potentiellen Vertragsadressaten befragt werden können, was viele der eben skizzierten Schwierigkeiten vermieden hätte.[78]

Diese Schwierigkeiten wirken sich direkt auf die Interpretation der von Stöhr beobachteten hohen Übereinstimmung seiner Befragungsteilnehmer aus: Zur zweiten Rechtsfrage fand die am häufigsten gewählte Lesart der Vertragsklausel (sog. Modalwert[79]) eine 92-prozentige Zustimmung, die beiden anderen Antwortmöglichkeiten – darunter auch diejenige, dass keine klare Antwort auszumachen sei – je eine etwa 4-prozentige. Berücksichtigt man allerdings, dass an der Studie von vornherein weniger als 3 % der Befragten überhaupt teilgenommen haben, ergibt die von Stöhr festgestellte 92-prozentige Übereinstimmung seiner *Teilnehmer* eben nur eine 2,5-prozentige Übereinstimmung unter allen *Befragten*, während 97,3 % der Befragten gar keine Stellung bezogen.[80]

---

[76] Das Bundesarbeitsgericht befasste sich in NZA 2008, 40 mit dem Bonussystem eines Finanzdienstleisters, in NZA 2012, 81 mit dem Arbeitsvertrag eines Sozialpädagogen in einem privaten Verein.

[77] *Döring/Bortz* (Fn. 70), 414.

[78] Dazu noch unten IV.3.

[79] Missverständlich *Stöhr* (568, 569): „Median". Dieses Maß setzt in empirischen Disziplinen jedoch mindestens ordinalskalierte (also fortlaufend aufreihbare) Daten voraus, während für nominalskalierte (kategoriale) Variablen nur der „Modus" als Lageparameter der zentralen Tendenz definiert ist: *Hamann* (Fn. 9), 75.

[80] Selbst wenn also – was nach dem bereits Gesagten (oben nach Fn. 68) durchaus zweifelhaft ist – Stöhrs Stichprobe völlig verzerrungsfrei die Mehrheitsverhältnisse unter seinen Befragten abbildete, so fänden sich allein an der Universität Marburg knapp

c) Nur ergänzend bleibt darauf hinzuweisen, dass Stöhrs Messmethodik auch kein Maß für die individuelle Entscheidungssicherheit (*confidence*) vorsieht. Damit werden Teilnehmer, die sich einer der beiden Lesarten „völlig sicher" sind, mit solchen gleichgesetzt, die dieselbe Lesart „gerade noch so" vorziehen.[81] Nähme man als *advocatus diaboli* an, dass sich erstere eher unter den Befürwortern eines Zahlungsanspruchs finden, letztere eher unter den Ablehnern, wäre schon begründungsbedürftig, warum die in Richtung einer Momenteingebung ratende Mehrheit den Ausschlag geben sollte. Jedenfalls dürfte die von Stöhr angebotene Kompromisskategorie „nicht klar zu beantworten" den Anteil der Teilnehmer unterschätzen, die von keiner Lesart mit gefühlter Gewissheit überzeugt waren. Zudem zeigt eine neuere Studie, dass bisweilen ein Fünftel der Befragungsteilnehmer die gestellte Frage blind, willkürlich oder nach sachfremden Erwägungen beantwortet,[82] weshalb die genauen Prozentverhältnisse ohne systematische Korrektur[83] nicht für bare Münze genommen werden können. Daher lässt Stöhrs Studie kaum einen Rückschluss darauf zu, wie viele seiner Teilnehmer von ihrer Antwort wirklich hinreichend überzeugt waren, um im Ernstfall Zeit und Geld für rechtliche Schritte oder das Risiko einer anderweitigen Konfrontation aufzuwenden. Das stellt den von Stöhr ausgemachten Beitrag des Transparenzgebots zur erleichterten Rechtsdurchsetzung *ex post* (s.o. vor Fn. 56) wiederum in Frage. Abhilfe hätte sich schaffen lassen, indem die Abfragen auf diskreten Antwortskalen (sog. *Likert items*) erfolgt wären,[84] womöglich sogar mit konkreten Antwortanreizen, um ein reflektiertes und überzeugungsgemäßes Abstimmungsverhalten sicherzustellen.

Zuletzt ließe sich auch fragen, welche Aussagekraft die isolierte Betrachtung einer einzelnen Vertragspassage in Kenntnis des konkreten Sachproblems hat, wenn Vertragsschluss und -durchsetzung im Arbeitsleben von Ri-

---

2.400 Befragte (= 8,02 % von 29.404), die die gegenständliche Klausel für unklar oder anspruchsbegründend halten – was Stöhr als „geradezu unhaltbar" ansieht (570).

[81] Vgl. jetzt auch *Ben-Shahar/Strahilevitz* (Fn. 6), 19: "Should courts count differently respondents who say that the language 'definitely' means X versus those who say that are less decisive?"

[82] *Ben-Shahar/Strahilevitz* (Fn. 6), 33: „respondents appear to be answering at random, misunderstand the language and survey questions, or focus entirely on the equities of the vignettes, suggesting that even in clear-cut cases a sizeable minority of respondents will go the other way."

[83] So der Vorschlag von *Ben-Shahar/Strahilevitz* (Fn. 6), 33 ("'normalize' the results of interpretation surveys by establishing benchmarks for preponderance […] for example, an average 'uncertain' response rate of 20 %, this should be the new 'zero' and only rates exceeding this baseline should count as true votes for ambiguity.").

[84] So nun auch *Ben-Shahar/Strahilevitz* (Fn. 6), 23.

sikobeurteilungen *ex ante* abhängen und Verbraucher zahlreiche Klauseln parallel und in ihrer oft komplexen Wechselwirkung bedenken müssen.[85]

4. Verarbeiten der Feststellungen

Im letzten Arbeitsschritt der empirischen Rezeption „muss der Rezipient seine Feststellungen zum Stand der empirischen Erkenntnis auf die Rechtsfrage zurückbeziehen. Das ist die ‚klassische' normative Arbeit, die die Rechtsfolgen zu einem (jetzt besser) bekannten Tatbestand erörtert."[86]

Angesichts der dargestellten methodischen Unklarheiten ist eher zweifelhaft, ob Stöhr darin zu folgen ist, dass die vermeintlich „gravierende Abweichung" seiner Befragungsergebnisse von der Arbeitsrechtsprechung hinreichenden Anlass bietet, „die Paradigmen der Transparenzkontrolle grundlegend zu überdenken" (570 f.). Wenn 2,5 % der Befragten eine andere Ansicht vertreten als das Bundesarbeitsgericht (bzw. 1,9 % in Fall 1, wo schon die Fragestellung von der gerichtlich beurteilten abwich), so mag das durchaus Grund zur Beunruhigung sein, nötigt aber vielleicht noch nicht dazu, die „Unwirksamkeit des Freiwilligkeitsvorbehalts" als „reines Zufallsgeschenk" aufzufassen, „das einen unverhältnismäßigen Eingriff in die Vertragsfreiheit des Arbeitgebers darstellt" (577).

IV. Praktische Instrumente einer „empirischen Herangehensweise"

Einen „Paradigmenwechsel" mahnt Stöhr nicht nur im Hinblick auf die arbeitsrechtliche Transparenzkontrolle an, sondern auch im Hinblick auf die allgemeine Methodik seiner „empirischen Herangehensweise" (571), für die er im Fortgang seines Beitrags vier Thesen entwickelt und abschließend feststellt (582 f.). Diese „empirische Herangehensweise" mag den Gerichten nicht ganz so neu sein wie Stöhr annimmt,[87] begegnet aber doch großen Herausforderungen, die andernorts schon für die Rechtswissenschaft allgemein[88] und von

---

[85] Vgl. *Korobkin* U. Chi. Law Rev. 70 (2003), 1203.

[86] *Hamann* (Fn. 9), 32.

[87] *Stöhr* zitiert den BGH zum Beleg der Tatsache, dass „bei der Auslegung von Willenserklärungen [.] der objektive Empfängerhorizont normativ zu ermitteln" sei (559), dabei hat der BGH ebenso ausdrücklich festgestellt, dass der für die Auslegung nach §§ 133, 157 BGB maßgebliche Begriff der „Verkehrssitte" gerade „keine Rechtsnorm, sondern die den Verkehr beherrschende tatsächliche Übung" bezeichne: BGH, LM § 157 BGB Nr. 1 (online unter t1p.de/w9lk) m.Verw. auf RGZ 49, 157, 162 u.a.

[88] *Hamann* (Fn. 9), 27–30 m.w.N. zu „fünf strukturellen Hindernissen".

Stöhr nun speziell für die Praxis formuliert wurden: „Die praktische Umsetzung der empirischen Methode könnte viele Gerichte vor Probleme stellen, da es […] an Vorgaben und Leitlinien mangelt" (579). Dieser Befund hat in der Vergangenheit bereits die oben erwähnte „pragmatische Rezeptionslehre" für Rechtsempirie inspiriert, die in Anlehnung an die „Leitlinien" der evidenzbasierten Medizin auch für das Recht „Grundsätze der empirischen Rezeption" liefern sollte.[89] Stöhr diskutiert zwei konkrete Umsetzungsvorschläge, denen vielleicht noch zwei weitere hinzuzufügen wären.

1. Sprachwissenschaft (Verständlichkeitstests)

Der erste Vorschlag, den Stöhr erörtert, ist die Ausrichtung an der Sprachwissenschaft. Dabei betrachtet er zwei Techniken, die 1948 bzw. 1974 entwickelt wurden, um die Verständlichkeit allgemeinsprachlicher Texte messbar zu machen: Den Flesch-Test (573) und den Hamburger Verständlichkeitstest (572). Beide Tests sollen bestimmte Eigenschaften von Texten quantifizieren, um deren Lesbarkeit vergleichbar zu machen. Sie gelten für juristische Fachtexte schon seit längerem nicht mehr als *lege artis*,[90] und auch Stöhr lehnt sie letztlich aus Gründen ab, die er mit den Adjektiven „subjektiv" und „schematisch" umschreibt (573) – wobei er auch offen lässt, wie Lesbarkeitstests überhaupt die in seinen Beispielfällen relevanten Widersprüche zwischen verschiedenen Passagen identifizieren und klären könnten.

Interessant wäre sicher noch gewesen, wie Stöhr sich zu den grundsätzlichen Fragen stellt, die die bisherige Anwendung sprachwissenschaftlicher Methoden im Recht (die sog. Rechtslinguistik) ergeben hat. So befassen sich sowohl Rechtstheorie[91] als auch juristische Methodik[92] eingehend mit den Erkenntnissen der Sprachwissenschaften. Dabei hat sich wiederholt gezeigt, dass „Objektivität", wie Stöhr sie für seine empirische Herangehensweise reklamiert, in der sprachgebundenen Rechtsarbeit nicht einlösbar ist und dass es nur darum gehen kann, verschiedene Subjektivitäten im Sinne einer Diskursrationalität zu vermitteln. Die Konsequenzen dieses Verständnisses in der von Stöhr kritisierten Arbeitsrechtsprechung wären sicher spannend ge-

---

[89] *Hamann* (Fn. 9), Zitate auf 53, 7 Fn. 52, 106.

[90] Vgl. *Neumann*, in: Grewendorf/Rathert (Hrsg.), Formal Linguistics and Law 2009, 55, die die von Stöhr genannten Techniken erörterte (58, 67) und resümierte: „readability is not considered a state of the art methodology anymore" (58 m.Verw. auf *Lerch*, in: ders. [Hrsg.], Die Sprache des Rechts Bd. 1: Recht verstehen, 2004, 239–283).

[91] Zuletzt *Kuntz* AcP 215 (2015), 387.

[92] *Müller/Christensen* (Fn. 30), insb. 190–201 (Rn. 166–184), 218–231 (Rn. 202–218).

wesen; Klärung könnte womöglich das rechtslinguistische Schrifttum zum Arbeitsrecht bringen,[93] das hier indessen nicht zu rezensieren ist.

Dort wird auch eine Methodik eingeführt, die es erlaubt, den „Sprachgebrauch" einer Sprechergemeinschaft in replizierbarer Weise zu erheben: Die sog. „Korpuslinguistik" macht sich große Textsammlungen zunutze, um mit Hilfe von Computern die bisherigen Gebrauchskontexte von Wörtern oder Wendungen (sog. Kollokationen bzw. *n-grams*) zu ermitteln; dies lässt sich nutzen, um die Bedeutung und den Bedeutungswandel von Rechtsbegriffen zu erkennen und nachzuzeichnen.[94] Da die Methodik allerdings ebenfalls eher granular ansetzt (auf der Ebene von Begriffseinheiten), kann sie die von Stöhr thematisierten Widersprüche zwischen verschiedenen Passagen eines Textes nicht ohne Weiteres klären. So bleibt letztlich doch nur die Befragung als taugliches Instrument übrig.

2. Umfrageforschung (Demoskopie)

Befragungen gibt es in verschiedenen Varianten. Stöhr lenkt den Blick zunächst auf die systematische, wissenschaftlich disziplinierte und an großen repräsentativen Stichproben durchgeführte Befragung (Demoskopie), indem er darlegt, dass solche Befragungen im Recht „nur selten durchgeführt" werden, „etwa zur Bestimmung der […] Verkehrsauffassung im Rahmen des wettbewerbsrechtlichen Irreführungsverbots oder der Verkehrsdurchsetzung einer Marke i.S.v. § 8 Abs. 3 MarkenG" (579). Dort sind sie zwar etabliert genug, dass alle großen Meinungsforschungsinstitute ein eigenes Produkt „empirische Rechtsforschung" anbieten.[95] Für die Arbeits- und AGB-Rechtsprechung hingegen unterstellt Stöhr zu Recht, dass sie von Demoskopie bislang kaum Gebrauch machen. Deshalb ist es hilfreich, dass Stöhr unter der recht allgemein gehaltenen Abschnittsüberschrift „Zur empirischen Methodik" zunächst einige Prinzipien der Rechtsdemoskopie zusammenstellt (580 f.).

In der neueren US-amerikanischen Literatur dient die Befragung demoskopischer Stichproben sogar just als Mittel der Wahl für die empirische Vertragsauslegung. Dort interpretiert man das Vertragsrecht *funktional* äquivalent zum Marken- und Wettbewerbsrecht: Verträge, Marken und Werbung seien sämtlich nur unterschiedliche Wege zur Kommunikation ähnlicher wirtschaftlicher Versprechen[96] – wofür mittlerweile auch das (europäisch

---

[93] Insb. *Vogel/Pötters/Christensen* (Fn. 15).

[94] *Vogel/Pötters/Christensen* (Fn. 15), 72 ff.

[95] *Hamann* (Fn. 9), 34 m.w.N.

[96] *Ben-Shahar/Strahilevitz* (Fn. 6), 11: "Contractual terms are one way in which parties communicate their promises. Another channel of communication is precon-

überformte) Kaufvertragsrecht in § 434 Abs. 1 Satz 3 BGB eine Stütze bietet.[97] Diese Parallele zwischen Vertrags- und Marken- bzw. Wettbewerbsrecht legt dann auch einen Methodengleichlauf nahe: Auch für die Vertragsauslegung kommt demnach eine Befragung repräsentativer Stichproben durch professionelle Marktforschungsinstitute in Betracht.[98]

Dabei bleibt bisher freilich die Kostenfrage etwas im Vagen: Selbst bei einer „praktischen und kostengünstigen" Umsetzung im Internet dürften auch heute noch Kosten im vierstelligen Bereich anfallen.[99] Günstiger wären womöglich wissenschaftlich genutzte Befragungsstichproben wie das *German Internet Panel* (*GIP*) der Universität Mannheim, oder mittelfristig der Aufbau eigener Infrastrukturen für die Rechtsempirie. Diese Möglichkeiten werden künftig noch näher auszuloten sein.

3. Informelle Befragung (Indizienverfahren)

Solche Überlegungen brauchte Stöhr nicht anzustellen, weil er die Demoskopie von vornherein meidet (sie auch nicht beim Namen nennt), da der „erhöhte Arbeitsaufwand […] im Rahmen der richterlichen Rechtsanwendung […] kaum geleistet werden" könne (581). Stattdessen schlägt er als Herzstück seiner „empirischen Herangehensweise" ein „abgestuftes Verfahren durch Indizien" (581) vor, wonach dem Richter in drei Schritten bzw. Stufen zu empfehlen sei,

> „im ersten Schritt das Gespräch mit mindestens einer weiteren Person zu suchen. Ergibt sich danach kein einheitliches Bild, ist auf der nächsten Stufe die Einschätzung eines größeren Personenkreises einzuholen […] z.B. Kollegen, Freunde und Verwandte via Email […]. Führt auch dies nicht zu verlässlichen Erkenntnissen, muss der Personenkreis auf einer dritten Stufe noch weiter ausgedehnt werden. Dazu sollten andere Personen – z.B. Referendare – um Unterstützung dergestalt gebeten werden, dass sie ihrerseits Befragungen in ihrem Bekanntenkreis durchführen." (582)

Das Kernanliegen dieses dreistufigen Befragungsverfahrens verdient Beifall: Laut Stöhr dient es der Prozessökonomie, indem es „auf einfachem und

---

tractual 'marketing' – the representations that a party makes to its potential contractual counterparts through advertising, branding, and various statements and disclosures".

[97] „Zu der Beschaffenheit [der Kaufsache] gehören auch Eigenschaften, die der Käufer nach den öffentlichen Äußerungen des Verkäufers, des Herstellers […] oder seines Gehilfen insbesondere in der Werbung […] erwarten kann".

[98] *Ben-Shahar/Strahilevitz* (Fn. 6), 23 ("The experiment was administered online to a nationally representative sample recruited by Toluna, a well-regarded survey research firm.").

[99] *Ben-Shahar/Strahilevitz* (Fn. 6), 22 nennen ihr Experiment "practical and inexpensive" im Vergleich zur früheren Demoskopie, berichten aber Kosten von 3,25 USD pro Teilnehmer (39 Fn. 132) bei etwa 1.300 Teilnehmern (23 f.).

schnellem Wege aufschlussreiche Ergebnisse" gewinnen helfe und dabei „auch und gerade juristische Laien" einbeziehe (582).

Ob (und wie) dieser Vorschlag allerdings über die derzeitigen Prozessrechtsvorschriften hinausgeht, wonach in Arbeitsrechtsstreitigkeiten ohnehin stets Berufs- und Laienrichter zusammenwirken,[100] bleibt offen. Womöglich hatte Stöhr eher den Zivilrichter im Amts- und Landgericht vor Augen, der nach § 22 Abs. 1 GVG bzw. § 348 Abs. 1 Satz 1 ZPO grundsätzlich als Einzelrichter entscheidet; hier wäre trotz der Mittelkürzungen der vergangenen Jahrzehnte, die an einigen Stellen überhaupt erst zur originären Einzelrichterschaft geführt haben, eine diskursive Überzeugungsbildung sicher hilfreich.

Gleichwohl verbleiben Zweifel, wie „aufschlussreich" die gewonnenen Erkenntnisse wirklich wären: Stöhr selbst weist darauf hin, dass „Ergebnisse, die auf solchen Indizien beruhen, […] nicht uneingeschränkt belastbar" seien und nur „Erfahrungswerte" darstellten, „die der Richter durch eigenen Sachverstand bzw. durch jedes beliebige Mittel eruieren kann" (582). Damit fragt sich indes, ob diese Herangehensweise noch „empirisch" genannt werden sollte, oder ob dies den Begriffsinhalt nicht unglücklich aufweicht und das Definitionskriterium der Replizierbarkeit ohne Not aufgibt. Nimmt man die Verpflichtungen einer stringenten empirischen Methodik ernst, so vermag das vorgeschlagene Verfahren kaum zu überzeugen:

Das auf der ersten Stufe empfohlene Gespräch mit Bezugspersonen ist für Dritte weder nach Auswahl des Gesprächspartners noch nach Form oder Inhalt des Gesprächs nachvollziehbar oder gar replizierbar (und dürfte kaum je dokumentiert werden). Sein konkreter Verlauf wird zwangsläufig von den – auch unterbewussten – Vorverständnissen des befragenden Richters abhängen. Der Richter wird im Zweifel Formulierungen verwenden, die sein (bewusst oder unbewusst) bevorzugtes Ergebnis erkennen lassen, und die Pfadabhängigkeit des Gesprächs wird dieses in Form und Inhalt bestärken. Aussagekräftige empirische Studien bemühen sich deshalb um eine transparent systematisierte Datenerhebungsmethode, also ein gewisses Maß an Standardisierung. Statt eines Gesprächs könnten also strukturierte Interviews oder Fragebögen in Betracht gezogen werden.[101] Die Stichprobe müsste dann aber so gewählt werden, dass sich der Einfluss von Zufallsschwankungen minimieren lässt, was jedenfalls mindestens zwanzig Befragte voraussetzen dürfte.[102]

Methodisch angreifbar ist auch der auf der zweiten Stufe hinzutretende Ratschlag, solange Meinungen zu sammeln, bis sich ein klares Bild ergibt. Ein

---

[100] §§ 6 Abs. 1, 16 Abs. 2, 35 Abs. 2, 41 Abs. 2 ArbGG.

[101] *Döring/Bortz* (Fn. 70), 358 ff., 398 ff.

[102] Zu dieser Faustregel *Döring/Bortz* (Fn. 70), 302 („oft zwischen 20 und 30" in der qualitativen Forschung) und relativierend *Hamann* (Fn. 9), 69 bei und in Fn. 79 m.w.N.

solches Verfahren unterläge der in der psychologischen Literatur hinlänglich belegten Wahrnehmungsverzerrung zugunsten vorgefasster Ansichten (*confirmation bias*)[103] und erhöht die Wahrscheinlichkeit falsch-positiver Ergebnisse, also solcher, die vorgefasste Vermutungen kontrafaktisch bestätigen.[104] Deshalb gehört es zu den mittlerweile konsentierten Praktiken empirischer Disziplinen, den Umfang der Datenerhebung und die Auswertungsmethoden möglichst vorab festzulegen.[105]

Auch die auf der dritten Stufe vorgeschlagene Erweiterung des Befragtenkreises begegnet methodischen Bedenken. Die Richtersoziologie der letzten Jahrzehnte zeigt, dass Richter „als Berufsgruppe insgesamt einen sehr kleinen und demographisch homogenen Teil der Gesellschaft" darstellen,[106] was demgemäß auch für ihr soziales Umfeld und die auf direktem Weg erreichbaren Gewährspersonen gelten muss. Auch die Einbeziehung von Laien und mittelbar Bekannten („Bekanntenkreis von Referendaren") ist angesichts der zu erwartenden Korrelationen in Bildungsgrad und Sozialisierung kaum geeignet, hinreichend kritische Gegenstimmen zutage zu fördern – ganz abgesehen von den auch hier relevanten Verzerrungen durch soziale Erwünschtheit und bewusste Konsenssuche. Vorzuziehen wäre demnach idealerweise eine repräsentative Stichprobe ohne Abdeckungsfehler (d.h. Über- oder Unterrepräsentation einzelner Bevölkerungsgruppen) gegenüber der jeweils relevanten (z.B. von Formulararbeitsverträgen betroffenen) Gesamtbevölkerung.[107]

Diese Einwände aus der Perspektive einer wissenschaftlich disziplinierten empirischen Methodik lassen sich zwar unter Verweis auf die Prozessökonomie beiseiteschieben, jedoch sollte dann das Attribut einer „empirischen" Herangehensweise und der Eindruck der damit verbundenen argumentativen Autorität vermieden werden. Eine so wie vorgeschlagen praktizierte, weitgehend freie Gesprächsführung müsste ihren Mehrwert gegenüber der bisher üblichen argumentativen Auseinandersetzung mit den von den Parteien vorgetragenen Tatsachen wohl noch erweisen, zumal sie eine für Prozessparteien und nachfolgende Instanzen weniger transparente Überzeugungsbildung be-

---

[103] *Nickerson* Rev. Gen. Psy. 1998, 175 ("If one were to attempt to identify a single problematic aspect of human reasoning that deserves attention above all others, the *confirmation bias* would have to be among the candidates for consideration […] it appears to be sufficiently strong and pervasive that one is led to wonder whether the bias, by itself, might account for a significant fraction of the disputes, altercations, and misunderstandings that occur among individuals, groups, and nations.") und 210.

[104] *Simmons/Nelson/Simonsohn* Psy. Sci. 22 (2011), 1359, 1361.

[105] Näher *Hamann* (Fn. 9), 69, 71.

[106] *Hamann* (Fn. 14), 184, 197 m.Verw. auf *Röhl*, Rechtssoziologie: Ein Lehrbuch, 1987, 46, 57 f.

[107] Allgemein dazu *Döring/Bortz* (Fn. 70), 294 f.

dingt. Denn während die „klassische" richterliche Begründung zumindest ihre Gedankenschritte offenlegen und damit angreifbar machen muss, würde das von Stöhr vorgeschlagene Befragungsverfahren wohl weder die Identität noch die genauen Äußerungen der Befragten transparent machen können, sondern letztlich auf ein kritikimmunes Autoritätsargument der Form „Ich habe … Personen befragt und eindeutig die Auskunft erhalten, dass …" hinauslaufen. Zugleich dürfte mit dem steigenden Gewissheitsgefühl des betreffenden Richters seine wahrgenommene Argumentationslast sinken. All dies ließe sich nur vermeiden, wenn die Befragung nach transparenten (und damit kritisierbaren) Methoden und mit sorgfältig dokumentierten (und damit revisiblen) Befragungsdaten stattfände.

Da diese Elemente in Stöhrs Indizienverfahren fehlen, bietet es wohl kaum eine größere Richtigkeits*gewähr* für die Rechtsanwendung. Ob es zumindest das richterliche Gewissheits*gefühl* dialektisch irritieren kann, weil empirische Studien „vor allem […] das Eingeständnis" erfordern, „etwas nicht zu wissen",[108] bleibt im Fall seiner Anwendung mit Spannung zu beobachten.

4. Experimentelle Befragung (Vignettenstudien)

Soweit Stöhrs „empirische" Herangehensweise demnach im Interesse der Prozessökonomie hinter den methodischen Standards der empirischen Forschung zurücksteht, lässt sich abschließend noch aufzeigen, wie die Rechtspraxis zu einer wirklich empirisch kontrollierten (d.h. replizierbaren) Herangehensweise voranschreiten könnte.

Den Schlüssel dazu liefert Stöhr selbst, wenn er von einer „experimentellen Befragung" (oben bei und in Fn. 64) spricht. Diese Bezeichnung passt zwar nach dem gängigen Sprachgebrauch nicht auf seine Studie, lässt sich aber durch eine geringfügige Anpassung mit Leben füllen. Insofern weist Stöhrs Vorschlag in die richtige Richtung und bedarf lediglich einer Ergänzung:

Eine experimentelle Befragung setzt voraus, dass unter den Versuchsteilnehmern nach dem Zufallsprinzip mehrere Gruppen gebildet werden, von denen jede eine andere Variation des relevanten Befragungstextes vorgelegt bekommt. Diese Variationen unterscheiden sich lediglich in je einem Detail, so dass Unterschiede im Antwortverhalten zweifellos auf diesen Detailunterschied zurückgeführt werden können. Diese Herangehensweise ist in verschiedenen Disziplinen unter ganz unterschiedlichen Bezeichnungen bekannt – unter anderem als Befragungsexperiment, Szenariostudie, faktorielle Befragung oder Vignettenexperiment.[109] Allgemein könnte man von „Dekompo-

[108] *Stöhr* 2017 (Fn. 5), 144.

[109] Nachw. für diese und weitere Begriffe bei *Hamann* (Fn. 9), 180 mit Fn. 267.

sitionsmethoden" sprechen, „weil sie Entscheidungen danach aufzugliedern versuchen, durch welche Entscheidungsfaktoren sie verursacht werden", aber mit der schon früher vorgeschlagenen Terminologie sei hier von einer „Vignettenstudie" die Rede.[110]

Diese Methode wäre insofern besser geeignet gewesen, die von Stöhr beabsichtigten Ursachenbehauptungen zu belegen, als sie die in Befragungen stets problematischen Verzerrungen durch die soziale Erwünschtheit bestimmter Antworten neutralisieren und dank standardisierter Experimentalbedingungen auch den Rückschluss auf kausale Veränderungen unter ansonsten gleichen Bedingungen ermöglichen kann.[111] Diese allgemeinen Vorzüge (sowie Nachteile) der Methode wurden zusammen mit ihrer konkreten Anwendung auf den Bereich der Arbeitsvertragsklauseln bereits an einem anderen Beispiel illustriert.[112] Auch der jüngste Vorstoß in Richtung einer umfragebasierten Vertragsauslegung verwendete Vignettenstudien in eingängigen Beispielen,[113] und resümierte, dass trotz eines gewissen Mehraufwands „die erforderliche Technik alles andere als Hexenwerk" sei.[114] Daher genügt an dieser Stelle der Hinweis, dass auch Stöhrs Fragestellung damit wohl einer wirklich empirischen Herangehensweise hätte zugeführt werden können. Nähere Einzelheiten lassen sich dann der einschlägigen Einführungsliteratur entnehmen.[115]

V. Fazit: Perspektiven der Empirie im Zivilrecht

Der empirische Vorstoß, den Alexander Stöhr gewagt hat, belegt ein interdisziplinäres Interesse und methodisches Innovationspotential, von dem das Zivilrecht in den kommenden Jahren noch zehren wird. Wie überaus zeitge-

---

110 Terminologie und Zitate bei *Hamann* (Fn. 9), 180.

111 Näher *Hamann* (Fn. 9), 178 m.w.N.

112 *Hamann* (Fn. 9), 180–185; vgl. auch *Ben-Shahar/Strahilevitz* (Fn. 6), 21 für kontradiktorische Vertragsrechtsprozesse in den USA ("choosing between the two competing surveys presented by the litigants").

113 *Ben-Shahar/Strahilevitz* (Fn. 6), 23 ("Wave 2 of the survey randomly assigned half the respondents to read slightly modified versions of most vignettes. In wave 2 we were trying to determine how subjects would react if legally relevant or legally irrelevant changes were made to the vignettes.")

114 *Ben-Shahar/Strahilevitz* (Fn. 6), 40: "There is the additional cost of designing the survey in a professional manner and persuading the court that it is done reliably. That said, the techniques necessary are not exactly rocket science."

115 Z.B. *Keuschnigg/Wollbring* (Hrsg.), Experimente in den Sozialwissenschaften, 2015, 294–367, und die schon bei *Hamann* (Fn. 9), 180 Fn. 262 zitierten; neuester Überblick an unverhoffter Stelle: *Schnell/Schulz/Atzmüller/Dunger* (Hrsg.), Ärztliche Werthaltungen gegenüber nichteinwilligungsfähigen Patienten, 2017, insb. 11 ff., 29 ff., 145 ff. („kommentierte Literaturliste").

mäß sein Vorstoß erfolgte, zeigt sich nun an der neuesten US-amerikanischen Rechtsliteratur, die unlängst mit identischer Zielsetzung und vergleichbarer Emphase eine empirisch inspirierte Methodenreform angemahnt hat, die Stöhrs Vorschlag in vielem gleicht, wenngleich sie methodisch einen anderen Weg einschlägt.

Beide Programmschriften regen in der Gesamtschau zu einer Diskussion über den Erkenntniswert empirischer Methoden für die Vertragsauslegung und über die noch zu entwickelnden methodischen Standards einer genuinen Rechtsempirie an. Darüber gelangt die Disziplin womöglich zu einer umfassenden „Erkenntnistheorie, die speziell auf das juristische Feld und seine eigenartigen Zwecksetzungen und Problemstellungen zugeschnitten ist",[116] deshalb ist Stöhrs Ansatz trotz aller Kritik im Einzelnen überaus begrüßenswert.

Wo solche Kritik nicht ausbleiben konnte, belegt sie nur die Geburtswehen einer werdenden Methodik,[117] und bestätigt den von Stöhr selbst artikulierten Bedarf nach „Vorgaben und Leitlinien" für die empirische Rechtspraxis, sowie nach neuen „Impulsen" für die juristische Ausbildung (579). Erst wenn „die juristische Methodenlehre das Problem der Empirie in Betracht zieht"[118] und systematische Arbeitsanleitungen für die juristische Ausbildung entwickelt,[119] kann der Zivilrechtsdiskurs unterschiedliche Weltbeschreibungen produktiv verarbeiten und zu einer diskursrationalen Synthese von Normen und Sachverhalten voranschreiten. Um mit den Worten Stöhrs zu schließen: „Für die Rechtswissenschaft wäre dies eine Bereicherung."[120]

---

[116] *Augsberg* 2015 (Fn. 21), 71, 88.

[117] Vgl. auch *Kuntz* AöR 141 (2016), 645, 651: „Aus dem Wagnis interdisziplinärer Arbeit entstandene Beiträge zu verreißen, ist einfach. Fehler finden sich immer, Kritik an Schwerpunktsetzung ist häufig Folge unterschiedlicher wissenschaftlicher Geschmäcker".

[118] *Starck* JZ 1972, 609, 614 (These 2; ähnlich These 6).

[119] Vorschlag dazu jüngst bei *Hamann* (Fn. 2).

[120] *Stöhr* 2017 (Fn. 5), 149.