BGH-GUTACHTEN

# Gutachten über Methodik und Bewertungskriterien für Psychologische Glaubwürdigkeitsgutachten

*Klaus Fiedler und Jeannette Schmid*

## 1 Vorbemerkung

Bevor wir – unserem Auftrag gemäß – zur Frage der Methodik von Glaubwürdigkeitsgutachten im allgemeinen und zum vorgegebenen Fragenkatalog im besonderen Stellung nehmen, möchten wir unseren eigenen Beitrag zu dieser Begutachtung in einer Vorbemerkung erläutern.

Unser Gutachten ist aus der Perspektive der wissenschaftlichen Grundlagenforschung verfaßt. Beide Verfasser dieses Gutachtens sind persönlich nicht in der diagnostischen oder forensischen Praxis tätig und somit von Berufs wegen auch nicht selbst mit Fallgutachten befaßt. Vor der Übernahme dieses Gutachten-Auftrages haben wir dies klargestellt und deutlich gemacht, daß unser Beitrag in erster Linie darin liegen wird, allgemeine wissenschaftliche Kriterien zu entwickeln, die an Glaubwürdigkeitsgutachten zu richten sind, und dies durch relevante Forschungsergebnisse zu begründen. Wir begrüßen es sehr, daß der Bundesgerichtshof hiermit die Grundlagenforschung einbezieht und neben der Frage, welche Verfahren unter Praktikern üblich sind, auch der normativen Frage nachgeht, welche Verfahren nach dem heutigen Stand der psychologischen Forschung eigentlich verwendet werden müßten. Tatsächlich werden wir zeigen, daß normative Richtlinien für die Bewertung von Gutachten in erheblichem Maße auf allgemein wissenschaftliche oder gar wissenschaftstheoretische Grundlagen zurückgreifen müssen, die von der Gutachtenpraxis zum Teil weit entfernt und unabhängig sind.

Nach den *Richtlinien für die Erstellung Psychologischer Gutachten (1995)* der Föderation Deutscher Psychologenvereinigungen 1988 haben " ... Personen, die direkt oder indirekt von Psychologischen Gutachten betroffen sind, einen Anspruch auf eine faire, wissenschaftlich fundierte und stets fachkundig angewandte gutachterliche Praxis. Von großer Bedeutung sind dabei die Transparenz und Nachprüfbarkeit der in Gutachten geäußerten Stellungnahmen ... " (Zitat aus dem Vorwort von Prof. Dr. Rudolf Egg und Uwe Wetter). Die zwei zentralen Elemente dieser Verpflichtung sind **Transparenz** und **wissenschaftliche Fundierung** – zwei Aspekte, die auch als Maximen der hier vorgeschlagenen Kriterien gelten können. Wir meinen ferner, daß die letztere Maxime der wissenschaftlichen Fundierung auch die Verpflichtung einschließt, die gutachterliche Praxis laufend im Lichte neuer Forschungser-

gebnisse und zeitgemäßer Methodik zu revidieren und bisher verwendete Verfahren durch wissenschaftliche Begleituntersuchungen zu begründen und zu validieren.

Obwohl wir überzeugt sind, aus theoretischer und methodischer Sicht eine Reihe von klärenden Beiträgen und normativen Empfehlungen bieten zu können, halten wir uns jedoch nicht in allen Punkten des Fragenkataloges für kompetent. Insbesondere gebietet die Tatsache, daß wir in der Psychologie des Kindes- und Jugendalters keine Experten sind, daß wir uns einer substantiellen Stellungnahme zum *Fragenkomplex II., der sich auf spezifisch kindliche und jugendliche Zeugen bezieht*, weitgehend enthalten. Allerdings wird dieser Komplex durch andere Teile unseres Gutachtens ohnehin relativiert bzw. in seiner Relevanz in Frage gestellt. Einige Quellenhinweise finden sich in der Tabelle auf S. 45.

In Format und Stil ist dieser Text so verfaßt, daß er möglichst ohne spezielle Fachkenntnisse nicht nur verstanden, sondern auch kritisch beurteilt werden kann. Wir geben also normative Kriterien wissenschaftlicher Diagnostik nicht nur wider, sondern versuchen gegebenenfalls, die Herleitung bzw. logische oder erkenntnislogische Verankerung dieser Kriterien wenigstens ansatzweise darzustellen. Zu diesem Zweck geben wir gelegentlich konkrete Beispiele zur Illustration für abstrakte Argumente. Unser Kommunikationsziel ist es – sowohl was die Lesbarkeit dieses Textes als auch die vorgeschlagenen Kriterien angeht – den mit der Entscheidung über die Stichhaltigkeit und formale Hinlänglichkeit von Gutachten befaßten Richtern nicht nur spezifische Kriterien anzubieten, sondern sie auch in die Lage zu versetzen, deren Berechtigung kritisch zu beurteilen.

Im ersten Teil des Gutachtens entwickeln wir die Grundlagen für das gesamte Konzept, das wir anbieten möchten. Danach berichten wir empirische Befunde, die unser Konzept stützen. Wir bieten eine Systematik von Kriterien zur Bewertung von Gutachten an und illustrieren diese Kriterien an Beispielen sowie an den aktuellen Strafsachen P. und O.. Abschließend gehen wir auf die einzelnen Fragen des vom BGH vorgegebenen Katalogs ein, indem wir zur Begründung jeweils auf die vorherigen Überlegungen zurückverweisen. Unsere Antworten auf alle Fragen sind zur Übersicht in einer Tabelle zusammengefaßt, die auch Verweise auf die betreffenden Textpassagen enthält.

## 2 Grundlagen des Gutachtens

Als wissenschaftlich tätige Psychologen benutzen wir als Grundlage für unser Gutachten in erster Linie die methodischen, empirischen und erkenntnislogischen Prinzipien der wissenschaftlichen Psychologie, gemäß dem heutigen Stand der Kenntnis. Die Prinzipien, die wir hier zur Bewertung von Gutachten anwenden, sind dieselben, die im allgemeinen auch für die Bewertung der Validität von Forschungsergebnissen und wissenschaftlichen

Erkenntnissen gelten. Sie werden insbesondere dann herangezogen, wenn in qualifizierten Zeitschriften über die Annahme von Forschungsbeiträgen entschieden wird, wobei in der Regel mehrere Expertengutachten eingeholt werden, wodurch die Veröffentlichung von unbegründeten und methodisch unsauberen Befunden zumindest in angesehenen Zeitschriften verhindert wird. Die heutige Psychologie hat in dieser Hinsicht – beispielsweise bei der Vergabe von Forschungsmitteln durch die Deutsche Forschungsgemeinschaft, im Wissenschaftsrat oder der Max-Planck-Gesellschaft – das Ansehen einer recht weit entwickelten Disziplin erlangt, die in ihrer methodischen Stringenz und kritischen Bewertung ihrer Ergebnisse durchaus mit den klassischen Naturwissenschaften verglichen werden kann. Da wir selbst vielfach an diesem inneren Bewertungssystem der Psychologie beteiligt sind – sowohl als Autoren wie auch in der Rolle als Gutachter oder Herausgeber – betrachten wir uns als repräsentative Vertreter des in der heutigen Psychologie gültigen Systems von Regeln.

Neben diesen allgemeinen Grundlagen, die in der Psychologie institutionalisiert sind und keiner speziellen Quellenangabe bedürfen, haben wir eine Reihe von Literaturquellen herangezogen, die in der abschließenden Literaturliste aufgeführt sind. Daneben haben wir Literaturrecherchen in elektronischen Datenbanken betrieben, um keine wichtigen Befunde zu übersehen. Zu einem gewissen Teil beruhen unsere Aussagen auch auf persönlichen empirischen Erfahrungen aus eigenen Gedächtnisexperimenten und Untersuchungen im Rahmen eines gemeinsamen Forschungsprojektes zur Psychologie des Lügens sowie auf konkreten eigenen Erfahrungen mit den dabei anzutreffenden methodischen Problemen.

### 2.1 Welchen erkenntnislogischen Status haben Glaubwürdigkeitsgutachten?

Zur Begründung unserer Argumentation müssen einige Grundbegriffe eingeführt werden, die in die **Wissenschaftstheorie** gehören, also in diejenige Disziplin, die sich mit den Erkenntnisregeln und den grundlegenden Regeln der wissenschaftlichen Methodik befaßt. Wie sich zeigen wird, ist dieser Ansatz keinesfalls akademischer Selbstzweck, sondern führt unmittelbar zu ganz zentralen Grundannahmen über die Logik und Begründung von Gutachten. Wie jede andere wissenschaftliche Systematisierung (Stegmüller, 1969) muß ein Glaubwürdigkeitsgutachten auf einer schlüssigen Argumentation aufgebaut sein, wobei **Schlußfolgerungen** aus einer oder mehreren **Beobachtungen** anhand von **gesetzesartigen Prinzipien** begründet werden müssen. Wenn die Güte oder Validität von wissenschaftlichen Begründungen – so wie in diesem Falle von Gutachten – bewertet oder auch kontrovers ausgehandelt werden, geht es im allgemeinen um Zweifel an folgenden Fragen:

*Reliabilität und Validität der Beobachtungen*: Wie zuverlässig sind die Beobachtungen? Wie genau sind Messungen? Wurden die Beobachtungen oder

Messungen sorgfältig registriert und verbal oder numerisch codiert? Können die Beobachtungen zweifelsfrei interpretiert werden? Sind die Beobachtungen repräsentativ oder einseitig verzerrt, selektiv und unvollständig?

*Gültigkeit und Anwendbarkeit der gesetzesartigen Prinzipien*: Als wie gut bestätigt können die zur Begründung herangezogenen Gesetzmäßigkeiten gelten? Handelt es sich um deterministische Regeln oder um probabilistische Annahmen, die nur mit einer bestimmten Wahrscheinlichkeit zutreffen? Gelten die betreffenden Gesetze universell oder nur innerhalb eines eingegrenzten Geltungsbereichs? Sind die Gesetze überhaupt anwendbar, das heißt, liegen die Beobachtungen innerhalb des Geltungsbereichs der Gesetze?

*Logische Korrektheit der Schlußfolgerungen*: Sofern die Beobachtungen gesichert und die verwendeten Gesetze gültig und anwendbar sind, ist im übrigen die Frage zu bewerten, ob die Ableitung der Schlußfolgerung aus den Beobachtungen mithilfe der Gesetze logisch ohne Widersprüche erfolgt. Diese Forderung ist nur auf den ersten Blick banal. Tatsächlich zeigen denkpsychologische Untersuchungen immer wieder, daß selbst motivierte und formal gebildete Menschen oft außerstande sind, auch nur einfache Regeln systematisch und logisch korrekt anzuwenden (Eddy, 1982; Fiedler & Hertel, 1994; Gigerenzer & Hoffrage, 1995; Wason, 1966). Noch viel schwieriger kann die Bewertung der logischen Korrektheit werden, wenn die betreffenden Gesetze nur probabilistisch sind (so daß die Wahrscheinlichkeitsrechnung herangezogen werden muß) oder wenn Folgerungen aus mehreren Gesetzen kombiniert und integriert werden müssen.

Dieselben Fragen stellen sich sinngemäß, wenn der wissenschaftliche Wert von Gutachten beurteilt werden soll. In Frage stellen und bewerten läßt sich dann ebenfalls, ob im Gutachten hinreichende Maßnahmen ergriffen wurden, um die Zuverlässigkeit und Validität der diagnostischen Beobachtungen und Testergebnisse zu sichern und nachvollziehbar zu dokumentieren, ob die zugrunde gelegten Gesetze oder Annahmen gut bestätigt und überhaupt anwendbar sind und ob dieser entscheidende Sachverhalt ausreichend klargestellt wird, ob die logische und gegebenenfalls mathematische Basis der Schlußfolgerung verstanden und kompetent dargelegt wird. Alle im weiteren entwickelten und auf die aktuell vorliegenden Gutachten angewandten Kriterien beziehen sich auf diese drei Klassen von methodologischen Problemen, also:

- Sicherung von Reliabilität und Validität der diagnostischen Beobachtungen;
- Berechtigung der Anwendbarkeit von Gesetzesannahmen und;
- Optimierung der Schlußfolgerung und Vermeidung logischer Fehler.

Beispielsweise könnte ein Glaubwürdigkeitsgutachten sich auf die Beobachtung beziehen, daß ein Zeuge sehr *detaillierte Angaben* zum Tathergang macht, eine Gesetzesannahme heranziehen, wonach sehr detaillierte Angaben

nur von Personen gegeben werden können, die reale Gegebenheiten berichten, und daraus den Schluß ziehen, daß der Zeuge die Wahrheit sagt. Die Bewertung des Gutachtens müßte sich dann mit den drei Fragen befassen, ob die Beobachtungen angezweifelt werden können (Waren die Angaben wirklich detailliert, gemessen an einem quantitativen linguistischen Standard?), ob das herangezogene Gesetz als empirisch gut bestätigt gelten kann und wo dies nachgewiesen wird (Gibt es gezielte Untersuchungsergebnisse hierzu, die heutigen Standards genügen? Gibt es vielleicht andere Gesetze, die das Gegenteil implizieren?) und, ob Probleme des logischen Schließens bedacht wurden (Wenn bekannt ist, daß wahre Aussagen oft detailliert sind, darf man den Umkehrschluß ziehen, daß detaillierte Aussagen wahr sind? Wie werden die Implikationen verschiedener Gesetze "verrechnet"?).

Angemerkt sei an dieser Stelle nur, daß hier auf Sachverständige im Grunde dieselben erkenntnislogischen Kriterien angewendet werden müssen wie auf Zeugen. Auch dann, wenn die Aussagen eines Augenzeugen bewertet werden, geht es im wesentlichen um die Verläßlichkeit seiner Beobachtungen (Ist der Augenzeuge sehtüchtig?), um die Relevanz von Gesetzen (Inwiefern beeinflußt Sehtüchtigkeit die Wahrnehmung des betreffenden Ereignisses?) und um die Logik der Folgerung (Wie werden reduzierte Sehtüchtigkeit und erhöhte Aufmerksamkeit zu einer Gesamtfolgerung kombiniert?).

### 2.2 Zwei Formen wissenschaftlicher Begründung

Betrachtet man nun Glaubwürdigkeitsgutachten im besonderen im Vergleich zu anderen wissenschaftlich fundierten Begründungen (z.B. Erklärung und Vorhersage von empirischen Forschungsresultaten), so muß eine pragmatisch sehr bedeutsame Unterscheidung zwischen zwei verschiedenen Formen der Begründung eingeführt werden. Diese Unterscheidung ist unseres Erachtens von zentraler Bedeutung für das Verständnis der Frage, welche der genannten drei Kriterienklassen speziell bei Glaubwürdigkeitsgutachten optimiert werden können. Es geht um die Unterscheidung von **deduktiv-nomologischen Beweisen** einerseits und **induktiv-statistischen Schlüssen** andererseits. (Wir gebrauchen diese Begriffe wie im folgenden definiert). Der Unterschied liegt nicht in der grundlegenden Form – beide Schlußformen beruhen auf denselben drei Konstituenten: Beobachtungen, Gesetze und Schlußfolgerung. Vielmehr liegt der pragmatische Unterschied in dem relativen Gewicht, welches den genannten drei Teilproblemen zukommt.

#### 2.2.1 Deduktiv-nomologischer Beweis

Ein deduktiv-nomologischer Beweis kommt, wie der Name besagt, durch eine deduktive Schlußfolgerung zustande, wobei eine Beobachtung (oder mehrere) unter eine allgemeingültige ("nomologische") Gesetzesaussage subsumiert wird, so daß die Schlußfolgerung bewußt und kontrolliert vollzogen werden kann. Die pragmatische Annahme hierbei lautet, daß die Gesetzesannahme wissenschaftlich so gut bestätigt ist, daß man sie mechanistisch an-

wenden kann. Da das Gesetz als gesichert und valide gilt, kommen in solchen Beweisen typischerweise singuläre Gesetze vor anstatt komplexer Gefüge aus mehreren Gesetzen. Eine oftmals stillschweigend mitgedachte Zusatzannahme besagt außerdem, daß die Gesetzesaussage nicht nur für sich gültig, sondern auch erschöpfend ist, das heißt, alle relevanten Faktoren für die Herleitung der Schlußfolgerung erfaßt, so daß das Gesetz auf einzelne Fälle unter variablen Bedingungen generalisiert werden kann. Mit anderen Worten, das Gesetz wird als notwendig und hinreichend zur Begründung eines Sachverhalts angesehen.

Wesentlich ist in jedem Falle die Einsicht, daß derartige Begründungen vom deduktiv-nomologischen Typ ganz entscheidend von der wissenschaftlichen Bestätigung der zentralen, explizit zu formulierenden Gesetzesannahmen abhängen. Wichtig für die Bewertung ist allerdings auch die Verläßlichkeit der Beobachtungen, aber dies steht außer Frage. Daneben liegt jedoch das Hauptgewicht der Bewertung eines deduktiv-nomologisch begründeten Gutachtens auf dem Nachweis der vorhandenen Evidenz für die verwendete Gesetzesaussage. Auch die logische Schlußform ist in der Regel nicht sehr problematisch, da zumeist einfache Gesetze zur Anwendung kommen.

Übertragen auf das obige Beispiel könnte ein deduktiv-nomologischer Beweis der Glaubwürdigkeit so aussehen:

| | |
|---|---|
| Beobachtung: | Zeuge gibt detailreiche Schilderung |
| Gesetz: | Wenn Schilderungen detailreich sind, entsprechen sie der Wirklichkeit |
| Schlußfolgerung: | Der Zeuge sagt die Wahrheit. |

Der logische Schluß birgt keine nennenswerten Probleme in sich. Die Beobachtung selbst sollte mit den gängigen diagnostischen Mitteln abzusichern sein (vorausgesetzt, die nötigen Methoden werden eingesetzt). Problematisch ist hier vor allem die Gesetzesannahme, mit deren Gültigkeit die gesamte Argumentation steht oder fällt. Ein so konzipiertes Gutachten stellt folglich enorm hohe Anforderungen an die Sicherung der Gesetzesaussage; in diesem Falle müßte wirklich stichhaltige empirische Evidenz für die Gültigkeit und für die hinreichende Generalität der Annahme nachgewiesen werden, daß Detailreichtum wahre Aussagen anzeigt.

### 2.2.2 Induktiv-statistischer Schluß

Betrachten wir im Vergleich dazu das durchaus verschiedene Prinzip des induktiv-statistischen Schließens. Anstelle der anspruchsvollen aber oft unrealistischen Annahme, daß gut bestätigte, universell anwendbare Gesetze existieren, aus denen gutachterliche Schlußfolgerungen logisch einfach abgeleitet

werden können, werden viele "Mini-Gesetze" herangezogen – im folgenden Indikatoren genannt – die für sich genommen alle nur von bescheidener Aussagekraft sind, obwohl sie im Erwartungswert (Durchschnitt) besser als der Zufall sein müssen. Entsprechend werden für die gleichzeitige Anwendung vieler solcher Indikatoren auch viele Beobachtungen benötigt:

| | |
|---|---|
| Beobachtungen: | B1: Aussage enthält viele räumliche Details |
| | B2: Aussage enthält **nicht** viele zeitliche Details |
| | B3: Aussage enthält viele soziale Details |
| | B4: Aussage enthält viele emotionale Details |
| | B5: Aussage enthält viele physikalische Details |
| | etc. |
| Gesetze: | G1: p(wahr/viele räumliche Details) > 50% |
| | G2: p(wahr/viele zeitliche Details) > 50% |
| | G3: p(wahr/viele soziale Details) > 50% |
| | G4: p(wahr/viele emotionale Details) > 50% |
| | G5: p(wahr/viele physikalische Details) > 50% |
| | etc. |
| Schlußfolgerung: | p(Aussage ist wahr) >>> 50% |

Anstatt sich auf eine gut bestätigte Gesetzmäßigkeit auf übergeordneter Ebene zu verlassen, werden viele schwach bestätigte Indikatoren auf unterer Ebene benutzt. Dabei kann die Menge der Indikatoren durchaus solche Kennzeichen erfassen, die andernorts wie nomologische Gesetze behandelt wurden. Das heißt, Gesetze und Indikatoren sind nicht essentiell verschieden; der Unterschied ergibt sich allein aus ihrer Einbettung in eine der beiden erkenntnislogischen Argumentformen. Trotz der bescheidenen Validität der einzelnen Indikatoren, kann die aus der Gesamtheit aller Indikatoren abgeleitete Schlußfolgerung eines Gutachtens jedoch einen beträchtlichen diagnostischen Wert erreichen, der weit höher liegt als die Gültigkeit der einzelnen schwachen Indikatoren. Logisch und mathematisch läßt sich dieser glückliche Umstand durch ein Prinzip begründen, das in seiner Stärke und Bedeutung dem gesunden Menschenverstand nicht unbedingt zugänglich ist, das aber die Grundlage für viele induktive Schlüsse bildet: das Prinzip der **Aggregation** (Rosenthal, 1987).

Da die Fehleranteile der einzelnen imperfekten Gesetze per definitionem unkorreliert (d.h., statistisch unabhängig) sind, die systematisch verläßlichen Anteile jedoch eine Gemeinsamkeit aufweisen (i.e., die zu erschließende Größe; in diesem Falle: die tatsächliche Wahrheit der Aussage) werden durch Aggregation die systematischen Anteile verstärkt, während sich die Fehler-

anteile der verschiedenen Elemente gegenseitig herauskürzen. Im obigen Beispiel wird sogar eine Beobachtung (B2: *Fehlen von zeitlichen Details*), die *gegen* eine wahre Aussage spricht, durch eine Mehrzahl von gegenläufigen Beobachtungen (B1, B3, B4 etc.) weggekürzt, so daß im Aggregat eine relativ klare Schlußfolgerung zugunsten der Wahrheit entsteht.

Gerade dann, wenn die elementaren Beobachtungen / Indikatoren von sehr begrenztem Wert sind (d.h. im Durchschnitt nur knapp über dem Zufall liegen), wirkt sich Aggregation besonders stark aus. So kann man mithilfe der Formel von Spearman und Brown (nach Walker & Lev, 1953) vorhersagen, daß bei einer durchschnittlichen Korrelation pro Einzelgesetz von nur r = 0.15 zwischen Beobachtungen und der zu erschließenden Größe durch Aggregation über 10 Beobachtungen eine Vorhersagbarkeit von r = 0.64 entsteht. Durch Aggregation über 20 Beobachtungen steigt die Korrelation als Maß für die Stärke der Gesetzmäßigkeit auf r = 0.78 (Korrelation ist ein statistisches Zusammenhangsmaß, das den maximalen Wert r=1 annimmt, wenn zwei Variablen perfekt zusammenhängen, und r=0, wenn zwei Variablen völlig unabhängig sind). Dieses hier sehr einfach erklärte Prinzip der Aggregation ist mathematisch und psychometrisch sehr eingehend untersucht und bildet die rationale Grundlage für viele diagnostische Verfahren, insbesondere auch etablierte Testverfahren.

Man könnte hieraus folgern, daß sich das schwerwiegende Problem des Nachweises von gut bestätigten Gesetzmäßigkeiten, welches deduktiv-nomologische Beweise so sehr erschwert, beim induktiv-statistischen Schließen durch einen einfachen Umformungstrick umgehen läßt: Scheinbar muß man ein globales Gesetz nur in viele kleine Komponenten zerlegen (dann als "Indikatoren" oder "Cues" umbenannt), an die dann keine großen Ansprüche zu stellen sind. Diese Darstellung ist jedoch irreführend. Tatsächlich sind induktiv-statistische Schlüsse keineswegs unproblematisch. Durch die Verteilung der Beweislast über viele kleine realistische Gesetzmäßigkeiten (Indikatoren) wird die Problematik lediglich verschoben. Während ein deduktiv-nomologischer Beweis in erster Linie ein gut bestätigtes Gesetz erfordert und dann angezweifelt werden kann, wenn das globale Gesetz nicht gesichert ist, hängen induktiv-statistische Schlüsse entscheidend von der adäquaten Auswahl und der ausreichenden Anzahl der betrachteten Indikatoren bzw. Beobachtungen ab. Mit anderen Worten, das Problem bei einer induktiv-statistisch begründeten Beweisführung liegt in der **Selektivität** und ausreichenden **Aggregation bzw. Kombination der elementaren Informationen**, nicht im (unrealistisch) hoch angenommenen Bestätigungsgrad eines einzelnen Gesetzes.

Induktiv-statistische Schlüsse profitieren zwar von der "eingebauten Tugend" der Aggregation (Herauskürzen von fehlerhafter Information), die auch mathematisch kontrollierbar und objektivierbar ist (Fiedler, 1996; Rosenthal,

1987). Sie können jedoch durch einseitige Informationssuche und einseitiges Hypothesentesten extrem fehlgeleitet sein. Indikatoren haben keine konstante, essentielle Bedeutung, sondern lediglich pragmatischen Wert. Dieselben Indikatoren können zur Diagnostik verschiedener Sachverhalte dienen; im Grund kann derselbe Indikator – je nach Kontext – sowohl als Indikator der Wahrheit wie als Indikator der Unwahrheit nützlich sein (Johnson, Bush & Mitchell, 1998). Um die willkürliche, ungerechtfertigte Verwendung von Indikatoren (als ob es sich um universelle Gesetze handelte) auszuschließen, kommt es daher entscheidend darauf an, das diagnostische Entscheidungsmodell zu explizieren (Cronbach & Gleser, 1965), innerhalb dessen die Indikatoren erst eine Bedeutung erhalten (z.B. Detailreichtum als Indikator tatsächlichen Erlebens versus Detailreichtum als Indikator von sprachlicher Raffinesse bei der Falschaussage). Die Verwendung von statistischen Indikatoren erfordert also notwendigerweise die Formulierung expliziter Modellannahmen oder Hypothesen (tatsächliches Erleben; raffinierte Sprache). Die Entscheidung zugunsten einer bestimmten Modellannahme (der Schluß von Detailreichtum auf tatsächliches Erleben) und der Ausschluß von alternativen Modellannahmen (der Schluß auf raffiniertes Lügen) muß grundsätzlich begründet und durch geeignete Methoden überprüft werden (siehe unten).

Als wesentliches Ergebnis bleibt somit festzuhalten, daß eine unabdingbare Voraussetzung für induktiv-statistischeSchlüsse aufgrund von Glaubwürdigkeits-Indikatoren die **Explikation und aktive Prüfung des zugrunde gelegten diagnostischen Modells** ist. Nur wenn eine gewählte Modellannahme hinreichend gesichert ist, so daß zumindest die diagnostische Richtung der Indikatoren bestimmbar ist, kann durch Aggregation über viele (einzeln schwache) Indikatoren ein Aggregationseffekt erwartet werden. Dieser Aggregationseffekt kann dann zu einer beträchtlichen Genauigkeit der diagnostischen Entscheidung führen.

Ein zweites Problem des induktiv-statistischen Schließens, neben der Explikation und Überprüfung des Modells, liegt in der **Selektivität der Indikatoren**. Eben weil die diagnostische Information über viele kleine Gesetzmäßigkeiten verteilt ist, kommt der repräsentativen Auswahl der Beobachtungen entscheidendes Gewicht zu. Zahlreiche psychologische Befunde zum induktiven Hypothesentesten zeigen, daß massive Fehlentscheidungen entstehen, wenn selektiv nur bestimmte Hypothesen betrachtet werden, während andere einfach außer Acht gelassen werden (Snyder, 1984; Tversky & Kahneman, 1974; Zuckerman, Knee, Hodgins & Miyake, 1995). Dies gilt für alltägliche Urteile und Entscheidungen ebenso wie für wissenschaftliche Erkenntnisse. Wenn die zu prüfende Hypothese lautet, daß Theorie A richtig ist, dann führt die selektive Auswahl von Beobachtungen sehr oft dazu, daß bestätigende Evidenz für die gewählte Hypothese A gefunden wird. Faßt man dagegen eine alternative Hypothese B ins Auge, die mit Hypothese A unvereinbar ist, dann wird über denselben selektiven Mechanismus oft Bestätigung für B gefun-

den. Hierfür gibt es zahlreiche und vielfältige Evidenz in der psychologischen Literatur. Die Kontrolle dieser systematischen Verzerrungen beim induktiven Schließen erfordert die systematische Suche von Beobachtungen für **alternative, kontrastierende Hypothesen** (Klayman & Ha, 1987).

Bezogen auf das oben verwendete Beispiel bedeutet dies etwa, daß die Suche nach Indikatoren von Detailreichtum (räumliche, zeitliche, soziale, emotionale, physikalische Details) eine einseitige Suche nach *Indikatoren der Wahrheit* einer Aussage darstellt. Je mehr Beobachtungen Detailreichtum anzeigen, um so mehr wird die Schlußfolgerung gestützt, daß die Aussage wahr ist. Alternativ könnte man nach *Indikatoren von Widersprüchlichkeit* suchen; mit wachsender Aggregation von derartigen Beobachtungen würde die umgekehrte Folgerung gestützt, daß die Aussage unwahr ist. Denn die betreffenden Gesetzesannahmen besagen vermutlich, daß Zeichen von Widersprüchlichkeit auf Unwahrheit schließen lassen. Ob dieselbe Aussage als wahr oder unwahr klassifiziert wird, sollte somit entscheidend davon abhängen, wie lange und gründlich nach Indikatoren von Detailreichtum einerseits und Widersprüchlichkeit andererseits gesucht wird.

Das Problem der Selektivität von Indikatoren (z.B. Auswahl von Symptomen im Verhalten oder von Interviewfragen) ist in der Umfrage- und Interviewforschung wohl bekannt (Blau & Katerberg, 1982; Ray, 1983; Semin, Rubini & Fiedler, 1995; Zuckerman et. al., 1995). Es kann die Ergebnisse von Untersuchungen massiv verfälschen. Wie schon erwähnt können die Ergebnisse diagnostischer Untersuchungen ganz entscheidend von den Fragen oder Indikatoren determiniert werden, die der Untersucher selbst auswählt bzw. focussiert. Oft existieren alternative oder gegensätzliche Modellannahmen, die andere Indikatoren nahelegen und so zu abweichenden Ergebnissen führen. Deshalb ist es im Rahmen von induktiv-statistischen Schlüssen unerläßlich zu prüfen, ob kontrastierende Hypothesen existieren und zu gegensätzlichen Schlüssen führen.

### 2.3 Welche Form haben gutachterliche Begründungen?

Aus diesen Vorüberlegungen sollte deutlich geworden sein, daß die Kriterien für die Bewertung von Gutachten davon abhängen, ob die gutachterlichen Schlußfolgerungen vom Typ eines deduktiv-nomologischen Beweises sind oder aber vom Typ eines induktiv-statistischen Schlusses. Im ersten Falle steht der Bestätigungsgrad eines globalen Gesetzes im Vordergrund der Bewertung, im letzteren Fall ergeben sich mögliche Probleme aus der Selektion der Indikatoren und den Gefahren einseitigen Hypothesentestens. Die wissenschaftliche Absicherung eines Gutachtens erfordert daher unterschiedliche Methoden und Maßnahmen:

- Wenn eine gutachterliche Argumentation auf einem deduktiv-nomologischen Beweis beruht, dann steht und fällt die Bewertung des Gutachtens mit dem Nachweis überzeugender Evidenz für

einzelne globale Gesetze, die universell und mechanistisch anwendbar sind.

- Wenn ein Gutachten jedoch auf der induktiven Aggregation über viele schwache Gesetze basiert, dann müssen Probleme der Selektivität und der nicht ausreichenden Aggregation kontrolliert und bewertet werden.
- Beiden Argumentationsarten gemeinsam ist die Notwendigkeit, die Verläßlichkeit der Beobachtungen und die Logik der Schlußfolgerung abzusichern.

Grundsätzlich können beide Varianten der wissenschaftlichen Begründung in der Glaubwürdigkeitsbegutachtung vorkommen. Ein Beispiel für ein deduktiv-nomologisches Argument, das in Gutachten enthalten sein kann, läge dann vor, wenn die Aussage eines Zeugen aus a-priorischen (z.B. logischen) Gründen gar nicht wahr sein kann (z.B. weil der Zeuge aus seiner räumlichen Perspektive etwas gar nicht gesehen haben kann). Ein weniger selbstverständlicher Fall läge vor, wenn der Zeuge äußerst spezifische Details über den Tathergang benennen kann, die objektiv bekannt sind aber absolut geheim gehalten wurden. Solch eindeutiges Tatwissen läßt zumindest den Schluß zu, daß der Zeuge tatsächlich anwesend gewesen sein muß (vgl. Tatwissenstest, Elaad, 1990).
Dennoch gehen wir in unseren weiteren Überlegungen davon aus, daß deduktiv-nomologische Beweise in psychologischen Glaubwürdigkeitsgutachten nur ganz ausnahmsweise eine Rolle spielen. Ferner dürfte in solchen Fällen ein psychologisches Gutachten meist nicht erforderlich sein, weil der Schluß auf eine Falschaussage offensichtlich ist. In den meisten Fällen ist somit die Annahme kaum gerechtfertigt, daß ein so gut bestätigtes Gesetz bekannt und anwendbar ist, daß die Wahrheit der Zeugenaussage wirklich deduziert werden kann. Im Normalfall werden Gutachten hingegen die erkenntnislogische Form eines induktiv-statistischen Schlusses annehmen. Daraus ergeben sich die oben skizzierten spezifischen Konsequenzen für die Frage nach den Kriterien eines wissenschaftlich angemessenen Gutachtens und die möglichen Fehlerquellen, denen Gutachten ausgesetzt sind.

Nach dieser Einführung in die Form wissenschaftlicher Begründungen werden im nächsten Abschnitt psychologische Befunde dargestellt, die den hier eingenommenen Standpunkt unterstützen. Zum einen wird gezeigt, daß die für die Aussagendiagnostik oft herangezogenen Gesetze kaum als gut bestätigte Gesetze gelten dürfen, die einen deduktiv-nomologischen Beweis begründen würden. Zum anderen wird gezeigt, daß induktiv-statistische Schlüsse auf die Glaubwürdigkeit auf der Grundlage von schwachen aber multiplen Gesetzesannahmen (multiple-cue inferences, Ambady & Rosenthal, 1992) in der Regel überzufällig genau sind und daher die Voraussetzungen für das Aggregationsprinzip erfüllen. Eine wichtige Voraussetzung für induktiv-statistisches Schließen ist indessen die repräsentative, nicht selek-

tierte Auswahl der Indikatoren. Werden die Indikatoren durch einseitiges Hypothesentesten in selektiver Weise verzerrt, so können induktive Schlüsse zu schwerwiegenden Fehlertendenzen führen. Auch hierfür werden empirische Untersuchungen angeführt. Welche Kriterien sich daraus für die Bewertung von Glaubwürdigkeitsgutachten ergeben, wird dann Gegenstand des übernächsten Abschnittes sein.

## 3 Empirische Evidenz

Verschiedene in der forensischen Praxis als Gutachter tätige Psychologen haben den Versuch unternommen, eine Systematik von Aussagemerkmalen zu entwickeln, die wahre von falschen Aussagen trennen (Arntzen, 1982; Szewczyk, 1973; Undeutsch, 1989). Die von verschiedenen Wissenschaftlern und Gutachtern empfohlenen und verwendeten Aussagemerkmale sind teilweise unterschiedlich, überlappen jedoch in der Regel sehr stark. Eine Integration und Kondensation der vorliegenden Kriterienkataloge, wie sie von Steller und Köhnken (1989) sowie Steller, Wellershaus und Wolf (1992) vorgeschlagen wird, erscheint in der nachfolgenden Tabelle.

**Realkennzeichen in Zeugenaussagen**
nach Steller, Wellershaus & Wolf (1992)

**Allgemeine Merkmale**

1. Logische Konsistenz
2. Unstrukturierte Darstellung
3. Quantitativer Detailreichtum

**Spezielle Inhalte**

4. Raum-zeitliche Verknüpfungen
5. Interaktionsschilderungen
6. Wiedergabe von Gesprächen
7. Schilderung von Komplikationen im Handlungsverlauf

**Inhaltliche Besonderheiten**

8. Schilderung ausgefallener Einzelheiten
9. Schilderung nebensächlicher Einzelheiten
10. Phänomengemäße Schilderung unverstandener Handlungselemente
11. Indirekt handlungsbezogene Schilderungen
12. Schilderung eigener psychischer Vorgäng
13. Schilderung psychischer Vorgänge des Täters

**Motivationsbezogene Inhalte**

14. Spontane Verbesserungen der eigenen Aussage
15. Eingeständnisse von Erinnerungslücken
16. Einwände gegen die Richtigkeit der eigenen Aussage
17. Selbstbelastungen
18. Entlastungen des Angeschuldigten

**Deliktspezifische Inhalte**

19. Deliktsspezifische Aussagenelemente

---

Alle 19 Merkmale bzw. Kriterien sind ihrer verbalen Ausrichtung nach als *Realkennzeichen* konzipiert; das heißt, das Auftreten dieser Kennzeichen in einer Aussage gilt als Hinweis auf die Wahrheit bzw. Glaubhaftigkeit der Aussage. Das Fehlen der Merkmale erhöht demgemäß die Wahrscheinlichkeit der umgekehrten diagnostischen Entscheidung als unwahr bzw. unglaubwürdig. Die in der Tabelle wiedergegebene Liste soll hier pars pro toto für eine größere Menge ähnlicher Realitätskriterien oder Indikatoren stehen, für die jedoch die folgenden Überlegungen in analoger Weise zutreffen.

Mehrere Praktiker und Forscher haben Untersuchungen durchgeführt und publiziert, in denen der Wert dieser Aussagemerkmale für die Diagnostik der Glaubwürdigkeit empirisch gemessen und überprüft wurde (Anson, Golding & Gully, 1993; Dahle, 1997; Köhnken & Wegener, 1982; Krahé & Kundrotas, 1992; Sporer & Küpper, 1995; Steller & Köhnken, 1989). Eine Bestandsaufnahme und kritische Bewertung dieser empirischen Erkenntnisse läßt nach unserer Überzeugung keinen anderen Schluß zu als, **daß die von fachlich ausgewiesenen Psychologen empfohlenen und von Praktikern benutzten Aussagemerkmale auf keinen Fall den Status von nomologischen Gesetzen beanspruchen dürfen. Andererseits scheinen diese Merkmale durchaus geeignet zu sein, als nützliche Indikatoren im Rahmen klar spezifizierter und kritisch geprüfter Modelle einen bedeutsamen statistischen Beitrag zur Wahrheitsfindung zu leisten.**

Diese Gesamtwertung stellt das Fazit aus einer Reihe von Beobachtungen dar:

### 3.1 Evidenz gegen die Verwendung als nomologische Gesetze

(1) Eine Verabsolutierung oder Etablierung dieser Merkmale als allgemeingültige Gesetze der Glaubwürdigkeitsdiagnostik allein aufgrund der Meinung von Experten, ohne kritische und empirische Prüfung, ist mit der Forderung nach wissenschaftlicher Fundierung nicht vereinbar. Dies folgt aus der fehlenden empirischen Evidenz.

(2) Meta-Analysen der Aussagemerkmale (Rosenthal, 1978) im Sinne der

obigen Tabelle, in der die insgesamt vorhandene Evidenz aus allen existierenden psychologischen Untersuchungen gewichtet und zusammengefaßt wird, wurden bisher nicht durchgeführt. Nach heutigen methodologischen Standards sind Meta-Analysen eine notwendige Mindestanforderung für die Bestätigung bzw. für die quantitative Bewertung von Gesetzen auf empirischem Wege.

(3) Kein einzelnes der sogenannten Realkennzeichen erreicht in irgendeiner Untersuchung für sich genommen eine quantitative Verläßlichkeit, welche es rechtfertigen würde, von einer nomologisch gesetzesartigen Beziehung zu sprechen, die sich auf Einzelfälle in spezifischen Kontexten generalisieren ließe. Typische Ergebnisse sind die folgenden:

In einer Feldstudie von Szewczyk und Littmann (1989) zeigte sich, daß die meisten von 12 verwendeten Kennzeichen eher in wahren Aussagen vorkamen. Allerdings wiesen zwei Kennzeichen (*ausschließliche Detailliertheit bei der Schilderung der Rahmenhandlung; global-vage Tatschilderung*) statistisch signifikant in die verkehrte Richtung, kamen also häufiger bei unwahren als bei wahren Aussagen vor.

Bei Bender (1987) wurden falsche Aussagen von Meineidigen mit wahren Aussagen von Zeugen verglichen, wobei die Klassifikation der Aussagen als wahr oder falsch als einigermaßen gesichert gelten kann. Bei insgesamt 4 von 10 verwendeten Einzelmerkmalen konnten die erwarteten Unterschiede nachgewiesen werden.

In einer Untersuchung von Krahé und Kundrotas (1992) mit Fallmaterial aus authentischen Vernehmungsprotokollen nach Vergewaltigungsanzeigen wurden eingestandene Falschaussagen mit anhand von Geständnissen fremder Täter als wahr klassifizierten Aussagen verglichen. (Die Problematik von Geständnissen besteht zwar, ist hier aber stark vermindert). Die diagnostische Grundlage bildeten die 19 von Steller und Köhnken (1989) vorgeschlagenen Realkennzeichen (siehe obige Tabelle). Nur drei der 19 Merkmale unterschieden signifikant zwischen wahren und falschen Aussagen, davon eines in der verkehrten Richtung. Ein ordinaler Vergleich (unabhängig von statistischer Signifikanz) zeigt eine Inversion der erwarteten Richtung (also häufigeres Auftreten bei falschen statt wahren Aussagen) in nicht weniger als zehn Fällen.

Nur sieben Inversionen bei 26 Kennzeichen findet Dahle (1997), aber dafür auch sehr bescheidene diagnostische Werte. Ganz analoge Befunde gelten interessanterweise auch für andere Listen von verbalen und nonverbalen Kennzeichen (Körpersprache, Stimme, Mimik etc.), die man zur als Diagnostica der Wahrheit versus Lüge in der psychologischen Forschung untersucht hat (z.B. Meta-Analyse von Zuckerman, DePaulo & Rosenthal, 1981).

Unabhängig davon, daß die verschiedenen empirischen Arbeiten nicht exakt gleiche Befunde liefern, scheinen die hier referierten ausreichend für die Demonstration, daß die Annahme der Gültigkeit dieser und ähnlicher Realkennzeichen – im Sinne nomologischer Gesetze – nicht berechtigt wäre und vor allem nicht generalisiert werden kann.

(4) Neben den als "Realkennzeichen" häufig zur Diagnostik der Glaubwürdigkeit herangezogenen verbalen Aussagemerkmalen erfüllen auch keine anderen bekannten Diagnostica das Kriterium von gut bestätigten empirischen Gesetzen, die im Sinne eines deduktiv-nomologischen Beweises Verwendung finden könnten. Dies gilt insbesondere für non-verbale Indikatoren (Zuckerman et al., 1981) und für die Messung von emotionalen und expressiven Indikatoren (Fiedler, 1999).

Für das Fehlen von universellen (i.e., mechanisch einsetzbaren) Gesetzen im Bereich der Wahrheitsdiagnostik gibt es mindestens zwei zwingende Gründe. Zum einen ist das Abweichen einer Aussage von der Wahrheit **kein einheitliches Phänomen**, sondern eine Sammelkategorie von vielerlei psychologischen Prozessen: Fehlerhafte Wahrnehmung eines Zeugen von Anfang an; Vergessen; konstruktive Gedächtnisverzerrung aufgrund der Konfusion mit anderem Weltwissen; nachträgliche Beeinflussung des Gedächtnisses durch Befragung und soziale Suggestion; bewußte Täuschungsabsicht; unbewußte motivierte Täuschung; Ratetendenz bei Urteilen unter Unsicherheit; Konfabulieren und andere imaginative Tendenzen bis hin zu pathologischem Realitätsverlust oder Halluzinationen; und andere. Da diese verschiedenen Quellen und Ursachen von Falschaussagen ihrem Wesen nach völlig verschieden sind, wäre es ungerechtfertigt und geradezu fahrlässig, eine invariante Gesetzmäßigkeit anzunehmen und bei jeder Art von Aussagen für die diagnostische Entscheidung zugrunde zu legen – ohne begründete Annahme von expliziten Modellen. Ein und dasselbe diagnostische Zeichen (z.B. Detailreichtum) kann mit Bezug auf ein Modell (ad-hoc produzierter Bericht ohne Vorbereitung) ein Indiz für eine wahre Aussage sein, während es innerhalb eines anderen Modells (phantasiereiche Konfabulation eines Kindes) normal ist und in einem dritten Modell (raffiniert vorbereiteter Täuschungsversuch) sogar ein Indiz für Unwahrheit sein könnte.

Der zweite a-priori-Grund für das Fehlen echter Gesetze liegt in der kaum vorhandenen Möglichkeit, die Gültigkeit solcher Gesetze empirisch zu validieren. Dieses Argument gilt analog zu einem zentralen Argument, das in der Sache 1 StR 156/98 und 1 StR 258/98 gegen den sogenannten Polygraphentest vorgebracht wurde und letztlich zu der Einsicht geführt hat, daß die meisten Anwendungen von Polygraphentests (i.e., verschiedene Formen des Kontrollfragentests) ungeeignet sind. Der Versuch, einen bestimmten Test oder einen anderen Indikator als festes Diagnostikum im Sinne eines universellen Gesetzes zu etablieren, das dann von jedem Gutachter ohne kritische

Prüfung eines spezifischen, auf den Einzelfall zugeschnittenen Modells gleichbleibend eingesetzt werden kann, wäre nicht nur aus den oben genannten Gründen unberechtigt. Ein solches Verfahren würde vor allem auch voraussetzen, daß ein solches Gesetz einem sehr aufwendigen quantitativen Prüfungs- und Normierungsverfahren unterworfen wird. Dazu wäre es insbesondere erforderlich, eine nicht-verzerrte, repräsentative Stichprobe von Aussagen zu kennen, deren wirklicher Wahrheitsgehalt zweifelsfrei bekannt ist. Diese Voraussetzung ist gerade bei denjenigen Aussagen, wo psychologische Wahrheitsdiagnostik eigentlich benötigt wird, nicht gegeben. Diejenigen Aussagen, die am Ende in eine Validierungsstudie eingehen, können in selektiver Weise so stark verzerrt sein, daß sie zu massiven Fehlschlüssen führen. Diese Gefahr ist besonders dann gegeben, wenn die Ermittlung der Wahrheit von dem zu validierenden Test oder Kennzeichen nicht unabhängig ist, so daß die Validität systematisch überschätzt wird (Fiedler, BGH Gutachten in o.a. Sache).

(5) Die negative Aussage, daß einzelne gut bestätigte psychologische Gesetze keine Wahrheitsdiagnostik deduktiv begründen können, schließt besonders auch solche Indikatoren ein, die als Subtests gängiger Persönlichkeitstests den Namen "Lügenskala" tragen (MMPI, FPI) und somit den Eindruck suggerieren, eine Eigenschaft "Ehrlichkeit" bzw. "Wahrheitsliebe" individueller Personen zu messen. Diese Subtests haben in erster Linie die Aufgabe, Tendenzen der Selbstdarstellung und der nicht realitätsgetreuen Darstellung *im Persönlichkeitstest* zu ermitteln. Informativ sind diese Subtests vor allem für die Identifikation von pathologischen Fällen sowie für die Diagnose einer bei allen Menschen mehr oder weniger stark ausgeprägten Tendenz der Selbstdarstellung ("soziale Erwünschtheit"), die mit den vielfältigen Gründen für eine Falschaussage im forensischen Kontext nicht das Geringste gemeinsam haben müssen.

Der Begriff "Lügenskala" in einem Persönlichkeitstest wie überhaupt der gängige Begriff der "Glaubwürdigkeitsbegutachtung" (etwa in der vorliegenden Fragestellung des BGH) könnte den Schluß nahelegen, daß Glaubwürdigkeit ein stabiles Persönlichkeitsmerkmal ist, welches eine Vorhersage der Ehrlichkeit bestimmter Personen bei beliebigen anderen Gelegenheiten gewährleistet. Auch für diese weit verbreitete Annahme gibt es keinerlei Berechtigung. Zwar wird die Existenz von kriminellen oder pathologischen Extremfällen (Personen, die notorisch oder pathologisch lügen) nicht angezweifelt; bei diesen Fällen, die jedoch eher Ausnahmen sind und selten große Probleme bei der Diagnostik aufwerfen – weder für Psychologen noch für Richter – kann man sicher mit einer großen Wahrscheinlichkeit eine Bereitschaft zur Verfälschung erwarten. Die Generalisierung indessen, daß bei allen Menschen aufgrund des Vorkommens einer Lüge oder Falschaussage in der Vergangenheit bzw. in einem diagnostischen Gespräch eine wissenschaftlich begründete Vorhersage der Wahrheit in einem aktuellen Einzelfall

möglich ist, entbehrt jeder Grundlage. Mehrere Untersuchungen im sozialpsychologischen Kontext (DePaulo et al., 1996; Turner et al., 1975) zeigen vielmehr, daß Abweichungen von der Wahrheit aus den verschiedensten Motiven bei virtuell allen Menschen unter bestimmten Bedingungen erwartet werden können. Mit anderen Worten, auch dieser Ansatz der Etablierung von idiomatischen Gesetzen (d.h. personspezifische Wahrheitstendenzen) hat aus wissenschaftlicher Sicht keinen Bestand.

Einschränkend sei nur hinzugefügt, daß eine Falschaussage im konkreten Kontext eines Gerichtssaales bzw. eines polizeilichen Verhörs in der Vergangenheit sehr wohl ein sehr nützlicher Indikator wiederum im Rahmen eines spezifischen Handlungsmodells sein kann.

### 3.2 Evidenz für eine induktiv-statistische Glaubwürdigkeitsdiagnostik

(6) Während es einerseits keinerlei Hinweise auf universell verwendbare Gesetze in der Glaubwürdigkeitsdiagnostik gibt – und wegen der Heterogenität des Gegenstandes auch nicht geben kann – stützen andererseits zahlreiche Befunde die Annahme, daß Aussagemerkmale wie die von Steller et al. (1992) in der obigen Tabelle sehr nützliche Indikatoren im Rahmen eines induktiv-statistischen **"multiple-cue"-Modells** (Lee & Yates, 1992) abgeben können. Obwohl der Wert einzelner Merkmale bzw. Indikatoren in der Regel sehr bescheiden bleibt, gestattet die Gesamtheit multipler Cues in vielen Untersuchungen eine hoch signifikante Diskrimination zwischen wahren und falschen Aussagen. Dieser Befund ist typisch für die Annahme eines probabilistischen Entscheidungsmodells, in dem durch Aggregation über multiple Indikatoren eine deutlich höhere Gesamtgenauigkeit erreicht wird. Die in der Literatur oft betonte statistisch gute oder befriedigende Trennbarkeit von wahren und falschen Aussagen betrifft stets die Gesamtheit vieler Indikatoren als Aggregat, aber niemals die Validität einzelner Indikatoren bzw. vermeintlicher Gesetze.

(7) Typisch für ein solches Modell mit multiplen Indikatoren, die für sich keine feste Bedeutung und Diagnostizität haben, ist auch die wechselhafte Funktion der Indikatoren, die sowohl Wahrheit wie Unwahrheit anzeigen können, was zu den oben beschriebenen Inversionen führt. Dies kommt in probabilistischen Umwelten nicht selten vor und spiegelt die Tatsache wider, daß die Indikatoren oder "Cues" keine feste, gesetzesartige (z.B. kausale) Reflexion der Wahrheit sind, sondern lediglich Korrelate, die je nach Modell unterschiedliche Funktion ausfüllen können (z.B. Detailreichtum als Symptom von authentischem Erleben oder von raffinierter Sprache). Übrigens führt die Aggregation über mehrere schwache Indikatoren auf so robuste Weise zu erhöhter Genauigkeit, daß einzelne invertierte Indikatoren von einer Mehrzahl richtig eingesetzter Indikatoren verdeckt werden (vgl. das Beispiel *der zeitlichen Details* in Abschnitt 2.2.2.).

(8) Daß verschiedene Autoren bzw. Gutachter mit teilweise unterschiedlichen Kennzeichen scheinbar ähnlich gut arbeiten, ist ebenfalls im Rahmen eines solchen statistischen Bezugsrahmens verständlich. Ein vorteilhafter Aspekt der Robustheit und des prinzipiellen Nutzens von multiplen Indikatorsystemen ist ihre Austauschbarkeit. Da die einzelnen Indikatoren keine essentiellen Ursachen oder Wirkungen des zu erfassenden Sachverhalts darstellen müssen, sondern lediglich schwach korrelierte Zeichen, liegt ein großer Vorteil derartiger Systeme in ihrer Flexibilität. Dieser als "vicarious functioning" bezeichnete Vorteil findet sich übrigens nicht nur in diagnostischen Modellen, sondern auch in vielen natürlichen Systemen, die unter Unsicherheit Lösungen finden und Entscheidungen treffen müssen, deren Effizienz angesichts der Schwäche der verwendeten Indikatoren überraschend hoch ist (Brunswik, 1955; Gigerenzer & Goldstein, 1996). Ein Beispiel ist etwa menschliches Tiefensehen (Entfernungssehen), wo für sich genommen schwache Indikatoren (Glanz der Oberfläche, Disparität der beiden Netzhautbilder etc.) zusammen erstaunliche Genauigkeit erzielen und den Ausfall einzelner Indikatoren leicht verkraften können. Diese Bezüge seien hier nur deshalb erwähnt, um deutlich zu machen, daß ein psychologischer und mathematischer Bezugsrahmen zur Erklärung der erstaunlichen Genauigkeit von Systemen schwacher Prädiktoren schon seit langem existiert und formal sehr weit entwickelt ist.

(9) So gibt es auch in der Grundlagenforschung – außerhalb der forensischen Praxis –gut bestätigte und durch Meta-Analysen (Ambady & Rosenthal, 1992) untermauerte Befunde, welche die Wirksamkeit schwacher Indikator-Systeme speziell bei der alltäglichen Glaubwürdigkeitsbeurteilung bestätigen. Empirische Analysen und Meta-Analysen zeigen, daß die Genauigkeit, mit der Täuschungen und Lügen aufgrund minimaler Information (d.h. anhand sehr schwacher Indikatoren) entdeckt werden, systematisch über der Zufallserwartung liegt (u.a., DePaulo, Lassiter & Stone, 1982; Fiedler & Walka, 1993; Manstead, Wagner & McDonald, 1986). Durch die gleichzeitige Nutzung mehrerer Indikatoren, die für sich genommen alle von sehr begrenztem Wert sind, kann ein deutlicher Gewinn an Diskriminationsleistung erzielt werden. Ob es sich um intuitive Glaubwürdigkeitsurteile handelt oder um quasi-systematische Auszählungen von Aussagenmerkmalen in einer Art Inhaltsanalyse ist hierbei nebensächlich. Wiederholt sei in diesem Zusammenhang nur, daß dasselbe Prinzip der Aggregation über viele Indikatoren bei fast allen psychologischen Tests eine wichtige Rolle spielt. Bei typischen Leistungs-, Persönlichkeits- oder Einstellungstests haben einzelne Testaufgaben eine sehr begrenzte Trennschärfe und damit auch eine sehr begrenzte Genauigkeit. Erst durch Aggregation der Testleistung über viele Indikatoren hinweg erreichen etablierte Tests ihre erwiesene Reliabilität und Validität. Aggregation über schwache Indikatoren ist also keine "unsaubere" Methode, sondern ein wissenschaftlich anerkanntes methodisches Prinzip (auch in der Nachrichtentechnik, den Computerwissenschaften oder anderen Disziplinen).

Das Prinzip der Aggregation ist wegen seiner Mächtigkeit und Robustheit von großer Bedeutung für jede Form der Diagnostik. Wenn die verschiedenen Indikatoren zumindest leicht überzufällig mit dem Vorliegen einer wahren Aussage korrelieren, dann steigt die Gesamtvalidität mit wachsender Zahl von Indikatoren auch dann an, wenn wenige einzelne Indikatoren invertiert sind, also einen negativen Beitrag leisten. Wegen dieser günstigen mathematischen Eigenschaften derartiger Indikator-Systeme erscheint die Chance, eine Menge von brauchbaren und wirksamen Indikatoren für die Glaubwürdigkeitsdiagnostik zu finden und zu nutzen, durchaus realistisch.

### 3.3 Fehlschlüsse durch selektive Nutzung von Indikatoren

Eine entscheidende Voraussetzung für die diagnostische Nutzung solcher Indikator-Systeme – und mitverantwortlich für die empirisch mehrfach beobachtete Genauigkeit solcher Systeme (Ambady & Rosenthal, 1992) – ist jedoch wie bereits oben klargestellt die repräsentative, nicht-selektive Auswahl der Indikatoren. Typisch für die Bedingungen, unter denen die Diskrimination von wahren und falschen Aussagen aufgrund minimaler Information erfolgreich war, ist die Nicht-Selektivität der beurteilten Beobachtungen (vgl. Brunswik's, 1955, Forderung nach "representative sampling").

Durch Einschränkung der Information auf wenige selektive Indikatoren, die einem bestimmten favorisierten Modell entsprechen, und Ignorieren anderer Indikatoren, die andere denkbare Modelle bestätigen könnten, werden unter Umständen erhebliche Fehler erzeugt. So zeigen unmittelbar mit Glaubwürdigkeit befaßte Experimente (z.B. Zuckerman, Koestner, Colella, & Alton, 1984), daß Aussagen eher für falsch gehalten werden, wenn Urteiler die Hypothese einer möglichen Lüge testen, während dieselben Aussagen eher für wahr gehalten werden, wenn die Hypothese einer wahren Äußerung focussiert wird. In der psychologischen Forschung im allgemeinen (Jussim, 1991; Koehler, 1991) und der Forschung zum Hypothesentesten in Gesprächen und Interviews im besonderen wurde vielfach demonstriert, daß die Ergebnisse systematisch in Richtung auf die Ausgangshypothese verzerrt sind (Snyder, 1984; Pyszczynski & Greenberg, 1988; Tversky & Kahneman, 1974; Zuckerman et al., 1995). Einer von mehreren Gründen für diesen sogenannten "confirmation bias" (Snyder & Swann, 1978) bzw. "autoverification effect" (Fiedler, Walther & Nickel, 1999) ist die einseitige, nicht-repräsentative Suche nach Indikatoren für die leitende Hypothese und die gleichzeitige Vernachlässigung von Indikatoren für alternative Hypothesen (Kunda, 1990; Semin & Strack, 1980; Snyder & Swann, 1978; Wason, 1966; inter alia). Eine beispielhafte Illustration der Gefahr selektiver Indikatoren und des Versäumnisses, alternative Modelle zu berücksichtigen, liefert der nun folgende Abschnitt.

### 3.4 Probleme bei der Nutzung multipler Indikatoren im Rahmen induktiv-statistischer Schlüsse – Beispiele und Illustrationen

Das Glaubwürdigkeitsgutachten des Diplom-Psychologen Dr. S. in der Sache AZ: 15 Js 1157/97, das in verschiedener Hinsicht als Negativbeispiel gelten kann, sei hier herangezogen, um die teilweise abstrakten Thesen über das Selektionsproblem und die Explikation diagnostischer Modelle zu verdeutlichen. Abgesehen von der bloßen Dokumentation verschiedener Gespräche und Aussagen der beiden Zeuginnen und der Anwendung einiger völlig unspezifischer Tests (Progessive Matrizen, Giessen-Test, Bilder Ergänzen) stützt sich der Schluß, daß die beiden Mädchen die Wahrheit sagen, unter anderem auf die Beobachtung von einigen Realkennzeichen (*eigene gefühlsbezogene Abläufe; unverstandenes Handlungselement; Benennung von örtlichen Gegebenheiten; Erinnerungslücken*). Mit Fokus auf **Real**kennzeichen (d.h., auf eine implizite Hypothese, die in einseitiger Weise auf Indikatoren der Wahrheit gerichtet ist) sucht der Gutachter einfach nach der Existenz irgendwelcher Kennzeichen dieser Art und schließt auf die Wahrheit der Aussage, weil zumindest einige dieser Kennzeichen in dem Text zu finden sind.

Das implizite Modell scheint hier anzunehmen, daß allein das vereinzelte Vorkommen solcher Kennzeichen die Wahrheit anzeigt, was sofort als unbegründet wenn nicht abwegig zu erkennen ist. Was hier offensichtlich unberücksichtigt bleibt, ist die Frage, wie viele derartige Kennzeichen in einem bestimmten Text gegebener Länge und gegebenen Inhalts zu erwarten sind, wie viele Kennzeichen das benutzte Indikator-System überhaupt umfaßt (d.h. wie viele überhaupt gefunden werden könnten) und nicht zuletzt auch, wie oft die Zeuginnen es unterlassen, relevante Realkennzeichen zu verwenden. Fraglich ist natürlich auch, wie viele Kennzeichen von Unwahrheit der Gutachter finden könnte, wenn er die alternative Hypothese der Unwahrheit verfolgend nach Indikatoren wie *Zögern, Ausweichen, selektives Vergessen der Zeugin* etc. suchen würde. Vielleicht wäre diese alternative Suche ergiebiger. In Ermangelung von linguistischen Normen darüber, welche Rate von Realkennzeichen in verschiedenen Textcorpora bestimmter Länge zu erwarten sind, wird dem kontrastierenden Hypothesentesten sogar entscheidende Bedeutung zukommen. Es gibt keinen normierten Grenzwert, ab welcher Zahl oder Dichte von Realkennzeichen eine Wahrheit beginnt. Folglich kann man die Wahrheitshypothese nicht absolut prüfen, sondern muß sie durch Kontrastieren verschiedener Hypothesen in ihrer relativen Plausibilität prüfen.

Zur Vorbereitung der im nächsten Abschnitt präsentierten Forderungen an sachgemäße Glaubwürdigkeitsgutachten können an diesem Beispiel weitere Probleme verdeutlicht werden. Zunächst einmal wird das Problem der Objektivität der Beobachtungen sowie deren Reliabilität durch die Stellungnahme von Prof. Dr. K. offenkundig. Hieraus wird deutlich, daß die Codierung oder Klassifikation bestimmter Aussagenelemente als *Erinnerungslücken*, *unverstandene Handlungselemente* oder *spontane Verbesserungen der Aus-*

*sage* durchaus subjektiv und problematisch ist. Köhnken stellt klar, daß die Äußerung, etwas nicht zu wissen, keinesfalls als unverstandenes Handlungselement zu interpretieren sein muß, und wirft ähnliche Fragen bezüglich der Beobachtung anderer Indikatoren auf. Dies zeigt sehr lebhaft die Notwendigkeit, die Reliabilität von Beobachtungen zu kontrollieren.

Sofern diagnostische Verfahren keine standardisierten Tests sind, deren Reliabilität bekannt und normiert ist, besteht eine jederzeit mögliche Methode darin, die Reliabilität aufgrund der aktuellen Beobachtungen zu schätzen. Eine Möglichkeit ist die Bestimmung der *internen Konsistenz*, also die Berechnung der Interkorrelation zwischen verschiedenen Indikatoren, die hypothetisch dasselbe messen. Ein Mindestmaß an interner Konsistenz (d.h. daß die Annahme gerechtfertigt ist, daß die Gesamtheit aller Indikatoren eine gemeinsame latente Größe messen) ist Voraussetzung für einen merklichen Aggregationseffekt. Leider wird diese psychometrisch wichtige Annahme jedoch in nahezu allen Untersuchungen zu Realkennzeichen und erst recht in der forensischen Gutachtenpraxis vernachlässigt (vgl. Wells & Loftus, 1991). Wenn in eine Untersuchung eine Vielzahl von Aussagen einbezogen werden, ist die interne Konsistenz leicht über Interkorrelation zwischen den Indikatoren über die Aussagen hinweg bzw. durch Faktorenanalyse zu bestimmen. Liegt nur ein einziger Fall vor (wie im vorliegenden Beispiel), dann ist die interne Konsistenz schwerer zu bestimmen. Wenn ein einzelnes Gespräch länger andauert, könnte man versuchen, die Interkorrelation verschiedener Indikatoren über verschiedene Abschnitte derselben Aussage hinweg zu korrelieren. Sollten Gutachter sich weiterhin auf die gebräuchlichen Realkennzeichen (vgl. Steller und Köhnken, 1989) stützen, so wäre es in jedem Falle wünschenswert, die interne Konsistenz dieses Systems von Indikatoren durch begleitende Forschung zu bestimmen und die Zahl und Auswahl der Indikatoren gegebenenfalls zu beschränken, so daß diese Forderung erfüllt werden kann.

Eine etwas andere Art, die Zuverlässigkeit zu bestimmen und systematische Beobachtungen gegenüber zufälligen Ereignissen abzusichern, besteht in der Replikation, analog zu dem Prinzip der Retest-Reliabilität. Wenn die in einer Aussage beobachteten Kennzeichen nicht klar genug sind und die interne Konsistenz nicht bekannt ist, sollten die Beobachtungen wenigstens durch Testwiederholung repliziert und kreuzvalidiert werden. (Kreuzvalidieren bedeutet, einen nicht vorhergesagten Befund durch einen zweiten, unabhängigen Test sichern). Von dieser Forderung einer methodisch adäquaten Diagnostik wird – nicht nur aus Kostengründen – leider zu wenig Gebrauch gemacht. Ein vorbildliches Merkmal eines positiv zu bewertenden Gutachtens ist immer der dezidierte, klar erkennbare Versuch, einmal festgestellte Beobachtungen zu replizieren bzw. daraus abgeleitete Folgerungen zu kreuzvalidieren.

Auch bei einzelnen Aussagen wird man aber in jedem Fall der Forderung nach Bestimmung der Beobachtungsobjektivität nachkommen können. In dem Beispiel-Gutachten hätte man das angebliche Vorliegen der Realkennzeichen leicht durch Berechnung der Codier-Übereinstimmung verschiedener Urteiler bestimmen können. Mit großer Wahrscheinlichkeit hätte sich gezeigt, daß die Klassifikation der von Köhnken hervorgehobenen Beobachtungen nicht sehr zuverlässig ist. Diese Forderung ist weder unrealistisch und "praxisfremd" noch zu teuer, weil die Codierung keine forensische Erfahrung verlangt und auch von Hilfskräften erlernt werden kann. (Codierer, welche die Objektivität bestimmen helfen, müssen schließlich nicht die verantwortliche Codierentscheidung treffen!)

Überhaupt gibt es gute Gründe, den Teil der Aussagendiagnostik, der objektiv sein soll, ohne Vorwissen des übrigen Falles von "blinden" Codierern vornehmen zu lassen – entgegen der unter Praktikern weit verbreiteten Auffassung, daß ein "verstehender Diagnostiker", der alle übrigen Daten kennt, neue Beobachtungen besser interpretieren kann. Methodologisch läuft ein solches "verstehendes Beobachten" häufig auf voreingenommenes, nichtobjektives Beobachten hinaus. Bemerkenswert in diesem Zusammenhang ist als nennenswerte Ausnahme die Untersuchung von Krahé und Kundrotas (1992), in der als einer der wenigen die Urteiler-Übereinstimmung (zwischen 4 Urteilern) bei der Feststellung der Realkennzeichen erfaßt wurde. Dort fallen die Urteiler-Übereinstimmungen sehr bescheiden aus; sie bewegen sich zwischen Kappa = 0.025 und Kappa = 0.356. (Kappa ist ein Übereinstimmungskoeffizient, der zwischen 0 und 1 variiert). Auch wenn die Urteiler in dieser Studie vorher nicht ausreichend trainiert waren, zeugen diese Daten doch von der grundsätzlichen Problematik der Sicherung der Beobachtungen. Ein wichtiger Teil der Bewertung von Gutachten wird sich in jedem Fall mit der Objektivität und Reliabilität der Beobachtungen befassen müssen.

Das betrachtete Beispiel-Gutachten ist ferner dazu angetan, das Fehlen von gezielten diagnostischen Modelltests zu illustrieren. Zugrunde gelegt wird hier implizit, ohne explizit ein Modell zu testen, ein in der Gutachtenpraxis weit verbreitetes klinisches Entscheidungsmodell. Die von einem Zeugen abgegebene Aussage wird in ihrer linguistischen und kognitiven Qualität an den Möglichkeiten gemessen, die der Zeuge aufgrund einer unspezifischen klinischen Untersuchung mitzubringen scheint. Angesichts einer allgemeinen Untersuchung der verbalen Fähigkeiten, des Gedächtnisses, der seelischen Belastbarkeit und der kognitiven Differenziertheit (typischerweise mit einigen gängigen Tests) wird die Entscheidungsfrage gestellt, ob ein Zeuge mit solchen Persönlichkeitseigenschaften eine Aussage mit dem gegebenen Niveau zu fabrizieren imstande wäre, wenn sie nicht ein wirklich erlebtes Ereignis widerspiegeln würde. Die Selektion der benutzten Indikatoren orientiert sich also an einem idiomatischen Modell der absoluten Kompetenzen und seelischen Möglichkeiten einer jeweiligen Persönlichkeit aus.

Ein solch starres, an stabilen Persönlichkeitsmerkmalen orientiertes Modell der Glaubwürdigkeitsdiagnostik ist in verschiedener Hinsicht mangelhaft und mit dem heutigen Stand der einschlägigen Forschung nicht vereinbar. Es ignoriert insbesondere moderne Erkenntnisse (Johnson & Raye, 1981; Loftus, 1979; Schwarz & Sudman, 1994) zu vier wesentlichen Gebieten der psychologischen Forschung: Lügenproduktion und -detektion, Gedächtnis, Antworttendenzen und suggestive Beeinflussung. Die Fähigkeit zur gefälschten Produktion einer Aussage – im Sinne einer persönlichkeitsspezifischen Beschränkung – ist nur eine von sehr vielen Bedingungen, von denen die Aussagenproduktion abhängt. Dabei ist diese Bedingung noch nicht einmal essentiell. Ein Zeuge kann eine sehr differenzierte, detailreiche, im Phantasiegehalt sehr lebhafte Aussage abgeben, obwohl er das ausgesagte Geschehen nicht wirklich erlebt hat und obwohl er selbst keinen sehr detailreichen Sprachstil besitzt. Die Erklärung kann einfach darin liegen, daß der Aussage eine fremderzeugte Geschichte zugrunde liegt, oder ein in Details verändertes anderes Erlebnis, oder die Aussage kann einfach den Einfluß wiederholter Befragungen (etwa durch Therapeuten) widerspiegeln, wobei viele Details und Vorstellungen von den Befragern suggeriert worden sind. Im übrigen gibt es eine Reihe anderer Motive und Ursachen für eine von den Tatsachen abweichende Falschaussage: Eigene Gedächtnistäuschungen oder Wahrnehmungstäuschungen bis hin zu Wahnvorstellungen des Zeugen, Selbstdarstellung, Bedürfnis nach Schutz des Selbstwertgefühls (McDowell & Hibler, 1987), nachträgliche Suggestionen (Köhnken & Maass, 1988; Loftus, 1979), mentale Vorstellungen und Imaginationen (Koehler, 1991), Antworttendenzen (Ja-Sage-Tendenzen) oder Quellenkonfusion von tatsächlich erlebtem und mental simuliertem Geschehen (Johnson, Hashtroudi, & Lindsay, 1993).

Die moderne psychologische **Lügenforschung** zeigt, daß Falschaussagen keine exklusiven Symptome bestimmter Menschen sind, die sich durch geringere Wahrheitsliebe von anderen unterscheiden, sondern daß alle oder zumindest viele Menschen im Alltag sehr häufig von der reinen oder vollen Wahrheit abweichen, wobei die Motive in der Regel nicht einmal eigennützig oder verwerflich sind (DePaulo, Kashy, Kirkendol, Wyer, & Epstein, 1996; Turner, Edgley & Olmstead, 1975). Hiermit soll nicht in Abrede gestellt werden, daß es extreme oder pathologische Fälle von notorischen Lügnern (oder halluzinierenden Psychotikern) gibt, die sich regelmäßig und ohne jeden Zweifel immer wieder in Widersprüche und Falschaussagen verstricken. Diese Ausnahmen sind jedoch – wegen ihrer Abnormalität – meist leicht zu identifizieren und bedürfen deshalb kaum einer Aussagenanalyse. Sie sollten indessen nicht den Schluß rechtfertigen, daß ein idiomatisches Modell der glaubwürdigen versus unglaubwürdigen Persönlichkeit eine generelle Grundlage für die Diagnostik von Falschaussagen im Einzelfall bietet.

Wird anstelle eines **klinischen oder naiv-charakterologischen Modells**, das

die Hypothese einer unglaubwürdigen Persönlichkeit in den Vordergrund stellt, die Begutachtung von einem häufig angemesseneren **gedächtnispsychologischen Modell** geleitet, so kann ein allgemeiner Gedächtnistest (Untertest aus einem gängigen Intelligenztest) mit dem Ziel, die allgemeine Gedächtnisfähigkeit eines Zeugen zu messen, kaum als angemessenes Verfahren gelten. Wenn die moderne Gedächtnisforschung der letzten beiden Jahrzehnte ein unstrittiges Ergebnis erbracht hat, dann ist es die Erkenntnis, daß die Vorstellung einer allgemeinen Gedächtnisstärke (memory strength) nicht mehr haltbar ist (Bjork, 1994). Was bei einer bestimmten Gedächtnisaufgabe (z.B. freie Wiedergabe; Recall) leicht und genau behalten wird, kann bei einer anderen Aufgabe (Recognition) vergessen oder verwechselt werden. Implizites und explizites Gedächtnis sind ebenso unabhängig wie prozedurales und deklaratives Gedächtnis. Sogar Hirngeschädigte mit massiven Gedächtnisausfällen in bestimmten Bereichen können bei anderen Funktionen (Sprache, älteres Wissen, implizites Gedächtnis) völlig intakte Leistungen zeigen (Nelson, 1992; Squire, 1986). Von einem undifferenzierten Test, der eine beliebige Gedächtnisfunktion herausgreift und als Baustein eines allgemeinen Intelligenzmodells interpretiert, sind für die Diagnostik der Glaubwürdigkeit von Aussagen keine besonderen Erkenntnisse zu erwarten.

In diesem Abschnitt wurde ausgehend von einem konkret vorliegenden Gutachten aufgezeigt, welche schweren und offenkundigen Probleme sich ergeben, wenn die diagnostischen Beobachtungen nicht kontrolliert und methodisch abgesichert werden und wenn die explizit oder implizit verwendeten Modelle des Diagnostikers nicht gesichert und kritisch geprüft werden. Im nächsten Abschnitt sollen die hier exemplarisch skizzierten Probleme systematisch zusammengestellt werden. Die resultierenden Übersichtstabellen können als Richtschnur für die Bewertung von Glaubwürdigkeitsgutachten nach wissenschaftlichen Kriterien verwendet werden.

## 4 Eine Systematik von Kriterien zur Bewertung von Gutachten

Dieser Abschnitt bildet das Kernstück unseres Papiers. Er enthält in allgemeiner, tabellarisch zusammengefaßter Form eine Übersicht von Kriterien, die eine nach wissenschaftlichen Maßstäben adäquate Untersuchung bzw. Begutachtung erfüllen sollte. Zugleich bilden diese Kriterien die Grundlage für die Bewertung und begründete Kritik solcher Gutachten.

Wie diese Systematik von Kriterien in den vorausgehenden Abschnitten hergeleitet wurde, sei hier noch einmal kurz rekapituliert. Ausgehend von derjenigen Disziplin, die sich mit Form und Logik von wissenschaftlichen Begründungen befaßt – der Wissenschaftstheorie – wurde eine grundlegende Unterscheidung zwischen deduktiv-nomologischem Beweis und induktiv-statistischen Schlüssen eingeführt. Es wurde sodann aufgrund der vorhandenen empirischen Forschung geschlossen, daß für deduktiv-nomologische Beweise in Glaubwürdigkeitsgutachten jegliche Grundlage fehlt. Auch ohne

gut bestätigte, universelle Einzelgesetze gibt es jedoch eine rationale Grundlage für eine leistungsfähige Diagnostik mithilfe induktiv-statistischer Schlüsse. Durch Aggregation über eine Reihe von probabilistischen Indikatoren kann eine diagnostische Entscheidung auch dann sehr verläßlich und genau sein, wenn die einzelnen Indikatoren nur einen bescheidenen Beitrag knapp über der Zufallsgrenze leisten. Tatsächlich bestätigen empirische Befunde, daß wahre und unwahre Aussagen durch Aggregation über mehrere Indikatoren oft mit einer befriedigenden Wahrscheinlichkeit getrennt werden können. Die Problematik bei solchen induktiv-statistischen Schlüssen liegt jedoch in der Selektion und Verläßlichkeit der Indikatoren. Eben weil ein universelles Gesetz nicht angenommen wird, muß die Eignung der jeweiligen Indikatoren im Rahmen eines klar definierten diagnostischen Modells begründet und anhand der vorliegenden diagnostischen Daten kritisch geprüft werden. Da dieselben Indikatoren im Kontext verschiedener Modelle unterschiedlichen Wert haben können, besteht das Ziel einer wissenschaftlich fundierten Diagnostik im kontrastierenden Vergleich verschiedener Modelle, die alternative Erklärungen für die vorhandenen Daten anbieten.

Hieraus ergeben sich zwei Schwerpunkte einer solchen Diagnostik auf der Grundlage von induktiv-statistischen Schlüssen: (a) Explikation von alternativen Modellen und Selektion von gezielten Indikatoren zur Prüfung dieser Modelle; sowie (b) Sicherung der diagnostischen Beobachtungen hinsichtlich dieser Indikatoren nach psychometrisch angezeigten Verfahren. Die im folgenden präsentierten und kommentierten Tabellen beziehen sich auf Gütekriterien für diese beiden Aspekte. Zunächst wird in einer Tabelle zusammengefaßt, was unter dem Gebot der Explikation eines diagnostischen Modells zu verstehen ist. Erläuterungen dazu folgen unmittelbar danach. Später wird in einer zweiten Tabelle zusammengestellt, welche Maßnahmen nach dem heutigen Stand der Kunst unternommen werden sollen bzw. können, um auf die Modelle bezogene diagnostische Beobachtungen zu selektieren und auf ihre Eignung hin zu überprüfen. Auch diese Tabelle wird anschließend erläutert.

### 4.1 Explikation diagnostischer Modellannahmen

Das Gebot, die diagnostischen Modellannahmen offenzulegen und nachvollziehbar zu machen, wird in der ersten Tabelle ausdifferenziert. Ansätze zur adäquaten Lösung des Problems werden aufgezeigt und Hinweise gegeben, wie die sachgerechte Erfüllung der Kriterien im Gutachten dokumentiert werden können.

**Explikation der diagnostischen Modellannahmen:**

| Problem | Ansatz der Problemlösung | Nachweis im Gutachten |
|---|---|---|
| Explizite Prüfung diagnostischer Modelle | a) Gedächtnismodelle: Inwiefern kann die Übereinstimmung vs. Abweichung zwischen Aussagen und tatsächlichen Sachverhalten gedächtnispsychologisch erklärt werden? | Gezielte Formulierung konstrastierender Hypothesen, die logisch und psychologisch eine Prüfung der Modelle anhand diagnostischer Daten gestatten. |
| | b) Emotionale Modelle: Inwiefern kann emotionale Belastung die Beziehung zwischen Aussagen und tatsächlichen Sachverhalten erklären? | |
| | c) Täuschungsabsicht: Gibt es Gründe zu der Annahme, daß eine Aussage die tatsächlichen Sachverhalte im Sinne einer Täuschung oder Lüge bewußt (trotz intakten Gedächtnisses) verfälscht? | |
| | d) Glaubwürdigkeit als Persönlichkeitsmodell: Kann ein idiomatisches Gesetz, wonach Lügen und realitätsverletztende Aussagen ein stabiles und vorhersagbares Persönlichkeitsmerkmal darstellt, ein psychologisch plausibles und diagnostisch trennscharfes Modell abgeben? | |
| Referentialität | a) Wissenschaftlicher Hintergrund für die Modellannahmen | a) Zumindest übersichtsweise Quellenangaben |
| | b) Erfahrungshintergrund des Gutachters als Grund für den diagnostischen Modellansatz | b) Umfang und Quelle der eigenen Erfahrung. Berufspraktische Standards |

| | | |
|---|---|---|
| Selektionsentscheidungen. Problem der Operationalisierung | a) Welche Teilmenge relevanter Hypothesen bzw. Gesetzesannahmen wird in die Untersuchung einbezogen? Wird die Auswahl bestimmt durch gezielte modellbasierte Überlegungen oder durch die Verfügbarkeit der vorhandenen Beobachtungen oder Daten?<br><br>b) Welche Beobachtungen, Testdaten oder Indikatoren werden zur Untersuchung welcher Hypothese herangezogen? | Wissenschaftliche (bekannte Befunde) oder pragmatische (Verfügbarkeit) Begründung für die Wahl der fokussierten Hypothese/ Beobachtungen. Explizite Gründe für die Exklusion anderer Hypothesen bzw. potentiell relevanter diagnostischer Daten. |

Die Tabelle enthält drei Spalten. In der ersten Spalte wird das Ziel – Explikation der diagnostischen Modellannahmen – in drei Aspekte zerlegt: Ein wissenschaftlich angemessenes und sorgfältiges Gutachten sollte klar zu erkennen geben, welches die Modellannahmen des Untersuchers sind (1. Teilaspekt), welches seine theoretischen oder erfahrungsbasierten Hintergründe für die Festlegung auf bestimmte Modellannahmen sind (2. Aspekt) und wie die Prüfung dieser Modellannahmen durch eine und klar zugeordnete Auswahl von diagnostischen Indikatoren bzw. Beobachtungen erfolgen soll (3. Aspekt). Eine klare Zuordnung der verwendeten Indikatoren zu expliziten Modellannahmen gibt Aufschluß darüber, ob einem Gutachten ein tragbares Konzept zugrunde liegt, das auf nachvollziehbare Weise kritisch geprüft wurde, oder aber konzeptionslos die gerade verfügbaren Daten oder die Befunde aus irgendwelchen Routinetests in willkürlicher Weise ausdeutet. Ob ein Gutachten als wissenschaftlich fundiert und methodisch stichhaltig gelten kann, wird so in den meisten Fällen unverkennbar sein.

Die mittlere Spalte zeigt mögliche Ansätze zum Umgang mit diesen drei Teilzielen auf. So wie die Annahme von universellen Gesetzen zur Wahrheitsdeduktion unberechtigt wäre, gibt es auch keine universell indizierte, stets zu befolgende Modellannahme. Die in der zweiten Spalte aufgeführten Klassen von Modellannahmen verstehen sich daher nicht als normativer Katalog, die in jedem einzelnen Falle nach einem bestimmten Schema zu testen sind. Dennoch meinen wir, daß ein sorgfältiges Glaubwürdigkeitsgutachten kaum darauf verzichten kann, auf bestimmte Modellklassen wenigstens einzugehen. Dabei ist zu beachten, daß die verschiedenen Modelle nicht unabhängig sind und teilweise in einer hierarchischen Beziehung zueinander stehen.

Ein ganz allgemeines Modell mag annehmen, daß eine Aussage von der Realität abweicht, weil ein Zeuge als informationsübertragendes System be-

stimmte Fehlfunktionen zeigt. Dieses Modell schließt als Spezialfälle unter anderem eine Wahrnehmungshypothese (der Zeuge unterlag einer Wahrnehmungstäuschung), eine Persönlichkeitshypothese (der Zeuge ist psychisch krank) und etwa eine Gedächtnishypothese ein (es handelt sich um fehlerhaftes Gedächtnis). Innerhalb einer solchen Gedächtniskonzeption kann man dann wiederum noch feinere Differenzierungen vornehmen. Die Gedächtnistäuschung kann auf Vergessen beruhen oder auf Suggestion von außen oder auch auf motivierten Prozessen (Verdrängen unerträglicher Inhalte). Wichtig ist, daß man in einer solchen Hierarchie von Modellen oder Hypothesen nicht spezielle Modelle testen darf, bevor man allgemeinere, übergeordnete Modelle betrachtet und alternative ausgeschlossen hat. In aller Regel wird dabei ein allgemeines Gedächtnismodell ein logisches Primat haben gegenüber spezielleren, logisch untergordneten Modellen wie pathologische Gedächtnisschwäche oder motiviertes Verdrängen.

Die dritte Spalte der Tabelle legt nahe, daß und wie die Befolgung und Umsetzung der Teilziele im Gutachten auch nachvollziehbar dokumentiert werden sollten. Ein Gutachter, der im einführenden Teil des Gutachtens zu erkennen gibt, daß er für gezielte, dem Stand der psychologischen Forschung entsprechende Modelle sensibel ist und alternative Modelle auf logisch stichhaltige Weise ausschließt, wird somit in den meisten Fällen deutlich von einem konzeptionslosen Gutachter zu unterscheiden sein, der in rigider Weise (und oft über viele heterogene Fälle hinweg) immer an derselben Routine festhält. Diese Unterschiede werden für ein mögliches Obergutachten maßgeblich und aufschlußreich sein.

Die Explikation der Modellannahmen stellt ein so prominentes Ziel einer sachgerechten und wissenschaftlich adäquaten Begutachtung dar, daß ein ausreichender Teil des Gutachtens diesem Ziel gewidmet sein sollte. Das heißt, zu Beginn eines Gutachtens sollte genügend Raum für die explizite Planung und Beschreibung der diagnostischen Vorgehensweise gewidmet werden. Dies schließt, wie die Tabelle zeigt, neben der Explikation der Modelle vor allem die Beschreibung und Begründung der Verfahren (Tests, Indikatoren, Beobachtungen) ein, die eine angemessene Prüfung der Modelle ermöglichen sollen. Eigens aufgeführt ist auch das Gebot, bei der Begründung des diagnostischen Verfahrens den theoretischen oder Erfahrungshintergrund anzugeben, aus dem der Gutachter seine Vorgehensweise ableitet. Hiermit ist nichts Unrealistisches gemeint! Freilich ist nicht gemeint, daß jedes Gutachten eine wissenschaftliche Originalarbeit sein muß. Dennoch meinen wir ganz entschieden, daß Referenzen auf relevante wissenschaftliche Literatur oder praxisbezogene Erfahrungen die Norm sein sollten. Für die Evaluation von Gutachten ist es eminent wichtig zu sehen, welche Quellen ein Gutachter verwendet, ob er sich fortbildet, ob er die nötigen Kenntnisse vor allem in Gedächtnispsychologie besitzt. Schon wenige Referenzen können hierfür sehr nützlich sein.

*Weitere Erläuterungen zur Tabelle:*
Bei jedem Zeugenbericht geht es letztendlich darum zu prüfen, wie hoch der Anteil an realer Erlebnisgrundlage für das berichtete Ereignis ist. Zu den in diesem Zusammenhang zu prüfenden Grundvoraussetzungen gehören neben simplen Prüfungen der Wahrnehmungsfähigkeiten unter den gegebenen Begleitumständen (Lichtverhältnisse, Sehschärfe, Blickwinkel, Dauer der Beobachtung) auch die Suche nach psychologischen Bedingungen, die die Wahrnehmung einengen können, wie z.B. bei Tatzeugen, die gleichzeitig Opfer sind und mit einer Waffe bedroht wurden (hier kennt man beispielsweise den sog. "Waffen-Fokus", d.h. die Person konzentriert sich so intensiv auf die Waffe, daß beispielsweise äußere Kennzeichen des Täters in den Hintergrund treten können (Cutler, Penrod & Martens, 1987; Maass & Köhnken, 1989). Die Rede ist hier lediglich von einem "Anteil an realer Erlebnisgrundlage", weil ein erlebtes Ereignis sowohl hinsichtlich seiner Wahrnehmung wie auch seiner Interpretation einer psychischen Bearbeitung unterliegt. Beides wird durch die Erwartungen der beobachtenden Person beeinflußt und mit vorhandenen Schemata zur Deckung gebracht. Die Frage, die hier zu prüfen ist, lautet demnach: „Gegeben das Ereignis hat sich wie berichtet zugetragen, wie objektiv war die Beobachtung?"

In dem Zeitraum zwischen dem Erlebnis und dem Erlebnisbericht wird ein Teil des Erlebnisses vergessen, d.h. ein Zugriff ist nicht mehr möglich. Gleichzeitig werden diese Lücken auch teilweise (und ohne aktives Zutun der befragten Person) wieder geschlossen, indem wiederum Schemata und Skripte über typische Ereignisabläufe herangezogen werden. Zusätzlich können neue Elemente eingebaut werden, die erst in Befragungen von den ermittelnden Personen quasi angeboten werden. Um diese drei Fehlerquellen: a) Vergessen, b) Rekonstruktion und c) Suggestion bestimmen zu können, ist die Berücksichtigung von **Gedächtnismodellen** unumgänglich.

Sowohl zum Zeitpunkt des Erlebnisses wie auch zum Zeitpunkt der Befragung muß eine erhebliche emotionale Belastung mitbedacht werden. Aus dieser ergeben sich sowohl für die Beobachtung wie für die Speicherung des Ereignisses bzw. seinen Abruf und schließlich für die Kommunikation der Erinnerung gegenüber der befragenden Person Auswirkungen, die ebenfalls unter Zuhilfenahme **emotionaler Modelle** diskutiert werden müssen.

Eine **Täuschungsabsicht** wäre ein weiteres zu prüfendes Modell. Auch die bewußt falsche Schilderung basiert auf Erinnerungen und Rekonstruktionen, allerdings wird anstelle einer Erlebnisgrundlage, die zum behaupteten Zeitpunkt entstand, eine andere Vorlage genutzt und mit Aspekten des aktuellen Falles verbunden. Ein typisches Beispiel ist das falsche Alibi, bei dem sämtliche berichtete Aspekte bis auf den angegebenen Zeitpunkt durchaus auf realen Erlebnissen beruhen können. Dem zu prüfenden Modell der absichtlichen Täuschung sind daher die Gedächtnis- und emotionalen Modelle

stimmte Fehlfunktionen zeigt. Dieses Modell schließt als Spezialfälle unter anderem eine Wahrnehmungshypothese (der Zeuge unterlag einer Wahrnehmungstäuschung), eine Persönlichkeitshypothese (der Zeuge ist psychisch krank) und etwa eine Gedächtnishypothese ein (es handelt sich um fehlerhaftes Gedächtnis). Innerhalb einer solchen Gedächtniskonzeption kann man dann wiederum noch feinere Differenzierungen vornehmen. Die Gedächtnistäuschung kann auf Vergessen beruhen oder auf Suggestion von außen oder auch auf motivierten Prozessen (Verdrängen unerträglicher Inhalte). Wichtig ist, daß man in einer solchen Hierarchie von Modellen oder Hypothesen nicht spezielle Modelle testen darf, bevor man allgemeinere, übergeordnete Modelle betrachtet und alternative ausgeschlossen hat. In aller Regel wird dabei ein allgemeines Gedächtnismodell ein logisches Primat haben gegenüber spezielleren, logisch untergordneten Modellen wie pathologische Gedächtnisschwäche oder motiviertes Verdrängen.

Die dritte Spalte der Tabelle legt nahe, daß und wie die Befolgung und Umsetzung der Teilziele im Gutachten auch nachvollziehbar dokumentiert werden sollten. Ein Gutachter, der im einführenden Teil des Gutachtens zu erkennen gibt, daß er für gezielte, dem Stand der psychologischen Forschung entsprechende Modelle sensibel ist und alternative Modelle auf logisch stichhaltige Weise ausschließt, wird somit in den meisten Fällen deutlich von einem konzeptionslosen Gutachter zu unterscheiden sein, der in rigider Weise (und oft über viele heterogene Fälle hinweg) immer an derselben Routine festhält. Diese Unterschiede werden für ein mögliches Obergutachten maßgeblich und aufschlußreich sein.

Die Explikation der Modellannahmen stellt ein so prominentes Ziel einer sachgerechten und wissenschaftlich adäquaten Begutachtung dar, daß ein ausreichender Teil des Gutachtens diesem Ziel gewidmet sein sollte. Das heißt, zu Beginn eines Gutachtens sollte genügend Raum für die explizite Planung und Beschreibung der diagnostischen Vorgehensweise gewidmet werden. Dies schließt, wie die Tabelle zeigt, neben der Explikation der Modelle vor allem die Beschreibung und Begründung der Verfahren (Tests, Indikatoren, Beobachtungen) ein, die eine angemessene Prüfung der Modelle ermöglichen sollen. Eigens aufgeführt ist auch das Gebot, bei der Begründung des diagnostischen Verfahrens den theoretischen oder Erfahrungshintergrund anzugeben, aus dem der Gutachter seine Vorgehensweise ableitet. Hiermit ist nichts Unrealistisches gemeint! Freilich ist nicht gemeint, daß jedes Gutachten eine wissenschaftliche Originalarbeit sein muß. Dennoch meinen wir ganz entschieden, daß Referenzen auf relevante wissenschaftliche Literatur oder praxisbezogene Erfahrungen die Norm sein sollten. Für die Evaluation von Gutachten ist es eminent wichtig zu sehen, welche Quellen ein Gutachter verwendet, ob er sich fortbildet, ob er die nötigen Kenntnisse vor allem in Gedächtnispsychologie besitzt. Schon wenige Referenzen können hierfür sehr nützlich sein.

*Weitere Erläuterungen zur Tabelle:*
Bei jedem Zeugenbericht geht es letztendlich darum zu prüfen, wie hoch der Anteil an realer Erlebnisgrundlage für das berichtete Ereignis ist. Zu den in diesem Zusammenhang zu prüfenden Grundvoraussetzungen gehören neben simplen Prüfungen der Wahrnehmungsfähigkeiten unter den gegebenen Begleitumständen (Lichtverhältnisse, Sehschärfe, Blickwinkel, Dauer der Beobachtung) auch die Suche nach psychologischen Bedingungen, die die Wahrnehmung einengen können, wie z.B. bei Tatzeugen, die gleichzeitig Opfer sind und mit einer Waffe bedroht wurden (hier kennt man beispielsweise den sog. "Waffen-Fokus", d.h. die Person konzentriert sich so intensiv auf die Waffe, daß beispielsweise äußere Kennzeichen des Täters in den Hintergrund treten können (Cutler, Penrod & Martens, 1987; Maass & Köhnken, 1989). Die Rede ist hier lediglich von einem "Anteil an realer Erlebnisgrundlage", weil ein erlebtes Ereignis sowohl hinsichtlich seiner Wahrnehmung wie auch seiner Interpretation einer psychischen Bearbeitung unterliegt. Beides wird durch die Erwartungen der beobachtenden Person beeinflußt und mit vorhandenen Schemata zur Deckung gebracht. Die Frage, die hier zu prüfen ist, lautet demnach: „Gegeben das Ereignis hat sich wie berichtet zugetragen, wie objektiv war die Beobachtung?"

In dem Zeitraum zwischen dem Erlebnis und dem Erlebnisbericht wird ein Teil des Erlebnisses vergessen, d.h. ein Zugriff ist nicht mehr möglich. Gleichzeitig werden diese Lücken auch teilweise (und ohne aktives Zutun der befragten Person) wieder geschlossen, indem wiederum Schemata und Skripte über typische Ereignisabläufe herangezogen werden. Zusätzlich können neue Elemente eingebaut werden, die erst in Befragungen von den ermittelnden Personen quasi angeboten werden. Um diese drei Fehlerquellen: a) Vergessen, b) Rekonstruktion und c) Suggestion bestimmen zu können, ist die Berücksichtigung von **Gedächtnismodellen** unumgänglich.

Sowohl zum Zeitpunkt des Erlebnisses wie auch zum Zeitpunkt der Befragung muß eine erhebliche emotionale Belastung mitbedacht werden. Aus dieser ergeben sich sowohl für die Beobachtung wie für die Speicherung des Ereignisses bzw. seinen Abruf und schließlich für die Kommunikation der Erinnerung gegenüber der befragenden Person Auswirkungen, die ebenfalls unter Zuhilfenahme **emotionaler Modelle** diskutiert werden müssen.

Eine **Täuschungsabsicht** wäre ein weiteres zu prüfendes Modell. Auch die bewußt falsche Schilderung basiert auf Erinnerungen und Rekonstruktionen, allerdings wird anstelle einer Erlebnisgrundlage, die zum behaupteten Zeitpunkt entstand, eine andere Vorlage genutzt und mit Aspekten des aktuellen Falles verbunden. Ein typisches Beispiel ist das falsche Alibi, bei dem sämtliche berichtete Aspekte bis auf den angegebenen Zeitpunkt durchaus auf realen Erlebnissen beruhen können. Dem zu prüfenden Modell der absichtlichen Täuschung sind daher die Gedächtnis- und emotionalen Modelle

logisch vorgeordnet.

Eine Prüfung der **Glaubwürdigkeit als Persönlichkeitsmodell** zieht hingegen Berichte über nicht-tatbezogene reale und fiktive frühere Ereignisse zum Vergleich heran, um zu prüfen, ob die Person zu (absichtlichen oder unabsichtlichen) Rekonstruktionen neigt. Aus einer solchen Neigung kann jedoch kein unmittelbarer Schluß auf den Realitätsgehalt der kritischen Aussage gemacht werden, sondern lediglich ein Hinweis darauf entnommen werden, welche spezifischen, Hypothesen weiter verfolgt werden müssen.

Diese Aufstellung zu prüfender Modelle verweist darauf, daß ein Verständnis von "Glaubwürdigkeit" im Sinne einer Abwesenheit von bewußter Verfälschung wesentlich zu kurz greift. Das Ziel der Prüfung ist immer der Realitätsgehalt der Aussage. Dies geschieht durch die Bildung kontrastierender Hypothesen, die logisch und psychologisch eine Prüfung der Modelle anhand diagnostischer Daten gestatten.

*Referentialität:* Für die Begründung der Modellannahmen sollte a) der aktuelle Wissensstand der Psychologie herangezogen und mit Quellen belegt werden. Es ist b) darüber hinaus vertretbar, zusätzlich Modellannahmen aus dem Erfahrungshintergrund des jeweiligen Sachverständigen zu entwickeln. Entsprechend ist auch diese individuelle Erfahrung nach Umfang und Quelle zu belegen. Grundsätzlich muß für den Leser eines Gutachtens die Quelle (Referenz) jeder Modellannahme zweifelsfrei ersichtlich sein. Belege durch ‚Allgemeinwissen' genügen diesen Vorgaben nicht, da das sogenannte Allgemeinwissen (auch als ‚Alltagspsychologie' bezeichnet), einen von Thematik zu Thematik unterschiedlichen Grad an Realgrundlage besitzt und ein gesellschaftlicher Konsens über Wahrscheinlichkeiten weder eine empirische Prüfung noch die spezifische individuelle Erfahrung des Sachverständigen ersetzen kann.

*Selektionsentscheidungen:* Aus dem individuellen Fall ergibt sich, welche Modellannahmen geprüft werden können (und müssen). Für den Sachverständigen ergibt sich daraus die Aufgabe, eine Selektion relevanter Hypothesen vorzunehmen und diese Selektion wiederum zu begründen, sei es durch theoretische Fundierung oder auch durch pragmatische Verfügbarkeit entsprechender Daten. Die Vorgabe, mit kontrastierenden Hypothesen zu arbeiten, bedingt auch, daß im Fazit des Gutachtens explizite Gründe für die Zurückweisung alternativer Erklärungen für das Zustandekommen der Aussage genannt werden.

Die Verfahren zur Prüfung der jeweiligen Modellannahmen müssen den Kriterien wissenschaftlicher Untersuchungsmethoden genügen. Dies bedeutet auch, daß ein Sachverständiger über den aktuellen Forschungsstand in der Testdiagnostik (für Tests mit forensischen Einsatzmöglichkeiten) informiert

sein muß und nach besten Möglichkeiten dafür zu sorgen hat, verbesserte Verfahren, sobald sie vorliegen, auch einzusetzen.

### 4.2 Sicherung der diagnostischen Beobachtungen

Welche operationalen Maßnahmen geeignet sind, um die diagnostischen Beobachtungen zu sichern und im Bezugsrahmen eines bestimmten Modells nach dem heutigen Stand der Methodologie zu prüfen, ist in einer weiteren Tabelle zusammengefaßt. Die Tabelle enthält wiederum Hinweise darauf, wie die Berücksichtigung dieser Kriterien im Gutachten dokumentiert werden kann.

**Beobachtung und Interpretation der Untersuchungsbefunde:**

| *Problem* | *Mögliche Operationalisierung* | *Nachweis im Gutachten* |
|---|---|---|
| Objektivität | a) Annahme der Quasi-Objektivität bestimmter Daten | a) Begründen |
| | b)Urteiler-Übereinstimmung ermittelt | b) Explizit angeben und bewerten |
| | c) Vergleichsmaßstäbe für die Bewertung und Quantifizierung relevanter Beobachtungen | c) Psychometrische oder praxis-bezogene Grundlage angeben |
| | d) Professionelle Durchführung | d) Vollständige Beschreibung der Prozedur. Protokoll aller wesentlichen und auf Anfrage Bereithaltung aller registrierfähigen Originaldaten |
| Reliabilität | a) Verwendung standardisierter Tests | a) Explizit angeben |
| | b) Interne Konsistenz aus hinreichend vielen aktuellen Indikatoren ermitteln | b) Verfahren und Resultat mitteilen |
| | c) Aktuelle Replikation kritischer Messungen | c) Verfahren mitteilen |

logisch vorgeordnet.

Eine Prüfung der **Glaubwürdigkeit als Persönlichkeitsmodell** zieht hingegen Berichte über nicht-tatbezogene reale und fiktive frühere Ereignisse zum Vergleich heran, um zu prüfen, ob die Person zu (absichtlichen oder unabsichtlichen) Rekonstruktionen neigt. Aus einer solchen Neigung kann jedoch kein unmittelbarer Schluß auf den Realitätsgehalt der kritischen Aussage gemacht werden, sondern lediglich ein Hinweis darauf entnommen werden, welche spezifischen, Hypothesen weiter verfolgt werden müssen.

Diese Aufstellung zu prüfender Modelle verweist darauf, daß ein Verständnis von "Glaubwürdigkeit" im Sinne einer Abwesenheit von bewußter Verfälschung wesentlich zu kurz greift. Das Ziel der Prüfung ist immer der Realitätsgehalt der Aussage. Dies geschieht durch die Bildung kontrastierender Hypothesen, die logisch und psychologisch eine Prüfung der Modelle anhand diagnostischer Daten gestatten.

*Referentialität:* Für die Begründung der Modellannahmen sollte a) der aktuelle Wissensstand der Psychologie herangezogen und mit Quellen belegt werden. Es ist b) darüber hinaus vertretbar, zusätzlich Modellannahmen aus dem Erfahrungshintergrund des jeweiligen Sachverständigen zu entwickeln. Entsprechend ist auch diese individuelle Erfahrung nach Umfang und Quelle zu belegen. Grundsätzlich muß für den Leser eines Gutachtens die Quelle (Referenz) jeder Modellannahme zweifelsfrei ersichtlich sein. Belege durch ‚Allgemeinwissen' genügen diesen Vorgaben nicht, da das sogenannte Allgemeinwissen (auch als ‚Alltagspsychologie' bezeichnet), einen von Thematik zu Thematik unterschiedlichen Grad an Realgrundlage besitzt und ein gesellschaftlicher Konsens über Wahrscheinlichkeiten weder eine empirische Prüfung noch die spezifische individuelle Erfahrung des Sachverständigen ersetzen kann.

*Selektionsentscheidungen:* Aus dem individuellen Fall ergibt sich, welche Modellannahmen geprüft werden können (und müssen). Für den Sachverständigen ergibt sich daraus die Aufgabe, eine Selektion relevanter Hypothesen vorzunehmen und diese Selektion wiederum zu begründen, sei es durch theoretische Fundierung oder auch durch pragmatische Verfügbarkeit entsprechender Daten. Die Vorgabe, mit kontrastierenden Hypothesen zu arbeiten, bedingt auch, daß im Fazit des Gutachtens explizite Gründe für die Zurückweisung alternativer Erklärungen für das Zustandekommen der Aussage genannt werden.

Die Verfahren zur Prüfung der jeweiligen Modellannahmen müssen den Kriterien wissenschaftlicher Untersuchungsmethoden genügen. Dies bedeutet auch, daß ein Sachverständiger über den aktuellen Forschungsstand in der Testdiagnostik (für Tests mit forensischen Einsatzmöglichkeiten) informiert

sein muß und nach besten Möglichkeiten dafür zu sorgen hat, verbesserte Verfahren, sobald sie vorliegen, auch einzusetzen.

### 4.2 Sicherung der diagnostischen Beobachtungen

Welche operationalen Maßnahmen geeignet sind, um die diagnostischen Beobachtungen zu sichern und im Bezugsrahmen eines bestimmten Modells nach dem heutigen Stand der Methodologie zu prüfen, ist in einer weiteren Tabelle zusammengefaßt. Die Tabelle enthält wiederum Hinweise darauf, wie die Berücksichtigung dieser Kriterien im Gutachten dokumentiert werden kann.

**Beobachtung und Interpretation der Untersuchungsbefunde:**

| *Problem* | *Mögliche Operationalisierung* | *Nachweis im Gutachten* |
|---|---|---|
| Objektivität | a) Annahme der Quasi-Objektivität bestimmter Daten | a) Begründen |
| | b)Urteiler-Übereinstimmung ermittelt | b) Explizit angeben und bewerten |
| | c) Vergleichsmaßstäbe für die Bewertung und Quantifizierung relevanter Beobachtungen | c) Psychometrische oder praxis-bezogene Grundlage angeben |
| | d) Professionelle Durchführung | d) Vollständige Beschreibung der Prozedur. Protokoll aller wesentlichen und auf Anfrage Bereithaltung aller registrierfähigen Originaldaten |
| Reliabilität | a) Verwendung standardisierter Tests | a) Explizit angeben |
| | b) Interne Konsistenz aus hinreichend vielen aktuellen Indikatoren ermitteln | b) Verfahren und Resultat mitteilen |
| | c) Aktuelle Replikation kritischer Messungen | c) Verfahren mitteilen |

| | | |
|---|---|---|
| Interne Validität | a) Ausschluß von nachträglicher Beeinflussung des Gedächtnisses | a) Angabe aller vorherigen Befragungen und Tests und gedächtnispsychologisch relevanter Bedingungen |
| | b) Ausschluß von Rate- oder Antworttendenzen durch gezielte Verfahren (z.B. Signalentdeckungs-analyse) | b) Deutlich machen, daß das Problem berücksichtigt wurde. Benutzte Verfahren angeben |
| | c) Ausschluß von Vergessen oder konstruktiver Gedächtnisveränderung (z.B. Gedächtnistests unter vergleichbaren Bedingungen; Kontroll-Tests für Alternativverklärungen) | c) Vollständige Dokumentation<br>Sensibilität für subtile Faktoren bei Gedächtnistäuschungen anzeigen |
| | d) Ausschluß von linguistischen Artefakten (Sprech- und Ausdrucksstil) als Alternativerklärung relevanter Beobachtungen | d) Abgrenzung epistemisch eindeutiger Hinweise auf die Erinnerung von Sachverhalten gegenüber einem kreativen, detailreichen Sprachstil. |
| | e) Kontrolle von Erwartungseffekten des Gutachters bzw. Untersuchers (falls nicht mit dem Gutachter identisch) | e) Auftraggeber und genaue Fragestellung.<br>Wurden objektive Teile der Untersuchung ohne Aktenkenntnis durchgeführt?<br>Sind die verschiedenen Verfahren operational unabhängig? |
| Konstrukt-Validität | a) Konvergente Validierung durch unabhängige Messungen derselben Merkmale | a) Kreuzvalidierung. Verweis auf interne Konsistenz. |
| | b) Divergente Validierung mit Bezug auf relevante Alternativmodelle | b) Expliziter Behandlung der Befunde über alternative Hypothesen |

*Erläuterungen zur Tabelle:*
Die in dieser Tabelle zusammengefaßten Maßnahmen zur Sicherung der Objektivität, Reliabilität und Validität von Beobachtungen sind in der Diagnostik weithin anerkannt und weniger ungewöhnlich als die zuvor betonte Forderung nach einer deutlicheren Offenlegung und Begründung der diagnostischen Modelle. Es erübrigen sich daher ausführlichere Kommentare zu den Kriterien in dieser weiteren Tabelle. Nur so viel sei deutlich hervor-

gehoben, daß auch bei der Sicherung von Beobachtungen und diagnostischen Daten eine Verbesserung des allgemein üblichen methodischen Aufwandes erwartet werden sollte und durchaus realistisch ist.

Für den Fall, daß standardisierte Tests verwendet werden, versteht es sich von selbst, daß die betreffenden Koeffizienten der Objektivität, Reliabilität und Validität genannt werden. Nur bei allgemein bekannten Tests kann dies unterbleiben. Aber auch andere Daten aus Gesprächen oder Beobachtungen können im Prinzip hinsichtlich ihrer Erfassung und Interpretation abgesichert werden, um zu vermeiden, daß subjektiv unsichere oder gar willkürliche Deutungen von diagnostischen Beobachtungen mit demselben Gewicht in den Befund eingehen wie zuverlässige und valide Daten. Häufig ist es möglich, die Übereinstimmung verschiedener Codierer (z.B. von Videoaufzeichnungen) zu bestimmen oder dieselben Daten mehrfach codieren und analysieren zu lassen. Die Kosten hierfür sind oft gering. Wenn die Art oder Menge der verfügbaren Daten eine aktuelle Bestimmung der Objektivität und Reliabilität nicht zuläßt, können Gutachter zumindest auf früher erhobene Ergebnisse (mit einem Codierverfahren oder mit bestimmten Urteilern bei ähnlichen Verfahren) verweisen. Das Problem der Sicherung von Beobachtungsdaten einfach zu ignorieren, ist indessen nicht akzeptabel (vgl. das Gutachten von Prof. Dr. K. in der Sache O.).

Hinsichtlich der Dokumentation dieser methodischen Maßnahmen im Gutachten gilt im allgemeinen, daß jede zusätzliche Angabe über den diagnostischen Wert der eingesetzten Verfahren nur von Vorteil sein kann und in der Regel wenig Raum beansprucht. Sofern der Gutachter ohnehin nach den methodischen Regeln der Kunst gearbeitet hat, verursacht diese Forderung ferner keinen zusätzlichen Zeitaufwand.

Was die Dokumentierung bzw. Protokollierung des vollständigen Materials angeht, so lautet unsere Empfehlung, hier das Prinzip der Verhältnismäßigkeit zu beherzigen. Während es sicher außer Frage steht, daß sämtliche Testwerte und quantitativen Messungen im Gutachten mitgeteilt werden müssen und daß auch aussagenanalytisch ausgewertete Gespräche wörtlich und vollständig zu protokollieren sind, wäre die extensive Dokumentation sämtlicher Gespräche und Beobachtungen bei allen Gelegenheiten und mit allen Bezugspersonen sicher kaum von Vorteil. Sämtliche Materialien in ein Gutachten einzuschließen, kann im Einzelfall gar von Nachteil sein, wenn auf diese Weise das Gutachten überladen und in seiner Lesbarkeit vermindert wird. Die Regel muß hier – wie allgemein in der Wissenschaft üblich – lauten, alle relevanten Materialien aufzubewahren und bei Bedarf vorlegen zu können. Hieraus ergibt sich auch eine Forderung nach systematischer Nutzung von zeitgemäßen technischen Verfahren (Videoaufzeichnungen, Tonband, CD ROM, Datenbank-Systeme etc. im Rahmen der gesetzlichen Möglichkeiten) der Registrierung und Speicherung.

Neben der Sicherung und Dokumentation der eigentlichen diagnostischen Beobachtungen sollte größter Wert gelegt werden auf eine **informative Beschreibung des vollständigen diagnostischen Kontexts**, soweit er für die Prüfung der Hypothesen und für die Interpretation der Daten erheblich ist. Dies bedeutet beispielsweise – mit Bezug auf die besondere Bedeutung von gedächtnispsychologischen Modellen – daß man den möglichen Einfluß von Gedächtnisprozessen auf Zeugenaussagen nur dann informiert verstehen und auswerten kann, wenn man systematisch erfaßt, wie häufig ein Zeuge zu einem bestimmten Thema befragt wurde, zu welchen Zeitpunkten dies geschah, welche Personen den Zeugen befragt haben, ob sie dieselben Fragen wiederholt und dabei insistiert haben, welches der Gesprächskontext war etc. Ohne diesen Hintergrund ist die kontextfreie Interpretation möglicher Gedächtniseinflüsse in einzelnen Gesprächen von geringem Wert. Analoges gilt für die Bedeutung des diagnostischen Kontexts und des gesamten Untersuchungsprozesses bei anderen Modellannahmen.

**Literatur**

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin, 111*, 256-274.

Anson, D.A., Golding, S.L., & Gully, K.J. (1993). Child sexual abuse allegations: Reliability of criteria-based content analysis. *Law and Human Behavior, 17*, 331-341.

Arntzen, F. (1982). *Psychologie der Zeugenaussage. Eine Einführung in die forensische Aussagepsychologie* (2. Auflage). Göttingen: Hogrefe.

Arntzen, F. (1983). *Psychologie der Zeugenaussage - System der Glaubwürdigkeitsmerkmale.* (2. überarbeitete und ergänzte Auflage). München: C.H.Beck'sche Verlagsbuchhandlung.

Bender, H.-U. (1987). *Merkmalskombinationen in Aussagen.* Tübingen: Mohr.

Bjork, R.A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A.P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185-205). Cambridge, MA: MIT Press.

Blau, G., & Katerberg, R. (1982). Agreeing responses set: Statistical nuisance or meaningful personality concept? *Perceptual and Motor Skills, 54*, 851-857.

Brickenkamp, R. (1997). *Handbuch psychologischer und pädagogischer Tests.* 2. Vollständig überarbeitete und erweiterte Auflage. Göttingen: Hogrefe.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62*, 193-217.

Ceci, S., Ross, D., & Toglia, M. (Eds.) (1989). *Perspectives in children's testimony.* New York: Springer.

Cronbach, L.J., & Gleser, G.C. (1965). *Psychological tests and personnel decisions* (2nd edition). Urbana: University of Illinois Press.

Cutler, B. L., Penrod, S. D., & Martens, T. K. (1987).Improving the reliability of eyewitness identification: Putting context into context. *Journal of Applied Psychology, 72,* 629-637.

Dahle, K.-P. (1997). Wege zu einem linguistischen "Wahrheitstest"? Perspektiven einer einzelfallexperimentellen Weiterentwicklung der Kriterienorientierten Aussageanalyse für die forensisch-psychologische Glaubwürdigkeitsdiagnostik. *Diagnostica, 43,* 3-26.

Dent, H., & Flin, R. (Eds.) (1992). Children as witnesses. Chichester: John Wiley & Sons Föderation Deutscher Psychologenvereinigungen. *Richtlinien für die Erstellung psychologischer Gutachten* (1995). Bonn: Deutscher Psychologen-Verlag.

DePaulo, B.M., Lassiter, G.D., & Stone, J.I. (1982). Attentional determinants of success at detecting deception and truth. *Personality and Social Psychology Bulletin, 8,* 273-279.

DePaulo, B.M., Kashy, D.A., Kirkendol, S.E., Wyer, M.M., & Epstein, J.A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology, 70,* 979-995.

Eddy, D.M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

Elaad, E. (1990). Detection of guilty knowledge in real-life criminal investigations. *Journal of Applied Social Psychology, 75,* 521-529.

Fiedler, K. (1999). Gutachterliche Stellungnahme zur wissenschaftlichen Grundlage der Lügendetektion mithilfe sogenannter Polygraphentests. *Praxis der Rechtspsychologie, 9*; S5-S44.

Fiedler, K., Armbruster, T., Nickel, S., Walther, E. & Asbeck, J. (1996). Constructive biases in social judgment: Experiments on the self-verification of question contents. *Journal of Personality and Social Psychology, 71,* 861-873.

Fiedler, K., & Hertel, G. (1994). Content-related schemata versus verbal-framing effects in deductive reasoning. *Social Cognition, 12,* 129-147.

Fiedler, K. & Walka, I. (1993). Training lie detectors to use nonverbal cues instead of global heuristics. *Human Communication Research, 20,* 199-223.

Fiedler, K., Walther, E., & Nickel, S. (1999). The auto-verification of social hypotheses: Stereotyping and the power of sample size. *Journal of Personality and Social Psychology, 77*; 1; 5-18.

Föderation Deutscher Psychologenvereinigungen (1988). *Richtlinien für die Erstellung Psychologischer Gutachten.* Bonn: Deutscher Psychologen Verlag.

Gigerenzer, G., & Goldstein, D.G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103,* 650-669.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instructions: Frequency Formats. *Psychological Review, 102,* 684-

704.

Johnson, M.K., Bush, J.G., & Mitchell, K.J. (1998). Interpersonal reality monitoring: Judging the sources of other people's memory.

Johnson, M.K., Hashtroudi, S., & Lindsay, D.S. (1993). Source monitoring. *Psychological Bulletin, 114*, 3-28.

Johnson, M.K., & Raye, C.L. (1981). Reality monitoring. *Psychological Review, 88*, 676-685.

Jussim, L. (1991) Social perception and social reality: A reflection-construction model. *Psychological Review, 98*, 54-73.

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94*, 211-228.

Koehler, D.J. (1991). Explanation, imagination, and confidence in judgment. *Psychological Bulletin, 110*, 499-519.

Köhnken, G., & Brockmann, C. (1988). Das Kognitive Interview: Eine neue Explorationstechnik (nicht nur) für die forensische Aussagenpsychologie. *Zeitschrift für Differentielle und Diagnostische Psychologie, 9*, 257-265.

Köhnken, G., & Maass, A. (1988). Eyewitness testimony: False alarms on biased instructions? *Journal of Applied Psychology, 73*, 363-370.

Köhnken, G., & Wegener, H. (1982). Zur Glaubwürdigkeit von Zeugenaussagen: Experimentelle Überprüfung ausgewählter Glaubwürdigkeitskriterien. *Zeitschrift für Experimentelle und Angewandte Psychologie, 29*, 92-111.

Krahé. B. & Kundrotas, S. (1992). Glaubwürdigkeitsbeurteilung bei Vergewaltigungsanzeigen: Ein aussagenpsychologisches Feldexperiment. *Zeitschrift für experimentelle und angewandte Psychologie, 39*, 598-620.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*, 480-498.

Lee, J.-W., & Yates, J.F. (1992). How quantity judgment changes as the number of cues increases: An analytical framework and review. *Psychological Bulletin, 112*, 363-377.

Loftus, E.F. (1979). *Eyewitness testimony*. Cambridge, MA: Harvard University Press.

Maass, A., & Köhnken, G. (1989). Eyewitness identification: Simulating the weapon effect. *Law and Human Behavior, 13*, 311-318.

Manstead, A.S.R., Wagner, H.L., & MacDonald, C.J. (1986). Deceptive and nondeceptive communications: Sending experience, modality, and individual abilities. *Journal of Nonverbal Behavior, 10*, 147-167.

McDowell, C.P., & Hibler, N.S. (1987). False allegations. In R.R. Hazelwood & A.W. Burgess (Eds.), *Practical aspects of rape investigations* (pp. 275-299). New York: Elsevier.

Nelson, T.O. (1992). *Metacognition: Core readings*. Boston: Allyn & Bacon.

Pyszczynski, T., & Greenberg, J. (1987). Towards an integration of cognitive and motivational perspectives on social inference: A biased hypothesis-testing model. In Berkowitz, L. (Ed.), *Advances in experimental social psychology* (Vol. 20, pp. 297-340). New York: Academic Press.

Qin, J., Quas, J. A., Redlich, A. D., & Goodman, G. S. (1997). Children's eyewitness testimony: memory development in the legal context. In N.Cowan & C. Hulme (Eds.), *The development of memory in childhood.* UK, Sussex: Psychology Press.

Ray, J.J. (1983). Reviving the problem of acquiescence response bias. *Journal of Social Psychology, 121*, 81-96.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin, 85*, 185-193.

Rosenthal, R. (1987) Judgment studies: Design, analysis, and meta-analysis. Cambridge: Cambridge University Press.

Schaaf, A., Kessler, J., Grond, M. & Fink, G. R. (1992). *Memo-Test.* Göttingen: Hogrefe.

Schwarz, N., & Sudman, S. (1994). *Autobiographic memory and the validity of retrospective reports.* New York: Springer.

Semin, G. R., Rubini, M. & Fiedler, K. (1995). The answer is in the question: The effect of verb causality on locus of explanation. *Personality and Social Psychology Bulletin, 21*, 834-841.

Semin, G.R., & Strack, F. (1980). The plausibility of the implausible: A critique of Snyder and Swann (1978). *European Journal of Social Psychology, 10*, 379-388.

Snyder, M. (1984). When belief creates reality. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 18, pp. 247-305). New York: Academic Press.

Snyder, M., & Swann, W.B. (1978). Hypothesis-testing strategies in social interaction. *Journal of Personality and Social Psychology, 36,* 1202-1212.

Spencer, J. R. & Flin, R. (1990). *The evidence of children. The law and the psychology.* London: Blackstone.

Spencer, J., Nicholsson, G., Flin, R., & Bull, R. (Eds.) (1990). *Children's evidence in legal proceedings.* Cambridge University Law Faculty.

Sporer, S. L., & Bursch, S. E. (1997). Kinder vor Gericht: Soziale und kognitive Voraussetzungen der Aussagen von Kindern. *Psychologische Rundschau, 48*, 141-162.

Sporer, S.L., Küpper, B. (1995). Realitätsüberwachung und die Beurteilung des Wahrheitsgehaltes von Erzählungen: Eine experimentelle Studie. *Zeitschrift für Sozialpsychologie, 26*, 173-193.

Squire, L.R. (1986). Mechanisms of memory. *Science, 232*, 1612-1619.

Stegmüller, W. (1969). *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosphie* (Bd. 1). Berlin: Springer.

Steller, M., & Köhnken, G. (1989). Criteria-based statement analysis. Credibility assessment of children's statements in sexual abuse cases. In D.C. Raskin (Ed.), *Psychological methods for investigation and evidence* (pp. 217-245). New York: Springer.

Steller, M., Wellershaus, P. & Wolf, T. (1992). Realkennzeichen in Kinderaussagen: Empirische Grundlagen der kriterienorientierten Aussagenanalyse. *Zeitschrift für Experimentelle und Angewandte Psychologie, 39*, 151-

170.

Szewczyk, H. (1973). Kriterien der Beurteilung kindlicher Zeugenaussagen. *Probleme und Ergebnisse der Psychologie, 46*, 47-66.

Szewczyk, H. & Littmann, E. (1989). Empirische Ergebnisse forensisch-psychologischer Begutachtungen zur Glaubwürdigkeit sexuell mißbrauchter Kinder und Jugendlicher nach einem 5-Aspekte-Modell. In J. Salzgeber, M. Stadler, G. Drechsel & C. Vogel (Eds.), *Glaubhaftigkeitsbegutachtung* (pp. 88-139). München: Profil.

Turner, R.E., Edgley, C., & Olmstead, G. (1975). Information control in conversation: Honesty is not always the best policy. *Kansas Journal of Sociology, 11*, 69-89.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Undeutsch, U. (1989). The development of statement reality analysis. In J.C. Yuille (Ed.), *Credibility assessment* (101-119). Dodrecht: Kluwer.

Walker, H.M., & Lev, J. (1953). *Statistical inference*. New York: Holt.

Wason, P.C. (1966). Reasoning. In B. Foss (Ed.), *New horizons in psychology* (pp. 135-151). London: Penguin.

Wells, G.L., & Loftus, E.F. (1991). Commentary: Is the child fabricating? Reactions to a new assessment technique. In J. Doris (Ed.), *The suggestibility of children's recollections: Implications for eyewitness testimony* (pp. 168-171). Washington, DC: American Psychological Association.

Zuckerman, M., DePaulo, B.M., & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. *Advances in Experimental Social Psychology, 14*, 1-59.

Zuckerman, M., Koestner, R., Colella, M.J., & Alton, A.O. (1984). Anchoring in the detection of deception and leakage. *Journal of Personality and Social Psychology, 47*, 301-311.

Zuckerman, M., Knee, C.R., Hodgins, H.S., & Miyake, K. (1995). Hypothesis confirmation: The joint effect of positive test strategy and acquiescence response set. *Journal of Personality and Social Psychology, 68*, 52-60.

## 5 Antworten auf die Frageliste des BGH an an die Sach verständigen in der Sache Az: 1 StR 618/98

| Frage | Antwort |
|---|---|
| I. Fragen zur Methodik bei Erstellung eines psychologischen Glaubwürdigkeitsgutachtens | |
| 1.1 Ist es erforderlich, angewandte Testverfahren hinsichtlich ihrer Indikation und Methodik zu erklären? | Ja. Für jedes Verfahren ist die Quelle (Autor, Jahr) anzuführen. Jedes Verfahren ist darüber hinaus hinsichtlich Zielsetzung, Prozedur und Gütekriterien kurz zu charakterisieren. |
| 1.2 Ist es erforderlich, die Ergebnisse der Testverfahren mitzuteilen (wenn ja, in welchem Umfang) oder genügt es, die Befunde interpretativ zu beschreiben? | Ja. Das Gebot der Transparenz bedingt, daß alle Ergebnisse der Testverfahren berichtet und (soweit möglich) mit den Testnormen verglichen werden können. |
| 2.1 Ist es ein methodischer Mangel, wenn die Sachverständige kein Wortprotokoll des Aussageberichts und der Befragung der Zeugin zur Sache erstellt? | Ja. Der Sachverständigen bietet nur ein Wortprotokoll die Möglichkeit, lückenhafte, mißverständliche und suggestive Befragungselemente im Nachhinein zu erkennen und die Schlußfolgerungen daraufhin zu korrigieren bzw. zu relativieren. |
| 2.2 Müßte ein solches Protokoll im Gutachten mitgeteilt werden? | Ja. Das Primat der größtmöglichen Transparenz gilt hier ebenso wie hinsichtlich der Ergebnisse der Testverfahren. |
| 2.3 Wie wirken sich entsprechende Mängel auf Ergebnisse und Nachprüfbarkeit des Gutachtens aus? | Der Nachvollzug der Stichhaltigkeit der Schlußfolgerungen wird verwehrt; die Tragfähigkeit der Ergebnisse und die Zulässigkeit der Zurückweisung von Alternativhypothesen kann nicht überprüft werden. |

| | |
|---|---|
| 3. Ist es erforderlich, die Dauer der Psychologischen Untersuchung festzuhalten und mitzuteilen? | Ja. Im Verlauf einer Untersuchung können Erschöpfungseffekte (Nachlassen der Motivation und der Aufmerksamkeit) auftreten, die auf die Bearbeitung der Tests und auf ihre Ergebnisse verfälschenden Einfluß nehmen. Neben der Bearbeitungsdauer ist daher auch die Reihenfolge der Tests zu vermerken. |
| 4.1 Ist es Aufgabe der Sachverständigen, insbesondere zur Analyse und Aufhellung der Entstehung und der Geschichte der Aussage alle Personen, mit denen die Zeugin über die mutmaßlichen Vorfälle gesprochen hat, informatorisch zu befragen? | Ja, im Rahmen des Möglichen und für die Modellprüfung Entscheidenden. Die Befragungen dienen nicht nur der Ermittlung etwaiger Aussage-Inkonsistenzen, sondern dienen auch der Ermittlung möglicher Gedächtnis-Intrusionen. |
| 4.2 Wären die Ergebnisse solcher Befragungen festzuhalten und im Gutachten mitzuteilen? | Ja. Hier gilt wie bei der Befragung der Zeugin selbst das Gebot der Transparenz. |
| 5.1 Welchen Wert haben Phantasieproben? | Keinen. Schlußfolgerungen auf der Basis von Phantasieproben (welche auf dem Täuschungsmodell basieren) erlauben es nicht, die Alternativhypothese des Gedächtnis-Modells auszuschließen. |
| 5.2 Ist es erforderlich, Sexualwissen und vorhandene Sexualerfahrungen zu explorieren? | Wenn für eine (bewußt oder unbewußt) verfälschte Aussage Sexualwissen Grundvoraussetzung ist, muß dies auch exploriert werden. |
| 5.3 Ist es statthaft, für die Beurteilung der Glaubwürdigkeit eines Zeugen sog. Außenkriterien heranzuziehen? | Ja, allerdings nur unter der Bedingung, daß es sich um echte Außenkriterien handelt, diese also nicht selbst Inhalt eines Sachverständigen-Gutachtens sind. |
| 6. Sind weitere Anmerkungen zur Erstattung psychologischer Glaubwürdigkeitsgutachten zu machen? | Geeignet sind alle Ergebnisse, die zur Prüfung gezielter Modellannahmen beitragen. |

| | |
|---|---|
| II Welches sind die wesentlichen sachlichen Kriterien für die Beurteilung der Glaubwürdigkeit insbesondere kindlicher und jugendlicher Zeugen? | Die Ermittlung der Glaubwürdigkeit von Zeugen, die im Kindes- oder Jugendalter sind, gehört zu den komplexesten Fragestellungen im forensischen Bereich. Insbesondere der Entwicklungsaspekt hat uns dazu bewogen, zu dieser Frage nicht Stellung zu nehmen und stattdessen auf einige theoretische und empirische Übersichten zu dieser Thematik zu verweisen (Ceci, Ross & Toglia, 1989; Dent & Flin, 1992; Qin, Quas, Redlich & Goodman, 1997; Spencer & Flin, 1990; Spencer, Nicholsson,, Flin, & Bull, 1990, Sporer & Bursch, 1997). |

*Anschrift der Verfasser:*
Prof. Dr. Klaus Fiedler und PD Dr. Jeannette Schmid
Psychologisches Institut
Universität Heidelberg
Hauptstraße 47-51
69117 Heidelberg