Diagnostica, 1994, *40*, Heft 1, 61–80

Zur Einzelfalldiagnose der Wertungskompetenz bei Fahrlässigen Brandstiftungen*

Wilfried Hommers

In 2 Untersuchungen mit 100 bzw. 88 Minderjährigen und 60 Erwachsenen werden Geschichten über drei – eine versehentliche, eine fahrlässige, eine absichtliche – Brandstiftungen mit bildlicher Unterstützung dargeboten. Durch Paarvergleiche zwischen bzw. Schätzurteile über die Geschichten wird ausgehend von der Wertungskomponente der Deliktsfähigkeit die Fähigkeit zur Unrechtserkenntnis für fahrlässige Schädigungen im Sinne der moralischen Differenzierungsfähigkeit untersucht, um einerseits die technische Durchführungsform und andererseits die einzelfalldiagnostische Auswertung nach psychometrischen bzw. varianzanalytischen Strategien zu erproben. Außerdem werden die Beziehungen der so erfaßten Wertungskomponenten zu Intelligenzuntertests geprüft. Aufgrund fehlender Korrelationen mit den Intelligenzmaßen empfiehlt sich die Anwendung des Verfahrens zur psychometrischen Begründung von Diagnosen zur zivilrechtlichen Verantwortlichkeit im Sinne des § 828 (2) BGB, damit sich die diesbezüglichen Diagnosen auch auf die Wertungskomponenten stützen können.

Determination of individual tort competency in cases of arson

Two studies with 100 or 88 minors and 60 or 49 adults employed paired comparisons or ratings to examine on a statistical or psychometrical basis the value-components of the major conditions of minors' tort competency, i. e. the ability to know right and wrong in the context of negligent as well as intentional harm. As some additionally assessed intelligence measures did not correlate with the tort competency measures, the latter should be used for psychometrically based competency diagnoses in liability law suits against minors according to § 828 (2) BGB, i. e. the German civil code.

Herleitung des Diagnoseparadigmas

Nach § 828 (2) des Bürgerlichen Gesetzbuches (BGB) ist der Minderjährige nach Vollendung des siebenten Lebensjahres bis zur Vollendung des 18. Lebensjahres nur unter gewissen Umständen nicht verantwortlich für den angerichteten Schaden aufgrund einer unerlaubten Handlung nach § 823 BGB, während jüngere Kinder nach § 828 (1) BGB immer nicht verantwortlich sind. Der Wortlaut des § 828 (2) BGB setzt dafür als Bedingung „wenn er bei der Begehung der schädigenden Handlung nicht die zur Erkenntnis der Verantwortlichkeit erforderliche Einsicht hat". Schon 1902 vom Reichsgericht in Zivilsachen und dann 1954 vom Bundesgerichtshof (vgl. die ausführliche Darstellung bei Undeutsch, 1967) wurde die zur Erkenntnis der Verantwortlichkeit erforderliche Einsicht, die Deliktsfähigkeit, näher bestimmt durch die beiden geistigen Entwicklungsstände, die zur Unrechtserkenntnis und zum Ver-

* Danksagung: Die empirischen Untersuchungen erfolgten unter einer Sachbeihilfe der Deutschen Forschungsgemeinschaft, Bonn (DFG Ho 920/2-2) an den Autor. Frau Dipl.-Psych. U Weist und Herr Dipl.-Psych. K. Feld halfen bei der Datenerhebung und bei der Auswertung. Frau M. Pirkner fertigte das Manuskript, die Tabellen und die Abbildung an. Zwei anonyme Gutachter gaben wertvolle Anregungen zur Überarbeitung der Erstfassung.

geltungspflichtverständnis befähigen: „diejenige geistige Entwicklung, die den Handelnden in den Stand setzt, das Unrecht der Handlung gegenüber dem Mitmenschen und zugleich die Verpflichtung zu erkennen, in irgendeiner Weise für die Folgen seiner Handlung einstehen zu müssen" (RGZ 1903, 53, S. 151f.). Für die Psychologie ergeben sich hieraus zwei empirisch zu bearbeitende Aufgaben: Prüfung der Entwicklung der geforderten Verantwortlichkeitskriterien im betreffenden Altersbereich (*de lege ferenda*) und Begutachtung von Minderjährigen, deren Verantwortlichkeit für angerichtete Schäden bezweifelbar erscheint (*de lega lata*). Der vorliegende Beitrag widmet sich den Erfassungsproblemen bei der Aufgabenstellung *de lege lata* (zur Fragestellung *de lege ferenda* vgl. Hommers, 1983 und 1991).

Da nach Bresser und Eisen (Eisen, 1977, S. 320) „zuverlässige, einhellige und begrifflich faßbare Kriterien" nicht vorzuliegen schienen, bediente sich die forensisch-psychologische Begutachtung neben der Exploration (vgl. Dauner, 1980, S. 122ff.) der Erfassung des allgemeinen intellektuellen Entwicklungsstandes (vgl. Ell, 1983, S. 65) oder Wissensprüfungen wie in den Mehrfachwahlantworten von Wille und Bettge (1971) und ließ im Extrem nur „eine grobe, dann auch allemal organisch begründbare Entwicklungsstörung" gelten (Bresser, 1972, S. 1294). Aber diese diagnostische Strategie muß gemäß der folgenden Überlegungen als eine nicht hinreichende Hilfsstrategie aufgefaßt werden.

Der besondere diagnostische Reiz der forensischen Begutachtungsaufgabe – insbesondere im behandelten Bereich, aber auch im allgemeinen – liegt in der geforderten doppelten Spezifität, maßgeschneidert gegenüber dem Individuum und gegenüber dem Begutachtungsanlaß, hier also dem Delikt, zu sein. Es sind also operationalisierbare Untersuchungsverfahren in doppelter Hinsicht *ad hoc* zu entwerfen: für das Individuum und für das Delikt. Damit wird aber die Verwendung standardisierter psychometrischer Tests, die zur Erfassung von Eigenschaftskonstrukten entwickelt wurden, grundsätzlich fragwürdig. Denn sie gehen in ihren Items auf die geforderte Deliktspezifität gar nicht (z.B. bei sogenannten kulturfreien Tests, in denen direkt auf den Alltag bezogenen Items fehlen) oder nicht hinreichend (z.B. in den an Wechsler oder Binet anschließenden Testkonzeptionen, in denen immerhin einige Items den Inhaltsbereich der unerlaubten Handlungen ansprechen, vgl. Untersuchung 2) ein.

Der fehlende Bezug auf das Delikt ist aber nur ein Teil der problematischen Verwendung standardisierter psychometrischer Tests. Denn die zur Erfassung der beiden Kriterien erforderliche Operationalisierung hat auch zu berücksichtigen, daß die rechtlich geforderten geistigen Entwicklungsstände das handlungsbezogene Wissen-Können und Werten-Können des Individuums umfassen. Die gesonderte Aufführung der Methoden zur Erfassung der „Ethischen Gefühlsbetonungen" (autoanamnestische Angabe, direkte Gefühlsproben, Exempla ficta, Fernaldsche Probe, ethische Begründungen) in der Systematik der Intelligenz- und Begabungsprüfungen des Psychiaters Ziehen (1923) unterstellte schon, daß aus dem Intelligenzdefekt oder aus dem Entwicklungsstand der anderen Bereiche der Intelligenz und Begabung zumindest nach dem damaligen Kenntnisstand nicht mit hinreichender

Sicherheit auf die Ausbildung der Wertorientierung geschlossen werden konnte. Empirisch zeigte sich denn auch, daß der Zusammenhang zwischen Intelligenztests und an Piaget (1932/1954) angelehnten Maßen des moralischen Urteilens bei Kindern höchstens mäßig stark (stets kleiner als r=.50) ausgebildet war (Müller, 1966; Hommers, 1983, S. 94ff.). Daher wäre er für individualdiagnostische Zwecke praktisch nur begrenzt verwendbar. Weiterhin blieb nach statistischer Ausschaltung des mäßigen Intelligenzeinflusses (Kurtiness & Pimm, 1983) der Zusammenhang zwischen Alter und Maßen des moralischen Urteilens in signifikanter Größe bestehen, so daß weitere der Entwicklung unterliegende Faktoren das moralische Urteilen bedingen müssen. Andererseits stützten diese Befunde die Vermutung, daß in den Zusammenhängen mit der Intelligenz methodisch bedingt nur oder vor allem die kognitive Komponente, das Wissen-Können, zum Ausdruck kam und das Werten-Können unerfaßt blieb. Also ist nicht nur der Schluß vom allgemeinen intellektuellen Entwicklungsstand auf die Deliktsfähigkeitskriterien, sondern auch der Schluß von reinen Wissensprüfungen auf die Deliktsfähigkeitskriterien problematisch. Schließlich erhob sich gegen die explorative Erfassung des Werten-Könnens das Bedenken, daß gerade unter der üblichen diagnostischen Strategie, die von einem Entwicklungsrückstand als Exkulpierungshypothese ausgeht, auch die verbale Äußerungsfähigkeit rückständig sein dürfte, so daß vermutlich auf diesem Kommunikationskanal keine validen Befunde zu erlangen sind.

Der vorliegende Beitrag versucht daher, eine doppelt spezifische und gleichzeitig psychometrische Alternative für die im Grunde unzulässigen diagnostischen Strategien zur Deliktsfähigkeitsbegutachtung aufzuzeigen. Der Ansatz der vorgestellten Alternative orientiert sich ganz eng am spezifisch forensisch-diagnostischen Leitziel der Kriterienausschöpfung: Für jede forensisch-diagnostische Aussage sind juristisch interpretierbare oder teilweise von dort vorgegebene Kriterien zu verwenden (Hommers, 1992). Die hier vorgestellte Alternative setzt dabei am Werten-Können an, da anscheinend bislang für das Werten-Können keine standardisierten Methoden bereitstanden. Da von allen unerlaubten Handlungen nach § 823 Bürgerliches Gesetzbuch (BGB) unter Kindern die Brandstiftung ein Delikt zu sein scheint, das relativ häufig zu forensisch-psychologischen Begutachtungen der zivilrechtlichen Verantwortlichkeit nach § 828 (2) BGB führte (Dauner, 1980; Ell, 1983), nimmt der vorliegende Beitrag zwar konkret auf die Beurteilung der Wertungsfähigkeiten anhand von fahrlässigen Brandstiftungen Bezug. Im Prinzip ist das Vorgehen aber auf andere Deliktarten übertragbar.

Die vorgestellte Methode knüpft an die bei Ziehen (1923) erwähnte Methode der direkten Gefühlsproben an. Dort wurden unmittelbar beobachtbare emotionale Reaktionen durch gezielte Fragen ausgelöst und mit einem klinischen Erfahrungsstandard verglichen. Die hier vorgeschlagene Verfahrensweise nimmt demgemäß in der Operationalisierung der abhängigen Variable des Untersuchungsverfahrens ebenfalls ein geringes Anforderungsniveau an die Äußerungsfähigkeit der Probanden an. Sie stellt aber an die Stelle der gezielten Fragen *Geschichten über verschieden verlaufende Brandstiftungen durch ein Kind und dessen Nachtatverhalten.* Die Merkmale der Gefühlsproben werden weiterhin abgeändert durch Verwendung

eines *quantitativ abgestuften Schätzurteils* auf einer Strafe- oder auf einer Gut-Böse-Skala und durch *Bezug auf den normativen Standard*, daß das Schätzurteil in Abhängigkeit von vorgegebenen, *mehrfaktoriell strukturierten Informationen* vom Urteiler *systematisch und rechtlichen Anforderungen gemäß variiert* werden kann. Die Fähigkeit zur systematischen Variation der wertenden Beurteilungen als Erfassungsmethode für die geforderte Unrechtserkenntnis zu wählen, ist dabei an dem klassischen Rechtsbegriff des Discernement (Unterscheidungsfähigkeit) orientiert. Discernement erfordert aber nicht nur Unterscheidung, sondern auch rechtlich richtige Unterscheidung, was sich über die festgelegte Richtung der Urteilsunterschiede berücksichtigen läßt. Die Verwendung mehrfaktorieller Stimulusgrundlagen erfolgt weiterhin bezogen auf die beiden Kriterien der Deliktsfähigkeit, die „zugleich" zu erfüllen sind. Durch mehrfaktoriell konstruierte Darbietungen soll also die enge Verbundenheit der Unrechtserkenntnis und des Verständnisses für die Vergeltungspflicht sowohl in der rechtlichen Forderung als auch im erwünschten alltäglichen Geschehen operationalisiert werden. Für die Erfassung der Unrechtserkenntnis wird Information über das Verschulden variiert, so daß Unterscheidungsfähigkeit hinsichtlich dieser Informationen Fähigkeit zur Unrechtserkenntnis anzeigt, wenn absichtliche oder fahrlässige Schädigung für moralisch schlechter oder strafwürdiger angesehen werden als unabsichtliche. Für das Verständnis der Vergeltungspflicht wird Information über das Nachtatverhalten des Täters variiert. Die rechtlich zu fordernde Unterscheidungsfähigkeit bildet das geringe Anforderungsniveau ab, das die höchstrichterliche Rechtsprechung an das Vergeltungspflichtverständnis stellte („irgendein Verständnis der Pflicht" und z. B. nicht das Verstehen des Unterschieds zwischen eigener Ersatzleistung und derjenigen durch eine Versicherung), wenn ein erfolgtes Einstehen für die Folgen eine unerlaubte Handlung als moralisch besser oder weniger strafwürdig erscheinen läßt, als wenn sie nicht von einem Einstehen für die Folgen gefolgt wird.

Das Ausmaß der systematischen Urteilsvariation bietet dann die Grundlage für den Einsatz von varianzanalytischen oder konsistenzanalytischen Prozeduren zur statistisch abgesicherten Erfüllung der zweiten Seite der geforderten Spezifität, der auf das Individuum bezogenen diagnostischen Entscheidungsfindung in der Forensischen Diagnostik. Trotz Abkehr von der Verwendung psychometrischer Testverfahren wird also ihr Vorzug in der zufallskritischen Bewertung von Untersuchungsergebnissen beibehalten. Dabei wird im Sinne von Huber (1973) für das statistische Entscheidungsverfahren sowohl der indirekte Ansatz (Untersuchung 1) als auch der direkte Ansatz (Untersuchung 2) der psychometrischen Einzelfalldiagnostik zur Schätzung des Meßfehlers verwendet.

Untersuchung 1

Die zu beurteilenden Geschichten enthielten zwei rechtlich wesentliche Informationen – zum Vergeltungspflichtverständnis und zur Unrechtserkenntnis – und eine unwesentliche über die Schadenshöhe der Brandstiftung, welche in einem Stimulusplan kombiniert waren. Die Information zum Vergeltungspflichtverständnis

wurde operationalisiert durch die Darbietung der Nachtatinformationen in den Stufen: Entschuldigung versus Nicht-Entschuldigung. Diese Operationalisierung erschien wegen der rechtlich mit geringem Niveau ansetzenden Forderung *irgendeines* Verständnisses der Vergeltungspflicht noch vertretbar, obwohl sie wegen der Rechtsfolge des materiellen Schadensersatzes unangemessen erscheinen kann. In Vorstudien, die materielle Ersatzleistungen als Urteilsgrundlagen verwendeten, erwiesen sich diese aber bei Minderjährigen genauso wirksam wie die Entschuldigungsinformation (Hommers, 1988a,b). Die Information zur Unrechtserkenntnisfähigkeit wurde operationalisiert durch die Darbietung der Verschuldensinformationen in den Stufen: Wut, Fahrlässig und Versehen. Als Anzeichen für die beiden ausgebildeten Wertungskomponenten sollten dann die gerichtet unterschiedlichen Beurteilungen der jeweiligen Stufen der beiden Faktoren gelten, die durch Anzeigen auf einer Strafe- bzw. auf einer Gut-Böse-Skala beobachtbar wurden. Die Verwendung der dritten Information zur Schadenshöhe hatte zwei Funktionen. Einerseits sollte sie den Urteilern die Möglichkeit zur Beachtung von Informationen bieten, die hinsichtlich der rechtlich definierten Deliktsfähigkeit unrelevant waren. Andererseits führte ihre Verwendung zu einer gerade ausreichenden Zahl von Urteilen für die konsistenzanalytischen Schätzung der Reliabilität der Beurteilungsunterschiede, insbesondere für die Bewertung der rechtlich relevanten Maße für die Deliktsfähigkeit.

Methode

Szenario. Die 160 Probanden waren 60 Kinder des von den deliktrechtlichen Begutachtungsaufträgen relevanten Altersbereichs zwischen 7 und 11 Jahren, 40 Jugendliche zwischen 12 und 17 Jahren sowie 60 Erwachsene zu Vergleichszwecken. Sie hörten zwölf Geschichten über einen Brandschaden und sahen dazu jeweils vier Bildtafeln, von denen jeweils zwei eine der drei Entstehungsweisen des Brandschadens (Verschulden), jeweils eine dritte einen der zwei Schadensumfänge (Schaden) und jeweils eine vierte eines von zwei Nachtatverhaltensweisen des handelndes Kindes (Entschuldigung) darstellten.

Die intentionale der drei Verschuldensbedingungen beschrieb eine ungerechtfertigte Affekttat: „Ein Kind, zu Besuch auf dem Bauernhof, nascht Kirschen und wird erwischt; zur Strafe soll es den Geräteschuppen aufräumen; darüber wütend zündet es im Schuppen einen Strohballen an". Die fahrlässige Verschuldensbedingung war als eine von einem Mißgeschick begleitete Mißachtung einer Warnung zu verstehen: „Trotz des Verbots vom Bauern spielt ein Kind in der Nähe des Geräteschuppens mit Feuer; das trockene Gras herum fängt plötzlich Feuer und setzt einen Strohballen im Schuppen in Brand". Die versehentliche Verschuldensbedingung war als eine Ungeschicklichkeit bei guten Motiven zu verstehen: „Ein Kind hilft die verlorene Geldbörse des Bauern im Schuppen suchen; in einer dunklen Ecke zündet es ein Streichholz an, um besser zu sehen; beim Bücken kommt es an einen Strohballen, der zu brennen anfängt". Jedes dieser Geschehnisse wurde mit zwei im wesentlichen schwarz-weißen, bezüglich Strohballen und Feuer aber farbigen Bildern unterlegt.

Die Schadenshöhe wurde durch ein Bild beschrieben, das entweder einen verkohlten Strohballen vor dem Geräteschuppen („Der Strohballen verbrennt") zeigte oder den verkohlten Strohballen und zusätzlich den verkohlten Eingang des Schuppens („Der Strohballen und ein Teil des Schuppens verbrennen").

Die beiden Stufen von Entschuldigung wurden wiederum mit einem Bild dargestellt. Der Text der Bedingung lautete: „Das Kind entschuldigt sich nicht und hilft dem Bauern nicht" (im Bild war ein deutlich abseits, mit verschränkten Armen und abgewendet vom Bauern stehendes Kind zu sehen) oder „Das Kind entschuldigt sich und verspricht dem Bauern bei der Ernte mitzuhelfen" (Bauer und Kind schüttelten in der bildlichen Darstellung einander die Hände). Das Hilfeversprechen wurde mit einer Sprechblase dargestellt, die das Kind mit einem Heuwagen und mit einer Gabel einen Strohballen hebend zeigte.

Die bildlichen Darstellungen waren so groß gehalten, daß sie zusammen mit dem Text der Geschichte und der Skala eine Vorlage der Größe DIN A 4 ergaben. Das Stimulusmaterial war zu einem Fragebogen zusammengestellt und die Erläuterungstexte wurden bei den beiden jüngsten Altersgruppen begleitend besprochen.

Aufgabe. Die Probanden sollten beurteilen, wie gut oder böse sie das in der Geschichte Dargestellte fanden, bzw. wieviel Strafe das Kind erhalten solle. Eine Aufforderung zur Identifikation des Urteilers mit dem Täter oder dem Geschädigten erfolgte nicht. Die Urteile wurden von der Hälfte der Probanden auf einer Gut-Böse-Skala abgegeben. Die schon in früheren Arbeiten des Verfassers verwendete Urteilsskala war eine Abbildung von je 13 aneinandergereihten weißen und schwarzen Streifen, die zur Mitte hin abnahmen. An beiden Enden war ein Schema mit einem fröhlichen oder mit einem traurigen Gesichtsausdruck zu sehen. Die andere Hälfte der Probanden urteilte auf einer 26stufigen Strafenskala. Die Probanden gaben ihre Urteile durch Ankreuzen des Streifens an, der zur Kennzeichnung der betreffenden Geschehensfolge einer Geschichte treffend erschien. Unterschiede zwischen den beiden Responsearten wurden hinsichtlich der folgenden Auswertungsaspekte nicht gefunden. Jedoch trat ein für die übrigen Ergebnisse unbedeutender Bodeneffekt (ordinale Verschulden-Entschuldigung-Interaktion) bei der Verwendung der Strafe auf, weil eine Entschuldigung bei versehentlicher Tat schon in den moralischen Gut-Bereich hineinragte, was aber auf der Strafeskala nicht dargestellt werden konnte.

Den Probanden wurde ihre Aufgabe erläutert durch zwei ebenfalls vierteilige Bilderreihen der Endanker. Der Begleittext des hohen Endankers lautete: „Ein Kind soll dem Bauern bei der Ernte mithelfen. Weil es dazu keine Lust hat, zündet es den Schuppen mit den Erntemaschinen an. Dabei verbrennt ein Teil des Schuppens und der Traktor wird beschädigt. Das Kind entschuldigt sich nicht und hilft dem Bauern nicht, sondern freut sich noch über den angerichteten Schaden." Der Begleittext zum niedrigen Endanker lautete: „Ein Kind hilft dem Bauern beim Aufräumen im Geräteschuppen. Plötzlich gibt es einen lauten Knall und einen hellen Blitz. Der Bauer bekommt einen großen Schreck, aber das Kind beruhigt ihn und sagt, daß nur die Stallampe durchgebrannt ist." Außerdem erhielten die Probanden drei der zwölf Geschichten zur Übung, bevor sie die zwölf Kombinationen beurteilten.

Ergebnisse

Individuelle ANOVA: Stimulus-Interaktionen als Fehlervarianzen. Ein erster, allerdings nicht hinreichender Zugang zur individuellen Kriterienprüfung besteht trotz einmaliger Beurteilung jeder Geschichte in der Prüfung der statistischen Signifikanz von Haupteffekten. Indem die Interaktionen der Stimulusfaktoren als Ausdruck von Zufallsvariationen aufgefaßt werden, können für jeden Probanden die Haupteffekte statistisch geprüft werden. Da die Prüfung der Stimulusinteraktionen in den drei Gruppen keine signifikanten F-Werte – außer für die Verschulden-Entschuldigung-Interaktion des Bodeneffekts mit einem Varianzanteil von 1% – erbrachte, konnte die bei diesen individuellen ANOVAs gemachte Annahme keine groben Verzerrungen bringen. Das Ergebnis ist in Tabelle 1 aufgeführt, die das Vorkommen von Beurteilungstypen bei den drei Altersgruppen auflistet.

Tabelle 1: Relative Häufigkeiten für drei Altersgruppen von einzeln oder kombiniert auftretenden individuellen Haupfeffekten (Beurteilungstypen) in der Beurteilung dreifaktorieller Stimuli aus Entschuldigung (EN)-, Schadenshöhe (SC)- und Verschulden (VE)-Informationen über eine Brandstiftung eines Kindes. Die statistischen Prüfungen der Effekte erfolgten individuell gegen die Stimulusinteraktionen unter Anwendung des 10%-Niveaus bei EN bzw. SC mit F(1,7)=3.59 und bei VE mit F(2,7)=3.26.

	Altersgruppe		
	7–11jährige (N=60)	12–17jährige (N=40)	21–40jährige (N=60)
Beurteilungs-Typ			
Nur Entschuldigung	44	10	8
Nur Schaden	3	–	–
Nur Verschulden	2	–	–
EN + SC	17	8	3
EN + VE	21	62	57
SC + VE	–	–	3
EN + SC + VE	8	20	25
Fehlanzeige	5	–	3

Anmerkung: Bei weiteren 20 untersuchten Vorschülern im Alter unter 7 Jahren war zu 80% nur EN signifikant und zu 10% EN + SC, sonst keiner der Haupteffekte. Auf die Untersuchung weiterer Vorschüler wurde daher mit diesem Design verzichtet.

Man sieht in Tabelle 1 im wesentlichen, daß Probanden mit genau einem signifikanten Haupteffekt fast nur in der jüngsten Probandengruppe vorkamen. Dafür kamen Probanden, die zwei oder drei Geschichteninformationen kombinierten, mit zunehmendem Alter häufiger vor. Schließlich wurde genau ein Verstoß gegen die Regel gefunden „wenn Verschulden signifikant (=beachtet), dann auch Entschuldigung", wodurch eine implizite Regel der höchstrichterlichen Rechtsprechung gestützt wurde, daß sich die Prüfung des Vergeltungspflichtverständnisses nach Vollendung des siebenten Lebensjahres erübrige (vgl. Hommers, 1983, S. 22ff.).

In Hinsicht auf eine deliktrechtliche Interpretation positiver Befunde in den Haupteffekten erscheint dieses Vorgehen für die Deliktsfähigkeitsbeurteilung aber problematisch. Denn selbst wenn der Urteiler seine Urteile variierte, so daß man ihm die Fähigkeit zur bloßen Unterscheidung unterstellen muß, entspricht diese Feststellung noch nicht der rechtlich geforderten gerichteten Unterscheidungsfähigkeit. Bei Ausbleiben des Wirksamkeitsbefundes einer Informationsart ist die Interpretation Unfähigkeit weiterhin aus zwei Gründen problematisch. Einerseits können systematische und u. U. interindividuell variierende Interaktionen der Stimuli in den Urteilen vorliegen. Andererseits kann angesichts nur eines Beurteilungsdurchganges die statistische Effizienz nicht hinreichend sein. Im Prinzip gäbe es dafür zwei Verbesserungsmöglichkeiten. Erstens könnten die Urteile eines Probanden graphisch analysiert werden, was u. U. schon Hinweise auf die geforderte Urteilsrichtung wie auf systematische Stimulusinteraktionen bringen könnte und dem Zivilrichter als Nachweis für die Deliktsfähigkeit genügen könnte. Zweitens könnte mit Hilfe weiterer Beurteilungsdurchgänge individuell die statistische Effizienz sowohl für die Haupteffekte als auch für die Interaktionseffekte erhöht werden, so daß auch aus zufallskritischer Perspektive das diagnostische Urteil Unfähigkeit besser gesichert erscheinen könnte. Diese Vorgehensweisen wären im Ernstfall einer Begutachtung zu empfehlen, da dann der höhere Zeitaufwand im Gegensatz zu dieser Studie gerechtfertigt erscheinen könnte. Die angeschnittene Thematik der statistischen Effizienz wird hier in der Untersuchung 2 wieder aufgegriffen, da dort aufgrund geänderten Vorgehens zwei Schätzurteile pro Geschichte erlangt werden.

Reliabilitätsschätzungen für summierte Differenzwerte. Da das individuelle ANOVA-Verfahren hinsichtlich der geforderten spezifischen Unterscheidungsleistungen im Verschulden nicht aussagekräftig ist und die geforderte gerichtete Unterscheidungsleistung durch die Signifikanz nicht direkt geprüft wird, wurde auch nach einem anderen Analyseansatz vorgegangen. Dieser übertrug die Vorgehensweise Psychometrischer Tests auf spezifisch gebildete Differenzen zwischen den Urteilen (vgl. Hommers, 1991).

Ähnlich wie bei Psychometrischen Tests ließen sich Summenscores bilden, deren Reliabilitäten mit Cronbachs ALPHA geschätzt werden konnten. Der Summenscore für die Differenz Fahrlässig-Versehen entstand z. B. durch die Aufsummierung der vier Differenzen, die bei jeder der vier möglichen Kombinationen der beiden weiteren Faktoren Schadenshöhe und Entschuldigung gebildet werden konnten (also summiert über „Scheunentor-Ja", „Scheunentor-Nein", „Hundehütte-Ja", „Hundehütte-Nein"). Entsprechend wurde für die vier anderen mit dem 3x2x2-Plan bestehenden Möglichkeiten zu derartigen Differenzwerten vorgegangen: Sechs Differenzen für den Entschuldigungseffekt, sechs Differenzen für den Schadenseffekt, vier Differenzen für Wut-Fahrlässig, vier Differenzen für Wut-Versehen. Die Annahme bei der Reliabilitätsschätzung über Cronbachs ALPHA ist, daß die aufsummierten Differenzen äquivalent sind, was bei fehlender Stimulusinteraktion gegeben wäre. Andernfalls handelt es sich bei der Reliabilitätsschätzung nur um eine untere Grenze der Reliabilität (Kristof, 1983).

Das in der Tabelle 2 aufgeführte Ergebnis der Reliabilitätsschätzungen ist wegen der Minimalschätzung durch ALPHA und weiterhin wegen der denkbaren Reliabili-

tätskorrektur nach Spearman-Brown (Testverlängerung) im Vergleich mit psychometrischen Tests günstiger zu bewerten, als es in den diagnostisch interessanten Zeilenblöcken für Entschuldigung, Wut-Versehen und Fahrlässig-Versehen erscheint. Inhaltlich besagt Tabelle 2 in diesen Zeilen, daß mit Hilfe des individualdiagnostischen Ansatzes der indirekten Meßfehlerbestimmung in den rechtlich relevanten Unterscheidungsleistungen Einzelfalldiagnosen schom im Grundschulalter vorgenommen werden können. Wegen der dort größeren Standardabweichungen sind in dem Altersbereich größere Urteilsunterschiede erforderlich als im Hauptschul- oder Erwachsenenalter.

Die Anwendung des gleichen Verfahrens zur Gewinnung von summierten Differenzwerten bei den beiden anderen, aber rechtlich irrelevanten Möglichkeiten (Wut-Fahrlässig und Schaden) wäre schließlich wegen der geringeren Reliabilitätsschätzungen ohnehin fragwürdig. Trotzdem haben sie einen Nutzen. Denn sie belegen, daß die summierten Differenzwerte nicht lediglich einen „Generalfaktor" der interindividuell unterschiedlichen Ausnutzung der Skala darstellen. Das gleichzeitige Bestehen und Nicht-Bestehen von Effekten der Geschichteninformationen kontrolliert sozusagen dieses denkbare Artefakt.

Tabelle 2: Reliabilitätsschätzungen nach Cronbach (REL), Mittelwerte (M) und Standardabweichungen (S) für 4 Differenzwertsummen bei Beurteilungen von 12 (3x2x2) Verschulden-Schaden-Entschuldigung Bild-Stimuli über einen minderjährigen Brandstifter.

		Altersgruppe 7–11jährige (N=60)	12–17jährige (N=40)	21–40jährige (N=60)
Differenzwertsumme				
Entschuldigung				
bei 6 Schaden-	REL	.76	.82	.76
Verschulden	M	9.5	9.9	8.7
Kombinationen	S	5.0	4.2	3.1
Wut-Fahrlässig				
bei 4 Schaden-	REL	.13	.52	.49
Entschuldigung-	M	1.2	1.6	2.8
Kombinationen	S	1.4	2.1	2.5
Wut-Versehen				
bei 4 Schaden-	REL	.60	.71	.62
Entschuldigung-	M	3.5	6.8	6.8
Kombinationen	S	4.5	4.0	3.4
Fahrlässig-Versehen				
bei 4 Schaden-	REL	.66	.69	.71
Entschuldigung-	M	2.3	5.2	4.0
Kombinationen	S	4.4	3.6	3.6
Schaden				
bei 6 Verschulden-	REL	.59	.16	.54
Entschuldigung-	M	0.4	1.6	1.6
Kombinationen	S	2.0	1.5	2.0

Diagnose der Unterscheidungsleistungen. Die Reliabilitätsschätzungen und die Standardabweichungen wurden nach dem indirekten Ansatz der Klassischen Testtheorie zur Bestimmung des Meßfehlers der Differenzwerte benutzt (vgl. Huber, 1973; Kristof, 1983). Mit diesem Meßfehler waren die individuellen Differenzwertsummen gegen die Nullhypothese „Keine Unterscheidung der in der betrachteten Differenz ausgewählten Stimulusinformationen" bei Verwendung des einseitigen Irrtumsniveaus (z.B. wie hier $p<.05$) zu prüfen.

Bezüglich der drei Verschulden-Stufen ergaben sich in Muster zusammenfaßbare Diagnosen der moralischen Unterscheidungsleistungen in der Unrechtserkenntnis. Mit den Reliabilitäten der Gesamtgruppe (N=160) wurde das Ergebnis der Tabelle 3 erlangt. Hervorzuheben ist daran, daß Minderjährige und Erwachsene relativ etwa gleichhäufig die Stufen Fahrlässigkeit und Versehen unterschieden. Die Prüfung der Altersunterschiede in den Häufigkeiten zeigte, daß Erwachsene relativ häufiger gerade diese beiden Stufen gleichsetzten und nur die Brandstiftung aus Wut im Urteil davon unterschieden. Außerdem war die Gleichsetzung aller drei Verschulden-Stufen in zu erwartender Weise altersabhängig.

Tabelle 3: Diagnosehäufigkeiten in Mustern für beurteilte Verschuldensbedingungen (W: Wut; F: Fahrlässig; V: Versehen) bei $p<.05$ einseitig als Kriterium für nicht gleich beurteilte Bedingungen mit Gesamt-Reliabilitäten und Gesamt-Standardabweichungen für die individuellen Effektstärken.

Muster	Minderjährige (N=100)	Erwachsene (N=60)	CHI^2
W=F<V	34	19	
W>F=V	5	15	12.00 $p<.001$
W>F>V	12	14	
W=F=V	35	8	6.55 $p<.02$
W>V,F?	5	0	
Andere	9	4	

Anmerkung: W>V,F? bedeutet, daß zwar Wut signifikant anders beurteilt wurde als Versehen, aber die Vergleiche dieser beiden mit Fahrlässigkeit nicht signifikant ausfielen.

Die moralische Unterscheidungsleistung zwischen den Stufen der Entschuldigungsinformation für das Vergeltungspflichtverständnis war demgegenüber zu 93% zu sichern, was wieder mit der impliziten rechtlichen Entwicklungstheorie eines der Unrechtserkenntnis vorauseilenden Vergeltungspflichtverständnisses konform ist (Hommers, 1983). Man wird also kaum zu begutachtende Probanden finden, die zu diesem Wertungsvorgang nicht fähig sind. Die Anwendung dieses individuellen Diagnoseprinzips auf den rechtlich irrelevanten Schadenseffekt führte zu einer Häufigkeit von 20% an gesicherten Unterscheidern der beiden Schaden-Stufen.

Diskussion. Aus zwei Gründen war der indirekte Ansatz zur Schätzung des Meßfehlers unbefriedigend: Erstens waren die Reliabilitätsschätzungen, anders als bei Hommers (1991), hier geringer (kleiner als REL = .80) ausgefallen als für eine Anwendung in der Individualdiagnostik optimal. Zweitens enthielt die Verwendung der Gruppenparameter für Reliabilität und Standardabweichung innerhalb des forensischen Kontexts prinzipielle Probleme, die nur bei Fehlen anderer Möglichkeiten in den Wind geschlagen werden durften. Auf diese Art konnte das eine der drei spezifischen Leitziele, die den Sonderstatus der Forensischen Diagnostik begründen, „Validitätsmaximierung und Beweisfähigkeit" nicht hinlänglich erfüllt werden, wonach Forensische Diagnostik im Gegensatz zu anderen Bereichen der Psychodiagnostik sich dadurch auszeichnet, daß sie maximale (nicht nur linear-korrelativ optimierte) Validität hinsichtlich ihres gerichtlichen Beweiswerts besitzen muß (vgl. Hommers, 1992). Vor Gericht kann nur ein individueller Maßstab unter § 828 (2) BGB maßgeblich sein. Hat z. B. ein Kind ungewöhnlich präzise Urteile gegeben, würde es bei geringen Differenzen als unfähig gelten aufgrund des in der Signifikanzprüfung angewendeten Gruppenparameters. Entsprechend könnte ein sehr variabel über gleiche Vorlagen urteilendes Kind bei gleichzeitig hinreichend großen Differenzen wegen des in der Prüfung angewendeten Gruppenparameters zu Unrecht als gesichert unterscheidungsfähig gelten. Veranschaulicht sieht man dieses Problem in der Tabelle 3 in den relativ großen Häufigkeiten völliger Gleichsetzungen der Stufen Wut, Fahrlässig und Versehen durch Erwachsene. Man darf annehmen, daß hierfür präzise, aber mit feinen Unterschieden urteilende Erwachsene der Grund waren. Daher wurde in der anschließenden Untersuchung eine volle Individualisierung der statistischen Absicherung der Unterscheidung der drei Verschulden-Stufen vorgenommen, indem pro diesbezüglicher Stimulusbedingung zwei Urteile vorlagen, was durch den besonderen Untersuchungsablauf der den Schätzurteilen vorangestellten Paarvergleiche sozusagen nebenbei erreicht wurde. Das geschah vor allem, um den Probanden die in anderen denkbaren Vorgehensweisen enthaltenen, möglicherweise monoton wirkenden Wiederholungen zu ersparen, ermöglichte aber auch den Methodenvergleich.

Untersuchung 2

Die Untersuchung 2 prüfte nur das Urteil in Hinsicht auf die Unrechtserkenntnisfähigkeit, da die Untersuchung 1 ergab, daß im wesentlichen alle Minderjährigen gemessen am Kriterium der gerichtet unterschiedlichen Wertung von hilfeversprechender Entschuldigung und Nicht-Entschuldigung das Vergeltungspflichtverständnis im Sinne der gering ansetzenden Anforderung der Rechtsprechung besaßen. Untersuchung 2 bezog weiterhin Vorschüler statt der Hauptschüler ein, um den aufgrund erheblicher Vereinfachung bzw. intensivierter Instruktion der Aufgabe vermutlich zwischen dem Vor- und Grundschulalter ablaufenden Entwicklungsvorgang der Unrechtserkenntnisfähigkeit darstellen zu können. Schließlich wurde mit Intelligenzleistungssubtests der IQ der Probanden erfaßt, um die These der Ersetzbarkeit des hier vorgestellten Verfahrens zu prüfen.

Methode

Abgewandeltes Brandstiftungsszenario. Als Untersuchungsmaterial diente wieder das Brandstiftungsszenario, es war jedoch anders aufgebaut als in Untersuchung 1. Das Schadensausmaß (Hundehütte oder ein Teil des Schuppens brennen ab) war nun nur als Gruppenfaktor variiert, da es sich in Untersuchung 1 als unwichtig für die Untersuchungsziele erwiesen hatte. Für jeden einzelnen Probanden war der Schaden also konstant. Die Informationen über das Nachtatverhalten waren durch Einführung einer Dritt-Entschädigungskomponente komplexer gestaltet: „Das Kind geht zum Bauern hin und entschuldigt sich bei ihm; und die Versicherung bezahlt den Schaden nicht" (Nur Entschuldigung) versus „Das Kind verweigert zwar die Entschuldigung, aber die Versicherung ersetzt den Schaden" (Nur Dritt-Entschädigung). Diese Bedingungen des Nachtatverhaltens wurden mit den drei Verschulden-Stufen Wut, Fahrlässig und Versehen unter konstantgehaltenem Schadensausmaß zu Geschichten kombiniert und sowohl auf einer bipolaren Gut-Böse-Skala beurteilt, als auch paarweise verglichen.

Jede Bedingungskombination aus variierendem Verschulden und Nachtatverhalten und konstantem Schaden wurde durch je fünf bebilderte 15 x 11 cm große Tafeln veranschaulicht, je zwei für die variierenden Bedingungen und eine für das Schadensausmaß. Diese fünf Tafeln mußten zunächst unter Verbalisierung der Geschichte von dem Probanden in die richtige Reihenfolge gebracht werden. Dann wurde eine weitere Bildgeschichte in ungeordneter Reihenfolge vorgezogen, die der Proband unter die erste Geschichte legen sollte, indem er sie ebenfalls in die richtige Bildreihenfolge brachte. Der zweite Stimulus unterschied sich nur im Verschulden vom ersten. Der Proband hatte nun die Aufgabe zu sagen, welche Geschichte schlimmer oder besser sei. War das geschehen, sollte er auch noch auf der vor ihm liegenden, 35 cm langen, aus zwei anschwellenden Dreiecken aus je 13 schwarzen bzw. weißen Streifen bestehenden Gut-Böse-Skala zeigen, wie gut oder böse jedes Geschehen war. Das erfolgte für jeweils alle drei Paare aus den drei Verschuldensbedingungen und zwar in zufällig bestimmter Reihenfolge, ob zuerst für die Bedingung Nur Entschuldigung oder zuerst für Nur Dritt-Entschädigung. Es lagen somit 6 Paarvergleichsurteile und 6 Meßwiederholungen für die Schätzurteile vor.

Ablauf der Untersuchung. Zusätzlich zu den Endankern, die auch in der Untersuchung 1 verwendet wurden, wurde eine Übungsphase vor den Paarvergleich eingeschoben, die in drei Schritten erfolgte:

– *Einzelinformationen:* Zuerst wurden die Probanden mit allen Informationen, die später in den kombinierten Geschichten vorkamen, und mit dem Gebrauch der Skala vertraut gemacht. Dazu erklärte der/die Versuchsleiter/in einzeln die Schaden-Stufen, d. h. welcher Schaden entstanden war, die Verschulden-Stufen, d. h. wie es zu dem Schaden kam und die Nachtat-Stufen, d. h. ob sich das Kind entschuldigte und ob die Versicherung für den Schaden aufkam. Zu den mündlichen Erläuterungen wurden jeweils die Bildtafeln gezeigt. Der Proband sollte sich jeweils merken, was die Bildtafeln darstellten. Nachdem der/die Versuchsleiter/in alle Informationen gegeben hatte, legte er/sie die einzelnen bebilderten Tafeln den Probanden erneut

vor mit der Bitte, nachzuerzählen, was die einzelnen Bilder darstellen würden und zu beurteilen, wie gut oder böse das war, was da passierte.

– *Bilderordnen:* Hier wurden dem Probanden fünf Bildtafeln (zwei über das Verschulden, eine, die den Schadensumfang darstellte und zwei über den Ersatz) in einer zufälligen Reihenfolge vorgelegt. Ihre Aufgabe war es, diese Bilder in die richtige Reihenfolge zu bringen, die Geschichte zu erzählen und sie dann zu beurteilen. Es folgten zwei weitere Geschichten mit den zwei übrigen Verschulden-Stufen, so daß jede Verschulden-Stufe einmal vorkam.

– *Paarvergleiche:* Im letzten Schritt der Übungsphase wurden die Paarvergleiche geübt. Dazu brachte der Proband die Bilder zweier Geschichten nacheinander in die richtige Reihenfolge, erzählte den Inhalt der Geschichten und wurde dann gefragt, in welcher der beiden das dargestellte Geschehen besser bzw. weniger böse sei, anschließend beurteilte er jede der Geschichten auf der Gut-Böse-Skala.

– *Hauptphase.* Nach insgesamt zwei derartigen Übungsbeispielen folgte die Hauptphase mit der Kombination von insgesamt zweimal 3 Paarvergleichen aller Verschuldensstufen untereinander und den jeweils an die Paarvergleiche anschließenden beiden Schätzurteilen.

– *Erfassung kognitiver Variablen.* Außer dem Brandstiftungsszenario wurden noch mehrere Zusatztests eingesetzt, um wegen der üblichen forensischen Begutachtungsstrategie die Beziehungen der Urteile zur Intelligenz zu prüfen. Die in zwei Listen zusammengestellten Fragen zur Gefahrenerkenntnis und zu Folgen von Regelverletzungen wurden dem Untertest „Allgemeines Verständnis" der Intelligenztests HAWIVA von Eggert (1975), HAWIK von Hardesty und Priester (1966), HAWIK-R von Tewes (1983) sowie dem Untertest „Soziales Erfassen und Sachliches Reflektieren" aus dem AID von Kubinger und Wurst (1985) entnommen. Zusätzlich wurden der Bilderordnen-Test (Untertest „Soziale und Sachliche Folgerichtigkeit") aus dem AID (im Mittel betrug T=46, 56, 53 für die drei Gruppen in der Altersreihenfolge) und der Mosaik-Test aus dem HAWIK-R (im Mittel betrugen die WP=8, 10, 11 für die drei Gruppen in der Altersreihenfolge) durchgeführt. Die Durchführungsreihenfolge war: Fragen zur Gefahrenerkenntnis, Fragen zu Folgen von Regelverletzungen, Bilderordnen-Test, Brandstiftungsszenario, Mosaik-Test.

Probanden. An der Untersuchung nahmen 88 Kinder teil: Eine Gruppe von 13 5jährigen mit einem Mittelwert von 5;6 Jahren (S=4 Monate) und 12 6jährigen mit M=6;5 Jahre (S=3), weiterhin 23 7jährige mit M=7;4 Jahre (S=3) und 40 8–9jährige mit M=8;7 Jahre (S=3). Zwei weitere 5jährige mußten aus der Auswertung ausgeschlossen werden, da sie durch die Aufgaben offensichtlich überfordert waren. Die Kinder stammten aus der Umgebung von Stuttgart und München und wurden einzeln zu Hause untersucht, was rund zwei Stunden pro Proband dauerte. Jedes Kind erhielt zur Belohnung für die Teilnahme 5 DM.

Ergebnisse

Direkte Meßfehlerschätzungen. Die pro Proband über die sechs Zellen des Verschulden-Nachtat-Planes aufsummierte Summe der Abweichungsquadrate wurde gemittelt und als individuelle Meßfehlervarianz für die drei Mittelwertunterschiede zwischen den Verschulden-Stufen verwendet. Die Ergebnisse über die direkten Meßfehlerschätzungen wurden in Tabelle 4 nach dem Alter der Kindergruppe differenziert aufgeführt.

Tabelle 4: Individuelle Meßfehlerschätzungen (Fehlervarianz bei df=6) von drei Altersgruppen

	5–6jährige N=25	7jährige N=23	8–9jährige N=40
Mittlere Fehlervarianz	55.71	27.33	13.10
Minimum	0.17	1.08	0.33
Maximum	180.75	112.67	86.67
Standardabweichung	43.46	30.18	18.22

Wie Tabelle 4 zeigt, bestanden alterskorrelierte Abnahmen im mittleren Meßfehler – schon innerhalb der drei Gruppen, $F(2,85)=18.68$, $P<.001$ –, im Maximum und in der Standardabweichung der individuellen Meßfehler. Darüberhinaus besagten die erheblichen Variabilitäten der individuellen Meßfehlervarianzen, daß eine Schätzung des Meßfehlers durch den indirekten Ansatz über die Gruppenreliabilität zu Fehlbewertungen der individuellen Differenzierungsfähigkeiten führen kann. Eine Bestimmung des Meßfehlers durch Wiederholung der Beurteilungen in einem vom Vorgehen mit den Kindern abweichenden Fragebogen-Format ergab mit 49 Erwachsenen eine nur geringfügig von den 8–9jährigen verschiedene mittlere Fehlervarianz. Der in Tabelle 4 sichtbare Alterstrend in den Fehlervarianzen scheint daher die Meßfehlerveränderung umfassend wiederzugeben.

Tabelle 5: Häufigkeiten von einzelfallstatistisch auf dem 10%-Niveau begründeten Diagnosen über die Differenzierungen der drei Verschulden-Stufen im Schätzurteil von drei Altersgruppen: Brandstiftungen aus Wut (W), aus Fahrlässigkeit (F), aus Versehen (V).

Muster	5–6jährige N=25	7jährige N=23	8–9jährige N=40
W=F>V	13	9	15
W>F=V	4	3	2
W>F>V	0	1	6
W=F=V	7	5	10
W>V,F?	1	3	6
Andere	0	2	1

Anmerkung: W>V,F? bedeutet, daß zwar Wut signifikant anders beurteilt wurde als Versehen, aber die Vergleiche dieser beiden mit Fahrlässig nicht signifikant ausfielen.

Diagnosen der Unterscheidungsfähigkeit. Die drei individuellen Mittelwerte der Verschulden-Stufen wurden paarweise varianzanalytisch in einem F-Test mit df=1 gegen die Meßfehlervarianz mit df=6 geprüft. Dabei ergaben sich die Muster der Tabelle 5 für die auftretenden Unterscheidungsfähigkeiten.

Die hinlänglich deutliche ($p<.10$) Differenzierung der (grob) fahrlässigen Brandstiftung von der versehentlichen (leicht fahrlässigen) gelang in allen drei Gruppen zu ca. 50%, was trotz der intensiveren Instruktion und dem zusätzlichen Paarvergleich in etwa dem Ergebnis von Untersuchung 1 entspricht (vgl. Tabelle 3: dort 46% bei den Minderjährigen). Daß Fahrlässig mit Versehentlich gleichgesetzt wird, nahm ebenfalls ähnlich wie in Tabelle 3 mit dem Alter ab: 16%, 13% und 5%, wenn Brandstiftung Aus Wut von diesen unterschieden werden konnte. Die klare dreifache Abstufung kam dagegen mit zunehmendem Alter häufiger vor: 0% über 4% zu 15%. Mangelnde Differenzierungsleistungen zumindest hinsichtlich Fahrlässig (W=F=V, W>V,F? und Anderes) waren dagegen recht häufig (ca. 40%) in allen drei Gruppen, aber etwas weniger häufig als in Tabelle 3.

Vergleich direkte versus indirekte Meßfehlerschätzung. Entsprechend zu Untersuchung 1 wurden Reliabilitätsschätzungen mit Cronbachs ALPHA vorgenommen. Diese fielen, vermutlich dank der intensiveren Instruktion, höher als in Untersuchung 1 aus: Für die Differenzwertsumme Wut-Fahrlässig .422, .756 für Wut-Versehen und .721 für Fahrlässig-Versehen. Der Vergleich dieser beiden Meßfehlerschätzungen erfolgte mit den Urteilen der Instruktions-, Haupt- und Endphase der Untersuchung, ohne daß Einflüsse der Untersuchungsphase oder der Meßfehlerschätzmethode gefunden werden konnten. Die Übereinstimmung beider Ansätze in den Diagnosen betrug 83% bzw. 89%. Die direkte Methode erschien dabei in den rechtlich relevanten Unterscheidungen als die progressivere.

Paarvergleich versus Schätzurteil: Es gab zwei Möglichkeiten die Zuverlässigkeit der Wahlurteile abzuschätzen. Einerseits lagen zwei Meßwiederholungen von Paarvergleichen vom Ende der Übungsphase und der Hauptphase vor und drei Meßwiederholungen bei sich ändernder Nachtatbedingung. Eine diesbezügliche Auswertung zeigte, daß die Übereinstimmungen dieser Wiederholungen im Vergleich von Wut und Versehen in allen Gruppen hoch waren, während die Paarvergleiche mit Fahrlässigkeit mit dem Alter stabiler wurden. Außerdem konnte innerhalb der Hauptphase die Transivität (Konsistenz) der drei Paarvergleiche pro Nachtatbedingung und ihre Stabilität über die Nachtatbedingung geprüft werden. Das diesbezügliche Auswertungsergebniss besagte, daß die 5–6jährigen zu 60% unzuverlässige Paarvergleiche zeigten, während das für die 7jährigen zu 39% und für die 8–9jährigen nur noch zu 15% galt.

Weiterhin konnte der Paarvergleich aufgrund seiner Stabilität und Konsistenz zur Diagnose der Unterscheidungsfähigkeit benutzt werden. Das wiederum konnte mit dem auf individuellem Meßfehler beruhenden Diagnoseergebnis der Schätzurteile verglichen werden. In der Tabelle 6 ist dieser Vergleich für die Unterscheidung von Fahrlässig und Versehen angegeben. Man sieht erstens die große Übereinstimmung von 45 Diagnosen (66%), zweitens daß relativ selten nur das Schätzurteil signifikant

ausfiel und drittens, daß das Wahlurteil relativ häufig stabil und konsistent war, obwohl das Schätzurteil nicht signifikant ausfiel. Das sprach dafür, daß das auf Stabilität und Konsistenz geprüfte Wahlurteil in der durchgeführten Vorgehensweise eher die Urteilsfähigkeiten zur Schau stellte als die statistische Prüfung der Schätzurteile. Die Häufigkeiten von Tabelle 6 ergaben eine signifikante Tripel-Interaktion von Alter-Schätzurteil-Wahlurteil mit $CHI^2 = 6.61$, $df = 2$, $P < .05$, welche auf dem sich ändernden Vorherrschen des Ortes maximaler Konkordanz (doppelt negativ bei Vorschülern, doppelt positiv bei 8–9jährigen) beruhte.

Tabelle 6: Gekreuzte Häufigkeiten der Diagnose „Fahrlässig wird von Versehentlich unterschieden" beim Wahlurteil und beim Schätzurteil

	Wahlurteil					
	sonst			stabil und konsistent		
Alter:	5–6	7	8–9	5–6	7	8–9
Schätzurteil						
signifikant	1	2	2	5	10	30
sonst	10	2	1	9	9	7

Für die Deliktsfähigkeitsbegutachtung würde das bedeuten, daß man drei Probandengruppen unterscheiden könnte: doppelt gesichert Fähige (diese bei 8–9jährigen relativ häufigen Probanden zeigen sowohl im Wahlurteil als auch im Schätzurteil ihre wertende Unterscheidungsfähigkeit im Unrecht an), fraglich Fähige (bei ihnen spricht nur eine Urteilsform für die Deliktsfähigkeit) und psychometrisch gesehen Unfähige (bei ihnen sprechen beide Urteilsformen trotz intensiver Instruktion gegen die Deliktsfähigkeitsannahme). Mit Alter der Kinder nahmen psychometrische Signifikanz und Wahlstabilität zu, was wiederum die Verwendbarkeit gerade im relevanten Altersbereich nahelegte.

Korrelationen mit Testdaten, Alter und Geschlecht. Von den 24 Produkt-Moment-Korrelationen zwischen den vier kognitiven (Intelligenzwert)-Variablen der Probanden und ihren drei mittleren Differenzwerten bzw. den drei mittleren Urteilen über die Bedingungen Wut, Fahrlässig und Versehen korrelierte gerade eine auf dem 5%-Niveau gesichert: Die Häufigkeit richtig beantworteter Gefahrenitems der Untertests zum Allgemeinen Verständnis mit der Differenz Fahrlässig-Versehen zu $r = .23$ bei $df = 86$, $p < .05$. Keine multiple Regression von diesen vier Intelligenzindizes, Geschlecht und Alter auf irgendeines der sechs genannten Urteilsmaße war statistisch zu sichern. Dagegen waren die vier kognitiven Variablen untereinander korreliert (zwischen $r = .34$ und $.723$) und ebenfalls die Unterscheidung Wut-Versehen mit Fahrlässig-Versehen ($r = .51$).

Die Mittelwerte der beiden Wissenslisten zur Gefahrenerkenntnis und zur Kenntnis der Restitutionspflichten nahmen mit dem Alter zu: $M = 7.1, 10.8$ bzw. 12.9 bei der aus 16 Fragen bestehenden Liste zur Gefahrenerkenntnis ($S = 2.6, 1.7, 1.8$); $M = 5.5, 8.4$ bzw. 10.4 bei der aus 13 Fragen bestehenden Liste zu den Restitutionspflichten ($S = 2.2, 1.6, 1.7$), jeweils für die Gruppen in Altersreihenfolge. Nur wer unter

den gefundenen MINIMAL-Werten seiner Altersgruppe bliebe (MIN=3, 7, 9 für die Gefahrenkenntnisse und MIN=0, 5, 7 für die Restitutionspflicht bei den drei Altersgruppen), wäre wissensmäßig auffällig.

Diskussion

Den pessimistischen Einschätzungen der wenigen sich direkt äußernden forensischen Psychologen ist auf empirischer Basis entgegenzutreten. Ausgehend von der Wertungskomponente der Deliktsfähigkeit gibt es recht zuverlässige Kriterien (vgl. die Ausprägung der Reliabilitätswerte); gibt es einhellige Kriterien, was durch die Übereinstimmung von Wahlurteil und Schätzurteil, die Stabilität innerhalb der drei Phasen von Untersuchung 2 und die Übereinstimmungen zwischen Untersuchung 1 und Untersuchung 2 in Diagnosehäufigkeiten und Diagnosemustern zum Ausdruck kam; und gibt es begrifflich faßbare Kriterien, was im Aufbau des Stimulusmaterials und der Aufgabe des Probanden zu erkennen war. Damit wurde dazu beigetragen, die unangemessene implizite oder explizite psychiatrische Orientierung an der krankheitsbedingten Bewußtseinsstörung (§ 827 BGB), den Rückzug auf den allgemein retardierten Entwicklungsstand, die ausschließliche Verwendung klinischer oder unpräziser Beurteilungsstrategien in der ersatzweise erfolgende Interpretation von Beantwortungen der Frage nach der Einstellung zur Tat oder der Frage nach Wiedergutmachungsabsichten (Dauner, 1980) entbehrlich oder ihrer konzeptionellen Bedeutung gemäß werden zu lassen.

Die zuvor dargestellten Untersuchungen verdeutlichen die systematische Erhebungsplanung als Voraussetzung der Untersuchungstechnik und dokumentieren zugleich die Ergebnisse, die mit dem statistischen Entscheidungsverfahren erlangt wurden. Dies blieb hier in der Durchführung auf das relativ häufig zu begutachtende Delikt der Brandstiftung von Minderjährigen begrenzt. Die untersuchungstechnische Methode ist aber gegebenenfalls auf andere Deliktformen übertragbar (Hommers, 1986a, b; 1988a, b), so daß die tatbezogene Deliktsfähigkeit auch in anderen Fällen hinsichtlich des Werten-Könnens untersucht werden kann.

Die wesentlichen Momente der vorgestellten Erfassungsmethodologie liegen offenbar in der Kombination experimenteller, statistischer und psychometrischer Aspekte. Daher dürfte die generelle, fast verständnisunwillig anmutende Kritik Bressers (1988) an einer statistische oder mathematisierend-experimentelle Methoden anwendenden Rechtspsychologie ungerechtfertigt erscheinen. Vielmehr kann und sollte gerade durch diese Methoden in der Tradition Marbes (1913a, 1913b, 1926) fortgefahren werden, der sein Gutachten zum Müllheimer Eisenbahnunglück vom 17. Juli 1911 eben auf experimentelle und quantifizierende Methoden stützte oder der deskriptiv einzelfall-statistisch vorgehend eine sechsfache Falschaussage von Mädchen gegen ihren Lehrer zusammenbrechen lassen konnte. Im Unterschied zur vorliegenden Arbeit konnte er dabei die hier ins Zentrum der Argumentation gerückte, psychometrisch angelegte Interenzstatistik noch nicht verwenden.

Die untersuchten wertenden Unterscheidungsfähigkeiten von Entschuldigung und Nicht-Entschuldigung (Vergeltungspflichtverständnis), von Versehen und Wut (Unrechtserkenntnis für absichtliche Schädigungen) und von Versehen und Fahrlässig (Unrechtserkenntnis für fahrlässige Schädigungen) erschienen nicht durch Intelligenztestwerte ersetzbar, da sie mit diesen nicht korrelierten. Damit wurde methodisch eine diagnostisch auswertbare Interpretation des Verhältnisses von Werten- und Wissen-Können gegeben, über deren Validität allerdings in forensischer Hinsicht letztlich die mit dem Begutachtungsresultat befaßten Zivilrichter zu entscheiden haben. Das Werten-Können wird offensichtlich auf einer kognitiven Basis erfaßt, die das Wissen-Können enthält. So muß ein Proband die Instruktion verstehen, daran anschließend die verschiedenen Stimulussequenzen verstehen, sie zuerst in die richtige Reihenfolge bringen und den für das Vergleichsurteil entscheidenden Stimulusteil erkennen. Dadurch erst sind die Voraussetzungen für die dann psychometrisch analysierten Wertungen geschaffen. Vom Versuchsleiter kann aber anhand des Verhaltens vor der Urteilsabgabe (z. B. beim Bilderordnen) kontrolliert werden, ob der Proband insofern Hinweise auf Mängel des Wissen-Könnens zeigt. Daher kann das Verfahren der Untersuchung 2 auch in dieser Hinsicht zur Begutachtung verwertet werden. Das Vorgehen der Untersuchung 2 ist besonders deswegen zu empfehlen, weil dann ein total individualisierter und urteils-methodologisch ein sich kontrollierender Maßstab angelegt werden kann, wodurch dem spezifisch forensisch-diagnostischen Leitziel der Validitätsmaximierung und Beweisfähigkeit weitgehend entsprochen wird. Das Vorgehen kann im übrigen ohne weiteres dahin abgeändert werden, daß die Paare der Bildergeschichten sich nicht mehr nur im Verschulden unterscheiden. Jedoch würden sich aufgrund der darin enthaltenen Erschwerung der Urteilsaufgabe vermutlich andere Ergebnisse wie in Tabelle 5 ergeben.

Die Untersuchungsergebnisse sprachen wegen der Alterstrends dafür, daß die Methode gerade dort anwendbar wird, wo es erforderlich werden könnte. Denn gerade im Alter von sieben Jahren reduzierte sich der individuell erfaßte Meßfehler drastisch, so daß er bei 8jährigen das Niveau Erwachsener erreicht hatte. Diese Verfahrensweise wäre außerdem wegen des geringen Anforderungsniveaus in der Äußerungsform sowohl mit Geistigbehinderten, wenn sie im Entwicklungsalter mindestens sieben Jahre alt sind (Hommers, 1989), als auch mit Taubstummen (Übersetzung in die Taubstummensprache durch den Untersuchenden) anwendbar, so daß die frühere explizite Aufnahme der Feststellung der sittlichen Entwicklung durch den Psychiater Ziehen (1923) in die Intelligenz- und Begabungsprüfungen wieder vollends aufgegriffen werden könnte.

Schließlich erscheint es auf dem bereitgestellten methodischen Hintergrund möglich, das Problem der rückbezogenen Verantwortlichkeitsbeurteilung (Wegener, 1981) empirisch zu bearbeiten. Das Problem nimmt Bezug auf ein weiteres der spezifisch forensisch-diagnostischen Leitziele von Hommers (1992), auf das Leitziel des Zustandsmodells: Forensische Diagnostik ist danach auf die Erfassung aktueller oder vergangener Prozeßzustände gerichtet, die sich als Folge von Dispositionen- und Umfeldeinflüssen ergeben. Das Problem entsteht nicht nur dadurch, daß

relativ große Zeitspannen zwischen Delikt und diagnostischer Untersuchung liegen, sondern auch dadurch, daß möglicherweise durch Reaktanz auf die direkten und indirekten Tatfolgen die Einsichtsfähigkeit erst erzeugt wird. Mit einer Längsschnittuntersuchung würde man Anhaltspunkte für die Grundraten individueller Änderungen in vorgegebenen Zeiten und Abschätzungen des Reaktanzeffekts durch die diagnostische Untersuchung gewinnen können. Mit einem gezielten Vergleich von Kindern, die Vorerfahrung mit einer (glimpflich verlaufenen oder auch haftungsrechtlich erheblichen) Brandstiftung haben, und Kindern ohne derartige Vorerfahrung, die zumindest hinsichtlich Intellekt und Geschlecht parallelisiert wurden, könnte das Problem der Reaktanz empirisch erhellt werden.

Bei der solchen Studien vorgreifenden Verwendung des vorgestellten Verfahrens in der forensisch-psychologischen Begutachtung sollte man noch folgendes bedenken:

– Die Wahl des statistischen Irrtumniveaus muß hier bei der zivilrechtlichen Verantwortlichkeit, also nicht überall in der Forensischen Psychologie, progressiv angesetzt werden, weil der Untersuchte einen Vorteil (nicht für den durch ihn verursachten Schaden haften zu müssen) zugesprochen bekommen soll, wenn er die Gesamtdiagnose „Fehlende Einsicht" im Unrechtsbewußtsein oder (was nach den Befunden selten sein wird) im Vergeltungspflichtverständnis erhält. Wo dieses statistische Niveau gesetzt wird, bleibt in der vorliegenden Arbeit eine unbearbeitete Frage. Möglicherweise wird man die statistische Effizienz zur Einschränkung des Beta-Fehlers durch geeignete methodische Maßnahmen erhöhen müssen. Dabei wäre aber der Zeitaufwand als Gegengewicht zu bedenken.

– Natürlich gilt auch hier „Ein Test ist kein Test" als Leitlinie für die gutachterliche Beurteilung des Probanden. Man wird als Meta-Kriterium die Konsistenz der *multimethod*-Untersuchungsergebnisse ansehen müssen, bei denen man auch Überarbeitungen der bei Ziehen (1923) erwähnten Methoden einsetzen kann. Die vorgestellte Verfahrensweise hat sicherlich gegenüber explorativen Methoden den Vorteil der Angabe von Irrtumswahrscheinlichkeiten. Klare Äußerungen eines Probanden sind jedoch deswegen nicht zweitrangig. Forensische Begutachtung ist wie andere Begutachtungen eine verschiedene Informationsquellen integrierende psychodiagnostische Tätigkeit im Sinne von Wegener und Steller (1986) und Steller (1988). Psychometrische Sicherung, wie sie hier vorgestellt wurde, ist nur erforderlich wegen des mit dem Vorgehen verbundenen Meßfehlers. Das hier exemplarisch verfolgte Ziel, psychometrische Einzelfalldiagnostik in der Forensischen Psychologie zu etablieren, setzt also in anderen Bereichen voraus, daß der Meßfehler der dort verwendeten Methoden identifiziert wird.

Literatur

Bresser, P. H. (1972). Die Beurteilung der Jugendlichen und Heranwachsenden im Straf- und Zivilrecht. In H. Göppinger & H. Witter (Hrsg.), *Handbuch der Forensischen Psychiatrie (Band II)*. Berlin: Springer.

Bresser, P. H. (1988). Über die Grenzen psychiatrischer Dokumentation: Was wird nicht abgebildet? *Forensia, 9,* 163–173.

Dauner, I. (1980). *Brandstiftungen durch Kinder. Kriminologische, kinderpsychologische und rechtliche Aspekte.* Bern: Huber.

Eggert, D. (1975). *Hannover-Wechsler-Intelligenztest für das Vorschulalter.* Bern: Huber.

Eisen, G. (1977).*Handwörterbuch der Rechtsmedizin für Sachverständige und Juristen. Band III. Der Täter, sein sozialer Bezug, seine Begutachtung und Behandlung.* Stuttgart: Enke.

Ell, E. (1983). *Wenn Kinder zündeln. Vorschläge zur Feuererziehung.* Tübingen: Katzmann.

Hardesty, F. P. & Priester, H. J. (1966). *Handbuch für den Hamburg-Wechsler-Intelligenztest für Kinder* (3. Auflage). Bern: Huber.

Hommers, W. (1983). *Die Entwicklungspsychologie der Delikts- und Geschäftsfähigkeit.* Göttingen: Hogrefe.

Hommers, W. (1986a). Non-Additivität als Beleg für die moralische Natur der Integration von Schaden und Ersatzleistungen. *Archiv für Psychologie, 138,* 71–90.

Hommers, W. (1986b). Zusammenwirken von Schaden und Ersatzleistung im moralischen Urteil. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 18,* 12–21.

Hommers, W. (1988a). Entschuldigung und Entschädigung für einen Diebstahl. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 20,* 121–133.

Hommers, W. (1988b). Die Wirkungen von Entschuldigung und Entschädigung auf Strafurteile über zwei Schadensarten. *Zeitschrift für Sozialpsychologie, 19,* 139–151.

Hommers, W. (1989). Das Urteil Geistigbehinderter über die Entschuldigung oder die Dritt-Entschädigung für einen „Diebstahl". *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 21,* 53–56.

Hommers, W.(1991). Das „Zündeln" im Urteil: Alterstrends und psychometrische Diagnostizierbarkeit der zivilrechtlichen Verantwortlichkeit nach § 828 BGB. *Zeitschrift für Differentielle und Diagnostische Psychologie, 12,* 163–175.

Hommers, W. (1992). Psychometrische Modelle für die Einzelfalldiagnostik in der Forensischen Psychologie. In L. Montada (Hrsg.), *Bericht über den 38. Kongreß der Deutschen Gesellschaft für Psychologie 1992 in Trier* (S. 12–13). Göttingen: Hogrefe.

Huber, H. P. (1973). *Psychometrische Einzelfalldiagnostik.* Weinheim: Beltz.

Kristof, W. (1983). Klassische Testtheorie und Testkonstruktion. In H. Feger & J. Bredenkamp (Hrsg.), *Messen und Testen* (Enzyklopädie der Psychologie, Band B/I/3) (S. 544–603). Göttingen: Hogrefe.

Kubinger, K. D. & Wurst, E. (1985). *AID. Adaptives Intelligenzdiagnostikum.* Weinheim: Beltz.

Kurtiness, W. & Pimm, J. B. (1983). The moral development scale: A Piagetian measure of moral judgment. *Educational and Psychological Measurement, 43,* 89–105.

Marbe, K. (1913a). Psychologische Gutachten zum Prozeß wegen des Müllheimer Eisenbahnunglücks. *Fortschritte der Psychologie und ihrer Anwendungen, 1,* 339–374.

Marbe, K. (1913b). Kinderaussagen in einem Sittlichkeitsprozeß. *Fortschritte der Psychologie und ihrer Anwendungen, 1,* 375–396.

Marbe, K. (1926). *Der Psycholog als Gerichtsgutachter im Straf- und Zivilprozeß.* Stuttgart: Enke.

Müller, R. (1966). *Sozialer Motivationstest SMT 4–9.* Weinheim: Beltz.

Piaget, J. (1932/1954). *The moral judgment of the child.* London: Routledge & Kegan Paul (deutsche Ausgabe bei Rascher: Zürich, 1954).

Steller, M. (1988). Standards der forensisch-psychologischen Begutachtung. *Monatsschrift für Kriminologie und Strafrechtsreform, 71,* 16–27.

Tewes, U. (1983). *Hamburg-Wechsler–Intelligenztest für Kinder, Revision 1983.* Bern: Huber.

Undeutsch, U. (1967). Delikthaftung junger Menschen. In U. Undeutsch (Hrsg.), *Forensische Psychologie* (S. 567–597). Göttingen: Hogrefe.

Wegener, H. (1981). *Einführung in die Forensische Psychologie.* Darmstadt: Wissenschaftliche Buchgesellschaft.

Wegener, H. & Steller, M. (1986). Psychologische Diagnostik vor Gericht. Methodische und ethische Probleme forensisch-psychologischer Diagnostik. *Zeitschrift für Differentielle und Diagnostische Psychologie, 7,* 103–126.

Wille, R. & Bettge, F. (1971). Empirische Untersuchungen zur Deliktsfähigkeit nach § 828 BGB. *Versicherungsrecht, 37,* 878–882.

Ziehen, Th. (1923). *Die Prinzipien und Methoden der Begabungs-, insbesondere der Intelligenzprüfung bei Gesunden und Kranken. Mit einem Anhang über Prüfung der ethischen Gefühle (5. umgearbeitete Auflage). Berlin: Karger.*