Universität Potsdam

Barbara Krahé, Silvia Kundrotas

Glaubwürdigkeitsbeurteilung bei Vergewaltigungsanzeigen : ein aussagenanalytisches Feldexperiment

first published in:
Zeitschrift für experimentelle und angewandte Psychologie, 39 (1992) 4, S. 598-620

Postprint published at the Institutional Repository of Potsdam University:
In: Postprints der Universität Potsdam
Humanwissenschaftliche Reihe ; 77
http://opus.kobv.de/ubp/volltexte/2009/3430/
http://nbn-resolving.de/urn:nbn:de:kobv:517-opus-34300

Postprints der Universität Potsdam
Humanwissenschaftliche Reihe ; 77

Glaubwürdigkeitsbeurteilung bei Vergewaltigungsanzeigen: Ein aussagenanalytisches Feldexperiment

Barbara Krahé und Silva Kundrotas*)

Institut für Psychologie der Freien Universität Berlin

Die Studie untersucht die Frage, ob sich anhand eines bereits bei kindlichen Zeugenaussagen bewährten Katalogs von Glaubhaftigkeitskriterien zwischen wahren und falschen Vergewaltigungsanzeigen unterscheiden läßt. Das Fallmaterial bestand aus dreißig authentischen Vernehmungsprotokollen, von denen je die Hälfte eingestandene Falschaussagen bzw. anhand von Tätergeständnissen oder Indizien als wahr klassifizierte Aussagen waren. Erfahrene Polizeibeamte aus dem Bereich Sexualstraftaten (N=30) beurteilten diese Fälle anhand des Kriterienkatalogs von Steller und Köhnken (1989) und gaben ein abschließendes Urteil über die Glaubwürdigkeit jedes Falles ab. Eine zweite Gruppe von Polizeibeamten (N=22) beurteilte die Fälle ohne Rückgriff auf Realkennzeichen als wahr bzw. falsch. Die diskriminanzanalytische Auswertung der Ergebnisse zeigt, daß der Kriterienkatalog insgesamt gut dazu geeignet ist, zwischen wahren und falschen Vergewaltigungsanzeigen zu differenzieren. Die Ergebnisse der schrittweisen Diskriminanzanalyse zeigen darüber hinaus, welche Glaubhaftigkeitskriterien im einzelnen einen signifikanten Beitrag zur Trennung zwischen Wahr- und Falschaussagen leisten. Der höhere Prozentsatz korrekter Klassifikationen in der Gruppe, die die Realkennzeichen verwendete, erreichte allerdings nur knapp statistische Signifikanz. Insgesamt stützt die vorliegende Studie die Grundannahme der Aussagenanalyse, daß die Analyse der sprachlichen Merkmale einer Aussage einen wesentlichen Beitrag zur Beurteilung des Wahrheitsgehalts einer Anzeige leisten kann.

Die Bestimmung der Glaubwürdigkeit von Aussagen stellt ein zentrales Problem in vielen Bereichen juristischer Urteilsfindung dar. Je stärker im Rahmen gerichtlicher Entscheidungsprozesse auf Zeugenaussagen zurückgegriffen wird, desto dringlicher ist die Notwendigkeit, den Wahrheitsgehalt einer Aussage im Einzelfall möglichst zuverlässig beurteilen zu können. Be-

*) Wir danken Herrn Kriminaloberrat Richter sowie den Mitarbeiterinnen und Mitarbeitern des Dezernats VB M III der Berliner Kriminalpolizei für ihre Unterstützung.

rücksichtigt man, daß Zeugenaussagen vor allem dann Gewicht bekommen, wenn eindeutige Sachbeweise fehlen, und Zeugen vielfach durch ihre Beteiligung am fraglichen Sachverhalt ein Interesse daran haben, eine bestimmte Sichtweise des Geschehens zu vermitteln, so wird die Untersuchung der Bedingungen zuverlässiger Glaubwürdigkeitsanalysen als genuine Aufgabe forensisch-psychologischer Forschung erkennbar.

Es verwundert daher nicht, daß sich die wissenschaftliche Psychologie seit ihren Anfängen mit dem Problem der Glaubwürdigkeitsbeurteilung beschäftigt hat (vgl. Köhnken, 1990). Zentraler Anwendungsbereich dieser Forschung ist dabei der Bereich der Sexualstraftaten, insbesondere die Beurteilung kindlicher Glaubwürdigkeit in Fällen sexuellen Mißbrauchs (z. B. Arntzen, 1983 a; Michaelis-Arntzen, 1981; Raskin & Esplin, 1991; Szewczyk & Littmann, 1989). Sexualdelikte zeichnen sich zumeist durch eine ungenügende Beweislage und das Fehlen unbeteiligter Zeugen aus, so daß für die Beurteilung des Sachverhalts oft nur die einander widersprechenden Aussagen von Beschuldigten und Opfern zur Verfügung stehen (Michaelis-Arntzen, 1981; Sanders, 1980, 114).

Grundsätzlich lassen sich zwei Perspektiven zur Analyse des Wahrheitsgehalts von Aussagen unterscheiden: (1) eine auf den Aspekt der *Glaubwürdigkeit* des Aussagenden gerichtete Perspektive, die auf die Identifizierung von Persönlichkeitskorrelaten genereller wahrheitsgemäßer bzw. wahrheitswidriger Aussagebereitschaft gerichtet ist, sowie (2) eine auf die *Glaubhaftigkeit* einer konkreten Aussage gerichteten Perspektive, bei der die Analyse der stilistisch-semantischen Merkmale der Aussage selbst im Mittelpunkt steht. Während auf der Grundlage des ersten Ansatzes keine Ergebnisse gewonnen werden konnten, die eine zuverlässige Differenzierung zwischen wahren und unwahren Aussagen erlaubten, hat die zweite Perspektive eine Reihe von Forschungsansätzen hervorgebracht, die Unterschiede in der sprachlichen Darstellung wahrer und unwahrer Aussagen anhand systematischer Kataloge von Aussagemerkmalen bestimmten (zur historischen Entwicklung der forensischen Aussagepsychologie vgl. Bender, 1987; Undeutsch, 1967, 1989). Grundlage dieses Ansatzes ist die Hypothese, daß sich Berichte über selbsterlebte von solchen über erfundene Ereignisse qualitativ unterscheiden, und zwar durch

> [...] Unmittelbarkeit, Farbigkeit und Lebendigkeit, sachliche Richtigkeit und psychologische Stimmigkeit, Folgerichtigkeit der Abfolge, Wirklichkeitsnähe, Konkretheit, Detailreichtum, Originalität und — entsprechend der Konkretheit jedes Vorfalls und der individuellen Erlebnisweise eines jeden Beteiligten — individuelles Gepräge. (Undeutsch, 1967, 125 f.)

Neben diesen von Undeutsch spezifizierten Kriterien haben andere Autoren eigene Kataloge von Glaubhaftigkeitsmerkmalen einer Ausage vorgelegt (z. B. Arntzen, 1983 a; Szewczyk & Littmann, 1983; Trankell, 1971), die jedoch eine Reihe inhaltlicher Überschneidungen aufweisen. Eine Integra-

tion und Kondensierung der vorliegenden Kriterienkataloge haben Steller und Köhnken (1989) in ihrem Katalog von 19 Realkennzeichen vorgelegt, der auch Grundlage der vorliegenden Untersuchung ist. Eine Aufstellung der 19 Kennzeichen in der deutschen Fassung von Steller, Wellershaus und Wolf (1992) findet sich in Tabelle 1.

Tabelle 1
Katalog der 19 Realkennzeichen (nach Steller, Wellershaus & Wolf, 1992, 153)

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 Vorgänge
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. Deliktspezifische Aussageelemente

Demnach ist eine Aussage mit um so größerer Wahrscheinlichkeit als tatsächlich erlebt, d.h. wahr, zu betrachten, je mehr Realkennzeichen in je stärkerer Ausprägung in der Aussage aufzufinden sind.

Wie Steller et al. (1992, 156) bemerken, fanden Glaubhaftigkeitskriterien über Jahrzehnte hinweg Eingang in psychologische Sachverständigenurteile,

ohne einer systematischen Validierung unterzogen worden zu sein. Inzwischen liegen einige Studien vor, die die Validität von Glaubhaftigkeitskriterien im Sinne ihrer Differenzierungsfähigkeit zwischen selbsterlebten und erfundenen Schilderungen überprüfen. Als methodische Strategien der Validitätsprüfung bieten sich vor allem zwei Wege an, deren Vor- und Nachteile umstritten sind (z. B. Arntzen, 1983 b; Köhnken & Wegener, 1985): (1) Feldstudien, in denen authentische Zeugenaussagen aus Gerichtsakten ausgewählt und einer kriterienorientierten Glaubhaftigkeitsbeurteilung unterzogen werden (z. B. Michaelis-Arntzen, 1981), und (2) Simulationsstudien, in denen die Variable „selbsterlebter versus nicht selbsterlebter Sachverhalt" experimentell variiert wird, um so die Hypothese zu prüfen, daß die vorgegebenen Realkennzeichen in selbsterlebten Sachverhalten häufiger identifiziert werden (z. B. Steller et al., 1992).

In der Feldstudie von Szewczyk und Littmann (1989) wurde eine Stichprobe von N = 173 Glaubwürdigkeitsgutachten einer Reanalyse im Hinblick auf 4 Aspekte der speziellen Glaubwürdigkeit unterzogen, wobei in etwa zwei Drittel der Gutachten die Glaubwürdigkeit bejaht und in einem Drittel erheblich bezweifelt bzw. verneint worden war. Im vorliegenden Zusammenhang ist insbesondere der Aspekt der inhaltlichen Aussagenanalyse von Bedeutung, bei der 12 Kriterien zugrunde gelegt wurden, die weitgehend in der Zusammenstellung von Steller et al. (vgl. Tabelle 1) repräsentiert sind. Die Mehrzahl der Kriterien fand sich erwartungsgemäß häufiger in glaubhaften als in unglaubhaften Aussagen. Zwei Inhaltsaspekte, ‚ausschließliche Detailliertheit bei der Schilderung der Rahmenhandlung' und ‚global-vage Tatschilderung' traten signifikant häufiger bei unglaubhaften Aussagen auf.

In einer anderen Reanalyse von Aussageprotokollen aus Gerichtsakten zu Meineidsverfahren wertete Bender (1987) die (falschen) Aussagen verurteilter Meineidiger im Kontrast zu den ihnen gegenüberstehenden (wahren) Aussagen von Zeugen in den entsprechenden Prozessen anhand von 10 Glaubhaftigkeitskriterien aus. Im Unterschied zu anderen Autoren und zu dem Katalog von Steller et al. enthält seine Kriteriologie drei „Phantasiekriterien", bei denen eine erhöhte Auftretenshäufigkeit in Falschaussagen erwartet wird: ‚Abstraktheit', ‚übertriebene Wahrheitsbeteuerung' und ‚Freudsche Versprecher'. Bei insgesamt vier der 10 Kriterien zeigten sich die erwarteten Häufigkeitsunterschiede, darunter bei den ersten beiden Phantasiekriterien.

Während Feldstudien wie diese zwar ein hohes Maß an ökologischer Validität aufweisen, leiden sie vor allem unter dem Mangel, daß den Beurteilern die gerichtliche Bewertung der Aussage als wahr bzw. falsch bei der Anwendung der Realkennzeichen bekannt war und außerdem die Zahl der wahren Aussagen im empirisch vorgefundenen Aussagematerial die Falschaussagen

bei weitem überwiegt (zur Kritik vgl. auch Köhnken, 1990). Schließlich wird die Frage der Reliabilität des aussagenanalytischen Vorgehens im Sinne der Inter-Rater-Übereinstimmung weitgehend ausgeklammert bzw. nur unzureichend problematisiert. In der von Raskin und Esplin (1991) berichteten Feldstudie war zwar der Beurteilerin, die die Glaubwürdigkeitsanalyse vornahm, der Wahrheitscharakter der Aussagen (wahre bzw. unwahre kindliche Aussagen zu sexuellem Mißbrauch) nicht bekannt, doch wurde das gesamte Fallmaterial nur von der einen Beurteilerin untersucht, so daß über die Reliabilität der Auswertung keine Aussagen möglich sind (kritisch dazu auch Wells & Loftus, 1991).

Laborexperimente als Simulationsstudien können diese Probleme zwar vermeiden, sind jedoch andererseits mit dem Problem fraglicher externer Validität konfrontiert, insofern als die Erlebnissituation etwa von Sexualdelikten nicht experimentell hergestellt werden kann und daher ein Ausweichen auf andere Erlebnisbereiche erforderlich ist. In der Studie von Steller et al. (1992) zum Beispiel wurden Schüler aufgefordert, selbsterlebte (wahre) bzw. erfundene (unwahre) Schilderungen zu insgesamt 7 belastenden Ereignissen zu verfassen: „eine Injektion bekommen", „eine Operation als Patient erleben", „Blut abgenommen bekommen", „Zähne gezogen oder Löcher gebohrt bekommen", „von einem Tier angefallen werden", „von einem anderen Kind verhauen werden" und „einen Unfall erleiden, der eine medizinische Behandlung erfordert". Die Ergebnisse zeigen, daß 9 der berücksichtigten 18 Realkennzeichen (Kriterium 19 wurde nicht erfaßt) signifikant häufiger in wahren Aussagen aufgefunden wurden, wobei jede Aussage von drei unabhängigen Beurteilern eingeschätzt wurde. Insbesondere die Merkmale der ersten drei Merkmalsgruppen (vgl. Tabelle 1) wiesen höhere Ausprägungsgrade in wahren Schilderungen auf. Für zwei Kriterien (17 und 18) zeigten sich erwartungswidrig höhere Ausprägungsgrade in unwahren Aussagen, jedoch war der Unterschied nur für Kriterium 17 signifikant.

Bei grundsätzlicher Bestätigung der postulierten Bedeutung der Realkennzeichen für die kriterienorientierte Aussagenanalyse weisen Steller et al. (1992, 166 f.) über ihre eigenen Befunde hinaus auf eine Reihe unbeantworteter Fragen bezüglich der Leistungsfähigkeit dieses Ansatzes in der forensischen Praxis hin, die im Rahmen der vorliegenden Untersuchung aufgegriffen werden sollen.

Ziel der Untersuchung ist die Überprüfung der diagnostischen Validität des von Steller und Köhnken (1989) vorgelegten Kennzeichenkatalogs, wobei folgende bislang vernachlässigte Gesichtspunkte im Mittelpunkt stehen:

(1) Die Anwendung des Ansatzes, der bisher überwiegend bei der Beurteilung kindlicher Aussagen zu sexuellem Mißbrauch bzw. in Simulationsstudien außerhalb des forensischen Bereichs erprobt wurde, auf Aussagen erwachsener Frauen zu Vergewaltigungsdelikten;

(2) Die Verwendung authentischer polizeilicher Vernehmungsprotokolle als Aussagenmaterial, wobei der Charakter jedes Protokolls als wahre bzw. falsche Aussage durch ein unabhängiges Außenkriterium (Eingeständnis der Frau im Falle der Falschbezichtigung, Beweislage im Falle der wahren Aussagen) abgesichert ist;
(3) Der Rückgriff auf eine Beurteilerstichprobe, nämlich Polizeibeamte, die mit der Glaubwürdigkeitsbeurteilung im Kontext von Vergewaltigungsdelikten in ihrem Berufsalltag konfrontiert wird;
(4) Die Einbeziehung einer Vergleichsgruppe, die die Glaubhaftigkeit von Aussagen ohne Verwendung der Realkennzeichen beurteilt;
(5) Die diskriminanzanalytische Prüfung der relativen Bedeutung der einzelnen Realkennzeichen mit dem Ziel, bei annähernd gleicher diagnostischer Leistungsfähigkeit zu einem sparsameren Satz von Kennzeichen zu gelangen.

Untersuchungshypothese ist die aussagenanalytische Kernannahme, daß zwischen wahren und unwahren Aussagen anhand stilistisch-semantischer Merkmale der Aussage in zuverlässiger Weise differenziert werden kann.

Methode

Untersuchungsmaterial

Aus den Akten der Berliner Kriminalpolizei wurden insgesamt 30 Vernehmungsprotokolle für die Untersuchung ausgewählt. Fünfzehn Protokolle bezogen sich auf Fälle, die aufgrund der Sachlage (Tätergeständnisse bzw. Sachbeweise) als glaubwürdig klassifiziert und an die Staatsanwaltschaft weitergeleitet worden waren. Die anderen fünfzehn Protokolle bezogen sich auf Falschanzeigen, die von den anzeigenden Frauen selbst am Ende der Vernehmung als unwahr eingestanden worden waren. Alle Protokolle wurden im Wortlaut übernommen und lediglich bezüglich persönlicher Angaben anonymisiert. Die durchschnittliche Länge der Protokolle betrug 3 Schreibmaschinenseiten.

Beurteiler

Als Beurteiler des authentischen Fallmaterials nahmen insgesamt 53 Mitarbeiter der Berliner Kriminalpolizei an der Untersuchung teil, davon 24 Frauen und 29 Männer. Das Durchschnittsalter lag bei 41.4 Jahren, die durchschnittliche Zahl der Dienstjahre betrug 16.8 Jahre. Bis auf je zwei weibliche und zwei männliche Untersuchungsteilnehmer hatten alle Beurteiler Vernehmungserfahrung in bezug auf Vergewaltigungsdelikte.

Untersuchungsablauf

Die Durchführung der Erhebung fand während der Dienstzeit statt. Die Versuchspersonen wurden in zwei Gruppen eingeteilt. In der Versuchsgruppe I („mit Kriterien") schätzten die Beurteiler jeden Fall anhand der 19 Realkennzeichen von Steller und Köhnken ein und gaben abschließend ein Urteil darüber ab, ob sie die Schilderung für eine wahre Aussage oder eine Falschbezichtigung hielten. In der Versuchsgruppe II („ohne Kriterien") gaben die Beurteiler zu jedem Fall lediglich ein Urteil „wahr" bzw. „Falschbezichtigung" ab. Alle Beurteiler waren über den Wahrheitscharakter der einzelnen Protokolle uninformiert.

Da es aufgrund der Länge der Protokolle und der Komplexität der Beurteilungsaufgabe nicht möglich war, jedem Beurteiler sämtliche Fälle vorzulegen, erhielten die 31 Beurteiler der Gruppe I jeweils 4 Fälle.[1]) In der Gruppe II wurde 12 Beurteilern eine Auswahl von 5 Fällen und 10 Beurteilern eine Auswahl von 6 Fällen vorgelegt. Auf diese Weise wurde jeder Fall von 4 unabhängigen Beurteilern eingeschätzt.

Der *Gruppe I* wurde ein Fragebogen mit vier Protokollen, jeweils gefolgt von den 19 Realkennzeichen, vorgelegt. Die Zusammenstellung der Viererkombinationen erfolgte nach dem Zufallsprinzip über wahre und falsche Aussagen hinweg. Die Realkennzeichen waren in der Alltagssprache formuliert und durch kurze Beispiele erläutert, wie in Tabelle 2 dargestellt. Die Verständlichkeit der Kennzeichenbeschreibungen wurde zuvor in einer Pilotstudie überprüft.

Zur Beurteilung der Ausprägung jedes Kennzeichens in dem jeweiligen Protokoll wurde den Beurteilern eine dreistufige Skala mit den Abstufungen „nein", „ja" und „ja, sehr" vorgegeben. Nach der Beurteilung jedes Falls bezüglich der 19 Realkennzeichen gaben die Beurteiler ein globales Urteil über den Wahrheitscharakter der Aussage („wahr" vs. „Falschbezichtigung") ab und wurden aufgefordert, dieses Urteil kurz in eigenen Worten zu begründen.

In der *Gruppe II* bestand die Aufgabe der Beurteiler lediglich darin, im Anschluß an jedes Protokoll ein dichotomes „wahr/falsch"-Urteil, gefolgt von einer kurzen Begründung abzugeben.

---

[1]) Ein Beurteiler wurde wegen unvollständiger Bearbeitung des Fragebogens aus der Stichprobe ausgeschlossen.

Tabelle 2
Alltagssprachliche Erläuterungen der Realkennzeichen

1) Innere Stimmigkeit und Folgerichtigkeit

z. B.: Ist die Aussage in sich stimmig?
Können alle berichteten Einzelheiten in einen passenden und folgerichtigen Zusammenhang gebracht werden?

2) Ungeordnete Darstellung des Geschehens

z. B.: Wird das Geschehen unzusammenhängend und sprunghaft erzählt?
Sind einzelne Teile des Sachverhalts über das Vernehmungsgespräch verstreut dargestellt?
Werden Ereignisse in falscher zeitlicher Reihenfolge geschildert, ohne daß allerdings Unstimmigkeiten entstehen?

3) Die Anzahl von genauen Berichten über Einzelheiten und Nebensächlichkeiten

z. B.: Finden sich Orts- und Zeitangaben?
Werden einzelne Personen, Dinge und Ereignisse ganz exakt beschrieben?
Werden Handlungen genau Schritt für Schritt beschrieben?

4) Schilderung des Geschehens im Zusammenhang mit örtlichen oder zeitlichen Gegebenheiten oder mit bestimmten Gewohnheiten

z. B.: Werden Gewohnheiten des Tagesablaufs des Täters, des Opfers oder anderer Personen geschildert?
Werden Besonderheiten der Landschaft, des Gebäudes, der Wohnung oder des Mobiliars zur Tatzeit geschildert?
Sind die Tatumstände mit diesen Besonderheiten verflochten dargestellt?

5) Wiedergabe von Handlungsketten

z. B.: Kommen Ketten von Handlungen und Reaktionen darauf im Bericht vor?
Werden Handlungen, die sich gegenseitig bedingen und aufeinander beziehen (Aktion →Reaktion), erzählt?

6) Wiedergabe von Gesprächen

z. B.: Werden Gespräche oder Äußerungen von beteiligten Personen geschildert?
Werden eventuell sogar Mißverständnisse wiedergegeben?
Benutzt die aussagende Person für sie unpassende, eventuell ihr unbekannte Worte bei der Wiedergabe von Gesprächen?

7) Schilderungen von Schwierigkeiten im Handlungsverlauf

z. B.: Wird von unvorhergesehenen Schwierigkeiten berichtet?
Wird es durch plötzlich erscheinende Personen oder andere Umstände für den Täter nötig, die Handlung zu unterbrechen oder abzubrechen?

8) Schilderung ausgefallener Einzelheiten

z. B.: Treten in der Aussage ungewöhnliche oder seltene Einzelheiten auf, die dabei aber nicht unmöglich sind?

9) Schilderung nebensächlicher Einzelheiten

z. B.: Werden Einzelheiten erzählt, die für den Handlungsverlauf oder die Tat unwichtig oder unnötig sind?

10) Sachgerechte Schilderung von unverstandenen Teilen des Handlungsablaufs

z. B.: Werden Handlungen erzählt, die die aussagende Person in ihrer Bedeutung nicht durchschaut hat?
Werden falsche Verbindungen hergestellt, weil Teile des Geschehens nicht richtig verstanden wurden?

11) Schilderung von anderen ähnlichen Vorgängen

z. B.: Wird von Handlungen berichtet, die dem jetzigen Geschehen ähnlich sind, aber zu anderer Zeit und mit anderen Personen stattgefunden haben?
Werden Gespräche von Personen wiedergegeben, die sich auf frühere ähnliche Handlungen beziehen?

12) Schilderung von psychischen Vorgängen des Zeugen bzw. des Opfers

z. B.: Werden Gedanken oder Befürchtungen zum Geschehen berichtet?
Wird von eigener Angst, Ekel, Scham, Freude, Neugier oder anderen Gefühlen berichtet?
Erwähnt die Zeugin z. B., daß sie gezittert hat oder errötete?

13) Schilderung von psychischen Vorgängen beim Täter

z. B.: Werden vermutete Gedanken oder Gefühle des Täters oder anderer beteiligter Personen geschildert?
Wird z. B. berichtet, daß der Täter komisch schaute, zitterte oder heftig atmete?

14) Plötzliche unaufgeforderte Berichtigungen der eigenen Aussage

z. B.: Drückt sich die Frau, ohne dazu aufgefordert zu werden, deutlicher und präziser aus?
Tauchen neue, klarere Erinnerungen auf?

15) Zugeben von Erinnerungslücken

z. B.: Gibt die Frau zu, sich an bestimmte Einzelheiten nicht mehr zu erinnern?

16) Einwände gegen die Richtigkeit der eigenen Aussage

z. B.: Stellt die Frau ihre eigene Glaubwürdigkeit bzw. die Glaubhaftigkeit ihrer Aussage in Frage? (Beispiel: Kein Mensch würde mir das glauben ...)

17) Selbstbelastungen

z. B.: Berichtet die Frau Unvorteilhaftes über sich selbst bzw. ihr Verhalten?
Bemüht sich die Frau z. B. durch Selbstkritik um wahrheitsgemäße Berichterstattung?

18) Entlastung des Angeschuldigten

z. B.: Wird auf eine Belastung des Täters verzichtet, obwohl es naheliegend wäre?
Werden beteiligte Personen von der Zeugin entlastet?

19) Ungewöhnliche oder unübliche Einzelheiten bezüglich des Delikts

z. B.: Wird von ungewöhnlichen Sexualpraktiken (z. B. Perversionen) berichtet, die der Frau bisher unbekannt gewesen sein müßten?
Wird das Geschehen in einer Weise geschildert, die dem Alltagswissen über Sexualverhalten widerspricht?

Ergebnisse

Gruppe I: „Mit Kriterien"

Analyse der Antworthäufigkeiten

Im ersten Schritt der Datenanalyse wurden die Häufigkeiten der drei Antwortkategorien „nein", „ja" und „ja, sehr" für die 19 Kriterien in den wahren und falschen Aussagen ermittelt. Die Hypothese signifikanter Unterschiede in der Auftretenshäufigkeit der 19 Realkennzeichen in wahren Fällen und Falschbezichtigungen wurde für jedes Kriterium mit Hilfe des Chi^2-Tests überprüft, wobei die „ja"- und „ja, sehr"-Antworten in eine Kategorie zusammengefaßt wurden.[2]) Da es für die aussagenanalytische Beurteilung der Glaubhaftigkeit auf die positive Identifizierung der Realkennzeichen ankommt, werden in Tabelle 3 die prozentualen Häufigkeiten der „ja"- und „ja, sehr"-Antworten an der Gesamtzahl der Ratings pro Kennzeichen und Fälle dargestellt.

Wie Tabelle 3 zeigt, treten nur bei 3 der 19 Kriterien signifikante Unterschiede zwischen wahren und falschen Aussagen auf. Die Kriterien 1 und 12 wurden signifikant häufiger in wahren Fällen identifiziert, das Kriterium 8 dagegen häufiger in Falschbezichtigungen. Dies deutet darauf hin, daß Schilderungen ausgefallener Einzelheiten eher im Sinne von „Lügen-Kriterien" interpretiert wurden.

Diskriminanzanalyse

Hauptziel der Datenanalyse war die Prüfung der Hypothese, daß sich wahre und falsche Vergewaltigungsanzeigen hinsichtlich ihrer semantischen Merkmale systematisch unterscheiden und diese Unterschiede mit dem verwendeten Katalog der Realkennzeichen auffindbar sind. Zur Beurteilung der diagnostischen Leistungsfähigkeit des Kennzeichenkatalogs wurden die Daten einer Diskriminanzanalyse unterzogen (vgl. Schuchard-Ficher et al., 1982). Dieses Verfahren erlaubt zu prüfen, inwieweit sich auf der Basis des Realkennzeichenkatalogs wahre und falsche Fälle zutreffend klassifizieren lassen. In die Analyse gingen sämtliche Beurteilungen (N=4) pro Fall ein, so daß für die 30 Protokolle insgesamt 120 Beurteilungen vorlagen. Es wurden getrennte Diskriminanzanalysen für folgende zwei Gruppierungsvariablen berechnet: (1) die anhand der Aktenlage vorgegebene *A-priori-Klassifi-*

---

[2]) Dies wurde aufgrund der relativ geringen Häufigkeiten der „ja, sehr"-Kategorie notwendig, die ansonsten wegen zu geringer Zellenbesetzungen keine entsprechende Überprüfung zugelassen hätten (vgl. auch Steller, 1989). Für die 19 simultanen Chi^2-Tests wurde eine Alpha-Adjustierung des Signifikanzniveaus (vgl. Hager, 1987, 177) vorgenommen.

Tabelle 3
Identifizierung der 19 Kennzeichen in wahren Aussagen und Falschbezichtigungen (Prozentsatz der „ja"- und „ja, sehr"-Antworten)

Kriterium	wahre Aussagen	Falschbezichtigungen	p*
1	90.0	59.3	< .01
2	28.8	48.3	n.s.
3	91.7	83.1	n.s.
4	66.1	58.3	n.s.
5	86.4	74.6	n.s.
6	58.3	41.7	n.s.
7	43.3	18.3	n.s.
8	30.0	55.0	< .01
9	16.6	36.6	n.s.
10	18.3	24.1	n.s.
11	11.7	16.6	n.s.
12	74.1	35.6	< .01
13	30.0	11.7	n.s.
14	20.0	16.6	n.s.
15	25.0	40.0	n.s.
16	5.0	8.3	n.s.
17	13.3	28.8	n.s.
18	11.7	22.8	n.s.
19	6.7	13.3	n.s.
M	41.4	36.5	

* Nach Alpha-Adjustierung.

kation der Fälle als wahr bzw. falsch und (2) die abschließenden *„wahr/falsch"-Urteile der Beurteiler* für jeden Fall. Betrachtet man zunächst die A-priori-Klassifikation der Fälle als wahr bzw. falsch, so zeigt sich, daß auf der Grundlage der 19 Realkennzeichen 80.8 Prozent der Fälle ihren jeweiligen Gruppen korrekt zugeordnet werden konnten. Es ergab sich bei einem Wilks' Lambda-Wert von .5504 eine hochsignifikante Trennung der beiden Fall-Gruppen anhand der 19 Realkennzeichen ($Chi^2_{(df=19)}$ 64.79; $p < .001$).

Wenn man die „wahr/falsch"-Urteile als Gruppierungsvariable betrachtet und die Frage prüft, wie gut sich diese Klassifikation durch die Urteile auf den 19 Realkennzeichen reproduzieren läßt, so zeigt sich für insgesamt 89.2 Prozent der Fälle eine korrekte Zuordnung. Bei dieser Analyse ergab sich bei einem Wilks' Lambda-Wert von .4493 ebenfalls eine signifikante Trennung der als wahr bzw. falsch beurteilten Protokolle auf der Grundlage der Beurteilung der Fälle anhand der Realkennzeichen ($Chi^2_{(df=19)}$ 86.80; $p < .001$).

Die durchschnittliche Korrelation zwischen den 19 Kriterien deutet mit einem Wert von $r = .17$ darauf hin, daß die Kennzeichen weitgehend unabhängige Aspekte der Aussage erfassen. Die mit Abstand höchste positive Korrelation ($r = .50$) bestand zwischen den Kriterien 1 und 3 (Logische Konsistenz und quantitativer Detailreichtum), die höchste negative Korrelation ($r = -.44$) bestand zwischen den Kriterien 1 und 2 (Logische Konsistenz und Unstrukturierte Darstellung). Auch die Ergebnisse von Bender (1987) sprechen dafür, daß die einzelnen Realkennzeichen jeweils spezifische, einander kaum überlappende Aspekte der Glaubhaftigkeit erfassen.

Angesichts der großen Zahl von Kriterien stellt sich über diese Überprüfung des Gesamtkatalogs hinaus die Frage nach der relativen Bedeutung jedes einzelnen Kriteriums bzw. der optimalen Kombination von Kriterien für die Differenzierung wahrer und falscher Aussagen. Hierzu wurden in einem zweiten Schritt mehrere stufenweise Diskriminanzanalysen berechnet, und zwar für die Gruppierungsvariable „A-priori-Klassifikation" (über alle Fälle sowie getrennt für wahre und falsche Aussagen) sowie für die Gruppierungsvariable „wahr/falsch-Urteil" über alle Fälle. Bei diesem Verfahren werden die einzelnen Kennzeichen nach ihrem relativen Beitrag zur Gruppentrennung schrittweise in die Analyse eingeführt, wobei die Analyse abgebrochen wird, wenn die Hinzufügung weiterer Kriterien nicht mehr zu einem signifikanten zusätzlichen Trennbetrag führt. Für die Gruppierungsvariable „A-priori-Klassifikation" zeigte diese Analyse, daß eine Kombination von 13 Kriterien zu einem ähnlich guten Ergebnis führt wie der vollständige Kriterienkatalog. Der ermittelte Wilks' Lambda-Wert von .5582 liegt nur geringfügig über dem für den Gesamtkatalog ermittelten Wert von .5504 und zeigt eine hochsignifikante Trennung zwischen wahren und falschen Fällen an ($Chi^2_{(df=13)}$ 65.01; $p < .001$). Auch der Prozentsatz korrekt klassifizierter Fälle ist mit 79.2 Prozent gegenüber der vollständigen Kriterienliste kaum verringert. Tabelle 4 gibt Aufschluß über die einbezogenen Variablen und ihre Trennfähigkeit.

Die Ergebnisse zeigen, daß die ersten 7 Schritte der Analyse zu einer deutlichen Verbesserung der Gruppentrennung (ausgedrückt durch den abnehmenden Wert von Wilks' Lambda) führen, die anschließend einbezogenen Variablen dagegen nur noch einen geringen Beitrag zur Verbesserung der Gruppentrennung leisten.

Die getrennten stufenweisen Diskriminanzanalysen für wahre und falsche Aussagen sollen Aufschluß darüber geben, ob zur Identifizierung von wahren Aussagen und Falschbezichtigungen dieselben Kriterien diagnostisch bedeutsam sind bzw. inwieweit sich unterschiedliche Kriterien als ausschlaggebend erweisen. Die Ergebnisse der separaten Diskriminanzanalysen für wahre Aussagen und Falschbezichtigungen sind in Tabelle 5 dargestellt.

Tabelle 4
Ergebnisse der schrittweisen Diskriminanzanalyse: „A-priori-Klassifikation" der Fälle als Gruppierungsvariable

Schritt	Kriterium	Lambda	Sig.	Standard. Diskriminanzkoeffizient
1	1	.8005	.0000	.4613
2	17	.7584	.0000	−.5581
3	12	.7019	.0000	.5578
4	7	.6692	.0000	.3424
5	13	.6508	.0000	.3245
6	14	.6196	.0000	.2912
7	15	.6051	.0000	−.3498
8	11	.5920	.0000	−.2888
9	18	.5830	.0000	.3249
10	8	.5759	.0000	−.2338
11	3	.5694	.0000	−.2436
12	6	.5663	.0000	.2001
13	16	.5582	.0000	−.2084

Tabelle 5
Ergebnisse der schrittweisen Diskriminanzanalyse getrennt für wahre Aussagen und Falschbezichtigungen: „A-priori-Klassifikation" der Fälle als Gruppierungsvariable

Schritt	Kriterium	Lambda	Sig.	Standard. Diskriminanzkoeffizient
Wahre Fälle				
1	1	.6855	.0000	.6317
2	8	.5719	.0000	−.6263
3	17	.5504	.0000	.4647
4	13	.5385	.0000	−.3521
5	19	.5221	.0000	.3315
6	15	.5075	.0000	−.2773
7	7	.4976	.0000	.2052
Falschbezichtigungen				
1	1	.7436	.0000	.8604
2	10	.6892	.0000	.5495
3	17	.6373	.0000	.3642
4	7	.6231	.0000	.2640
5	14	.6075	.0000	.2585

Für die wahren Fälle erwies sich ein Katalog von 7 Kriterien als hinreichend, wobei der Wilks' Lambda-Wert von .4977 ($Chi^2_{(df=7)}$ 38.03; $p < .001$) noch deutlich niedriger ist als bei der 13-Kriterien-Lösung bei Berücksichtigung aller Fälle. Bei den Falschbezichtigungen ergab sich eine 5-Kriterien-Lösung mit einem Wilks' Lambda-Wert von .6076 ($Chi^2_{(df=5)}$ 27.65; $p < .001$). Der Prozentsatz der korrekt klassifizierten Fälle in beiden Analysen ist in Tabelle 6 dargestellt.

Tabelle 6
Ergebnisse der Klassifikationsanalyse für wahre Fälle und Falschbezichtigungen (richtige Zuordnungen in Prozent)

		A-priori-Klassifikation wahr	falsch
Diskriminanzanalytische Klassifikation	wahr	88.33	21.67
	falsch	11.67	78.33

Hier zeigt sich, daß auch bei einem stark reduzierten Kriterienkatalog ein sehr hoher Prozentsatz der Fälle korrekt klassifiziert wird. Dabei ist festzuhalten, daß die Kriterien 1, 7 und 17 sowohl bei den wahren Aussagen als auch bei den Falschbezichtigungen eine zentrale Rolle spielen. Die Kriterien 8, 13, 15 und 19 tragen lediglich bei den wahren Aussagen zur Identifizierung des Wahrheitsgehaltes bei, die Kriterien 10 und 14 dagegen lediglich bei den Falschbezichtigungen.

Für die Gruppierungsvariable „wahr/falsch-Urteil" zeigte sich, daß eine Kombination aus 8 Variablen eine beinahe ebenso gute Diskriminanzleistung erbringt wie der Gesamtkatalog. Wilks' Lambda ist mit einem Wert von .4732 ($Chi^2_{(df=8)}$ 85.29; $p < .001$) nicht wesentlich erhöht (der Vergleichswert beträgt .4493). Der Prozentsatz korrekt klassifizierter Fälle bleibt mit 88.3 Prozent gegenüber 89.2 Prozent fast unverändert. Die Ergebnisse im einzelnen sind in Tabelle 7 enthalten.

Hinsichtlich der beibehaltenen Kriterien zeigt sich, daß von den 8 Kriterien nur 5 auch in der Analyse der A-priori-Klassifikation der Fälle enthalten sind. Drei Kriterien (3, 4 und 10) tragen nur für die Variable „Urteil", nicht jedoch für die Variable „A-priori-Klassifikation" zur Trennung zwischen wahren und falschen Aussagen bei.

Diese Ergebnisse zeigen, daß es ohne nennenswerten Informationsverlust möglich ist, den ursprünglichen Katalog von 19 Realkennzeichen auf eine

Tabelle 7
Ergebnisse der schrittweisen Diskriminanzanalyse:
„wahr/falsch"-Urteil der Versuchsteilnehmer als Gruppierungsvariable

Schritt	Kriterium	Lambda	Sig.	Standard. Diskriminanzkoeffizient
1	1	.6048	.0000	.9103
2	8	.5657	.0000	−.4007
3	10	.5361	.0000	.3373
4	7	.5159	.0000	.3350
5	14	.4989	.0000	.3112
6	3	.4916	.0000	−.3160
7	4	.4776	.0000	.2126
8	5	.4732	.0000	.1493

sparsamere Anzahl zu reduzieren. Betrachtet man die Kriterien im einzelnen, so zeigt sich bei beiden Analysen die herausragende Bedeutung von Kriterium 1 (Logische Konsistenz), die auch durch die Häufigkeitsanalysen unterstrichen wird. Darüber hinaus erwiesen sich die Kriterien 7, 12, 13, 14, 15 und 17 als besonders bedeutsam für die Diskriminierung zwischen a priori klassifizierten wahren und falschen Aussagen, während die subjektiven wahr/falsch-Urteile der Beurteiler insbesondere durch die Kriterien 7, 8, 10 und 14 differenziert werden.

Die in beiden Analysen gefundene leichte Überlegenheit der Ergebnisse für die „wahr/falsch"-Urteile der Vpn als Gruppierungsvariable läßt sich als Konsistenzeffekt interpretieren, da hier sowohl Kennzeichenratings als auch Gruppierungsvariable von den Beurteilern erhoben wurden, während bei der A-priori-Klassifikation die Gruppierungsvariable über ein gegebenes Außenkriterium unabhängig von den Beurteilern definiert wurde.

Beurteiler-Übereinstimmung

Während die bisher berichteten Befunde die diagnostische Tauglichkeit des Realkennzeichenkatalogs bei der Aussagebeurteilung weitgehend bestätigen, darf ein problematischer Aspekt nicht übersehen werden. Dieser Aspekt betrifft die Frage der Beurteilerübereinstimmung bei der Anwendung der Realkennzeichen. In die oben genannten Analysen gingen pro Fall die Ratings aller vier Beurteiler ein, ein Vorgehen, das auch von Steller et al. (1992) und anderen Studien verwendet wird. Das Ausmaß der Beurteilerübereinstimmung wird in diesen Analysen nur indirekt berücksichtigt, indem eine geringere Beurteilerübereinstimmung eine größere Varianz der Be-

urteilungen bedeutet, wodurch der Nachweis eindeutiger Effekte der Kennzeichenbeurteilung erschwert wird. Zur Beurteilung der Reliabilität des Verfahrens sollte die Beurteilerübereinstimmung jedoch auch direkt überprüft werden. Nur eine hohe Übereinstimmung unabhängiger Beurteiler stellt sicher, daß die Ergebnisse der Aussagenanalyse hinreichend unbeeinflußt von spezifischen Merkmalen (Verständnis der Realkennzeichen, Verständnis der Aussage, Persönlichkeitsmerkmale etc.) einzelner Beurteiler interpretierbar sind. Da in der forensischen Praxis Glaubwürdigkeitsbeurteilungen gemeinhin von einzelnen oder wenigen Gutachtern vorgenommen werden, kommt diesem Aspekt der Reliabilität des Verfahrens hohe anwendungsbezogene Bedeutung zu. Steller (1989) berichtet zufriedenstellende Übereinstimmungen zwischen den Urteilen einer studentischen Rater-Stichprobe und Experten-Urteilen, wobei jedoch das verwendete Prozentmaß der Übereinstimmung den Einfluß der Zufallsübereinstimmung unberücksichtigt läßt. In der vorliegenden Untersuchung sind die Ergebnisse zu diesem Punkt weniger vielversprechend. Die Beurteilerübereinstimmung wurde mit Hilfe des Kappa-Koeffizienten (Fleiss, 1971) überprüft, der das Ausmaß erfaßt, in dem die durch die Rater erzielte Übereinstimmung über die nach Zufall zu erwartende Übereinstimmung hinausgeht. Der höchste Kappa-Wert lag bei .35 (Kriterium 7), der durchschnittliche Zuwachs an Übereinstimmung betrug lediglich .14 (vgl. Tabelle 8).

Tabelle 8
Kappa-Koeffizienten der Beurteilerübereinstimmung

Kriterium 1:	.1132	Kriterium 11:	.2708
Kriterium 2:	.0700	Kriterium 12:	.3051
Kriterium 3:	.0426	Kriterium 13:	.1237
Kriterium 4:	.0765	Kriterium 14:	.0502
Kriterium 5:	.0663	Kriterium 15:	.1959
Kriterium 6:	.1540	Kriterium 16:	.0246
Kriterium 7:	.3563	Kriterium 17:	.2513
Kriterium 8:	.2207	Kriterium 18:	.0764
Kriterium 9:	.1525	Kriterium 19:	.0952
Kriterium 10:	.0411		

Einschränkend ist allerdings festzuhalten, daß die Beurteiler in der vorliegenden Studie zwar eine kurze schriftliche Erläuterung der einzelnen Kennzeichen erhielten, jedoch keinem ausführlichen Rater-Training unterzogen werden konnten. (In der von Steller, 1989, berichteten Studie erhielten die Beurteiler ein 90minütiges Training in der Anwendung des Kennzei-

chenkatalogs.) Damit unterstreichen die Ergebnisse zur Inter-Rater-Reliabilität die Notwendigkeit, die jeweiligen Gutachter in einem gezielten Beurteiler-Training in der Anwendung der Realkennzeichen zu schulen. Die Vernehmungserfahrung von Polizeibeamten erwies sich nicht als ausschlaggebend für die korrekte Identifikation der Fälle. Die Korrelation zwischen der Zahl der Dienstjahre und dem Prozentsatz der korrekt identifizierten Fälle war lediglich $r = .04$.

Gruppe II: „Ohne Kriterien"

Zur Beurteilung des diagnostischen Wertes der kriterienorientierten Aussagenanalyse gehört auch der Nachweis, daß Aussagebeurteilungen anhand von Realkennzeichen zuverlässiger zwischen wahren und falschen Aussagen differenzieren als Beurteilungen ohne Rückgriff auf semantische Kriterien. Die „wahr/falsch"-Urteile der Versuchsgruppe II, die ohne Rückgriff auf die 19 Realkennzeichen abgegeben worden waren, wurden zu diesem Zweck mit den „wahr/falsch"-Urteilen der Gruppe I verglichen. Die Ergebnisse dieses Vergleichs sind in Tabelle 9 dargestellt.

Bei Verwendung der Realkennzeichen ergibt sich erwartungsgemäß ein höherer Prozentsatz korrekt identifizierter Fälle als ohne Kennzeichen, doch erreicht dieser Unterschied nur marginale statistische Signifikanz ($Chi^2_{(df=1)} = 3.21$; $p = .07$).

Tabelle 9
Prozentsatz korrekter Zuordnungen mit bzw. ohne Verwendung der Realkennzeichen

	Versuchsgruppe	
Zuordnung	mit Realkennzeichen	ohne Realkennzeichen
korrekt	74.2	63.3
falsch	25.8	36.7
N	120	120

Diskussion

Ziel der vorliegenden Untersuchung war die Anwendung der kriterienorientierten Aussagenanalyse auf die Beurteilung authentischer Vernehmungsprotokolle angezeigter Vergewaltigungen. Mit den im Rahmen eines Feldexperiments erfaßten Kriterienratings jedes Falles durch mehrere unabhängige Experten, die hinsichtlich des Wahrheitscharakters der Aussagen blind waren, mit der Definition des Wahrheitsgehalts über ein eindeutiges Außenkriterium sowie der Berücksichtigung einer gleich großen Zahl wahrer bzw. unwahrer Aussagen werden im vorliegenden Untersuchungsansatz die an bisherigen Felduntersuchungen identifizierten Mängel vermieden (vgl. Köhnken & Wegener, 1985). Gleichzeitig sind die Ergebnisse in hohem Maße ökologisch valide, insofern sie Beurteilungsprozesse untersuchen, die einen festen Bestandteil des Alltagshandelns von Polizeibeamten ausmachen.

Wie die Ergebnisse der Diskriminanzanalysen zeigen, spricht die hohe Übereinstimmung zwischen Kriterienratings und unabhängig davon definiertem Außenkriterium deutlich für die diagnostische Relevanz der Realkennzeichen. Dabei sind diejenigen Analysen, die sich auf die A-priori-Klassifikation der Fälle beziehen, theoretisch von größerer Bedeutung als die Befunde bezüglich des Rater-Urteils als Gruppierungsvariable. Hier kann die hohe Übereinstimmung zwischen Kriterienratings und abschließendem „wahr/falsch"-Urteil zumindest teilweise als Resultat intraindividueller Urteilskonsistenz interpretiert werden.

Betrachtet man die wahren Fälle und die Falschbezichtigungen getrennt, so zeigt sich, daß einige Kriterien bei beiden Kategorien von Fällen einen signifikanten Beitrag zur Identifizierbarkeit leisten, andere Kriterien dagegen speziell für wahre Aussagen bzw. für Falschbezichtigungen aussagekräftig sind. So sind die Kriterien 1, 7 und 17 (Logische Konsistenz, Schilderung von Komplikationen im Handlungsverlauf und Selbstbelastungen) für wahre und falsche Aussagen gleichermaßen relevant: Ihr Vorkommen spricht zugunsten des Wahrheitsgehaltes einer Aussage, ihr Fehlen spricht dagegen. Demgegenüber sind die Kriterien 8, 13, 15 und 19 (Schilderung ausgefallener Einzelheiten, Schilderung psychischer Vorgänge des Täters, Eingeständnis von Erinnerungslücken und deliktspezifische Aussagenelemente) nur bei der Beurteilung wahrer Fälle diagnostisch bedeutsam und erlauben es, „Treffer" (als wahr klassifizierte Aussagen) von „Verpassern" (als falsch klassifizierte Aussagen) zu trennen. Bei den Falschbezichtigungen sind an der Differenzierung von „korrekten Zurückweisungen" (als falsch klassifizierten Aussagen) und „falschem Alarm" (als wahr klassifizierten Aussagen) speziell die Kriterien 10 und 14 (Phänomengemäße Schilderung unverstandener Handlungselemente und spontane Verbesserungen der eigenen Aussage) beteiligt. Die separaten Diskriminanzanalysen für wahre und falsche

Aussagen machen damit gegenüber der Analyse über alle Fälle deutlich, daß es sich bei der Beurteilung der beiden Aussagetypen nicht um parallele Urteilsvorgänge auf der Basis eines identischen Katalogs von Kriterien handelt, sondern um unterschiedlich akzentuierte Beurteilungsprozesse. Auf der Grundlage der beiden jeweils „maßgeschneiderten" Kriterienkombinationen werden jedoch für beide Gruppen von Fällen gute Identifizierungsleistungen erreicht, wobei der Prozentsatz korrekt klassifizierter Fälle bei den Falschaussagen etwas niedriger liegt als bei den wahren Aussagen. Dieser Unterschied sollte allerdings nicht überbewertet werden, da etwa Szewczyk und Littmann (1989) eine etwas höhere Treffer-Rate bei falschen im Vergleich zu wahren Aussagen fanden und insgesamt die Klassifikationsgenauigkeit als hoch zu bezeichnen ist.

Bei der Bewertung der Ergebnisse sind jedoch nicht nur die diskriminanzanalytischen Befunde, sondern auch die Häufigkeitsanalysen im ersten Schritt der Datenauswertung zu berücksichtigen. Hier zeigte sich, daß nur 10 der 19 Kennzeichen häufiger — davon 2 signifikant häufiger — in wahren als in falschen Aussagen identifiziert wurden. Das Kriterium 8 (Schilderung ausgefallener Einzelheiten) wurde signifikant häufiger in Falschaussagen identifiziert. Dieser Befund läßt sich zu den Ergebnissen von Szewczyk und Littmann (1989) in Beziehung setzen, die die Schilderung von Einzelheiten insbesondere bei der Rahmenhandlung als charakteristisches Merkmal unglaubhafter Aussagen feststellten. In der Studie von Steller et al. (1992) ergab sich für insgesamt 9 Kriterien ein signifikant höheres Vorkommen in wahren Aussagen, Kriterium 17, „Selbstbelastungen", trat signifikant häufiger in unwahren Aussagen auf. Steller (1989, 145) hebt hervor, daß die häufige Identifizierung der Realkennzeichen in wahren Aussagen Voraussetzung für die Überprüfung der Hypothese ist, daß Beurteiler unter Verwendung der Realkennzeichen überzufällig häufig zu zutreffenden Glaubhaftigkeitseinschätzungen gelangen. Die diskriminanzanalytischen Befunde der vorliegenden Studie erfahren daher durch die Ergebnisse der Häufigkeitsanalyse insofern eine Einschränkung, als die Voraussetzung der höheren Auftretenshäufigkeit der Kriterien in wahren Aussagen nur für einen Teil der Kennzeichen erfüllt ist. Ein Grund für die uneindeutigen Ergebnisse der Häufigkeitsanalysen im Vergleich zu früheren Untersuchungen liegt sicherlich darin, daß die Polizeibeamten der vorliegenden Untersuchungen keinem systematischen Training in der Handhabung der Kriterien unterzogen wurden und damit ihr Urteilsverhalten in nicht unwesentlichem Maße durch andere, unkontrollierte Einflußvariablen bestimmt wurde. Die vorliegenden Ergebnisse können insofern nur als erster Schritt zur Bestätigung des Realkennzeichen-Katalogs bei der Beurteilung von Wahr- und Falschaussagen erwachsener Personen zum Delikt der Vergewaltigung betrachtet werden.

Mit dieser Einschränkung zeigen die Ergebnisse jedoch, daß die verwendeten Realkennzeichen ein taugliches Instrumentarium zur Differenzierung zwischen wahren und falschen Aussagen darstellen. Gleichzeitig ergeben die Befunde der schrittweisen Diskriminanzanalysen Anhaltspunkte für eine Reduktion des Kennzeichenkatalogs, die — sofern sie sich in weiteren Studien bestätigen läßt — zu einem insgesamt sparsameren Vorgehen in der Aussagenanalyse führen könnte.

Forschungsarbeiten zur Überprüfung von Kriterien zur inhaltsorientierten Glaubwürdigkeitsdiagnostik können einen wichtigen Beitrag zur Verbesserung forensischer Beurteilungsprozesse leisten, indem sie systematische Richtlinien zur Diskriminierung zwischen wahren und unwahren Aussagen vorgeben. Es ist jedoch nicht zu übersehen, daß die theoretische Untermauerung dieses stark praxisorientierten Forschungszweiges bislang fast völlig fehlt, so daß etwa Köhnken und Wegener (1985, 115) in der Entwicklung einer „empirisch fundierten Theorie der Zeugenaussage" eine zentrale Aufgabe künftiger Forschung sehen. Eine solche Theorie müßte u. a. in der Lage sein, vom bisher unverbundenen Nebeneinander von Glaubhaftigkeitskriterien zu theoretisch begründbaren Annahmen über Beziehungen zwischen Glaubhaftigkeitskriterien zu gelangen und damit dem Problem der „Elementarisierung" der Aussagenanalyse zu begegnen (Hengesch, 1989). Gleichzeitig beinhaltet die Forderung nach theoretischer Fundierung des aussagenanalytischen Vorgehens auch die Notwendigkeit, die Beziehung zwischen Aussage und anderen konstitutiven Merkmalen der Beurteilungssituation zu spezifizieren. Während die Suche nach zuverlässigen Kriterien zur Trennung zwischen wahren und falschen Aussagen als primär pragmatische Aufgabe angesehen werden kann, sind Fragen nach dem sozialen Kontext sowohl der Aussageentstehung als auch der Aussagerezeption durch den Begutachtenden im Zuge einer theoretischen Analyse der beteiligten Ausdrucks- und Eindrucksprozesse zu beantworten. Ein möglicher theoretischer Rahmen für die Glaubhaftigkeitsbeurteilung wird von Köhnken (1990, Kap. 3) skizziert, der Glaubwürdigkeit bzw. Unglaubwürdigkeit als Kommunikationsphänomen betrachtet, zu dessen Verständnis neben dem Kommunikationsinhalt selbst die Personen von Kommunikator *und* Rezipient in die Analyse einzubeziehen sind.

Bezogen auf die Vergewaltigungsproblematik zeigen McDowell und Hibler (1987) einen möglichen Ansatz auf, in dem sie von der Frage nach der funktionalen Bedeutung einer falschen Vergewaltigungsanzeige für die psychologische Situation der Frau ausgehen und Falschanzeigen im Lichte des Bedürfnisses nach Schutz des Selbstwertgefühls und positiver Selbstdarstellung interpretieren. Als allgemeiner Bezugsrahmen könnte der „sozial-konstruktivistische" Ansatz herangezogen werden, um den Prozeß der Entstehung von Aussagen und der Beurteilung ihrer Glaubhaftigkeit zu konzep-

tualisieren (Gergen, 1985; vgl. auch Hengesch, 1989). Die Kernaussage des sozial-konstruktivistischen Ansatzes wird von Gergen zusammengefaßt:

> The terms in which the world is understood are social artifacts, products of historically situated interchanges among people. From the constructionist position the process of understanding is not automatically driven by the forces of the nature, but is the result of an active, cooperative enterprise of persons in relationships. (Gergen, 1985, 267).

Es ist klar erkennbar, daß ein solcher theoretischer Hintergrund zu einem fundamentalen Umdenken in der aussagenanalytischen Forschung führen würde, insofern als die Beurteilung der Glaubhaftigkeit einer Aussage nicht länger als beurteiler-unabhängiges Merkmal der Aussage gesehen werden könnte, sondern als Ergebnis eines Prozesses der Wahrheitskonstruktion zwischen Interaktionspartnern, in den Charakteristika des Aussagenden und des Beurteilers ebenso eingehen wie die Aussage selbst. Zwischen einem solchen Ansatz einerseits und den praxisorientierten Erwartungen an eine psychologische Glaubwürdigkeitsanalyse andererseits, liegt ein breites Spektrum möglicher und notwendiger theoretischer Entwicklungen aussagenanalytischer Forschung.

Summary

This study explores the issue of whether allegations of rape can be correctly identified as true or false on the basis of a set of content criteria used successfully in previous studies with respect to children's accounts of sexual abuse. A sample of 30 authentic protocols of police interrogations of women reporting rape were selected for the study. Fifteen cases were classified as false accounts because the complainants eventually confessed to having fabricated their allegations. The remaining cases were classified as true rapes on the basis of medical and other corroborating evidence. A group of experienced police officers (N=30) judged these cases in terms of their credibility on the basis of the 19 content criteria suggested by Steller & Köhnken (1989) and gave an overall judgment of the true vs. false nature of each statement. A second group of police officers (N=22) judged the credibility of the statements without reference to these criteria. The findings from a series of discriminant function analyses showed that the set of criteria as a whole is successful in distinguishing between true and false rape allegations. The results of the stepwise discriminant function analyses reveal which criteria in particular contribute to the correct identification of true and false cases. However, the higher percentage of correct classifications in the group using the content criteria was only marginally significant compared to the second group judging the cases without reference to the criteria. Altogether, the

findings of this study support the basic tenet of statement analysis that the linguistic features of a statement contain essential clues for assessing its credibility.

Literatur

Arntzen, F. (1983 a). *Psychologie der Zeugenaussage. Systematisierung der Glaubwürdigkeitsmerkmale.* München: Beck.

Arntzen, F. (1983 b). Die Grenzen experimenteller Verfahren in der forensischen Aussagepsychologie. *Zeitschrift für Experimentelle und Angewandte Psychologie, 30,* 523—528.

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

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76,* 378—382.

Gergen, K. J. (1985). The social constructionist movement in modern psychology. *American Psychologist, 40,* 266—275.

Hengesch, G. (1989). Das Dilemma der Glaubwürdigkeitsbeurteilung. *Zeitschrift für die gesamte Strafrechtswissenschaft, 101,* 611—626.

Hager, W. (1987). Grundlagen einer Versuchsplanung zur Prüfung empirischer Hypothesen der Psychologie. In G. Lüer (Hrsg.), *Allgemeine experimentelle Psychologie* (pp. 43—264). Stuttgart: G. Fischer.

Köhnken, G. (1990). *Glaubwürdigkeit.* München: Psychologie Verlags Union.

Köhnken G. & Wegener, H. (1985). Zum Stellenwert des Experiments in der forensischen Aussagepsychologie. *Zeitschrift für Experimentelle und Angewandte Psychologie, 32,* 104—119.

McDowell, C. P. & Hibler, N. S. (1987). False allegations. In R. R. Hazelwood & A. W. Burgess (Eds.), *Practical Aspects of Rape Investigation* (pp. 275—299). New York: Elsevier.

Michaelis-Arntzen, E. (1981). *Die Vergewaltigung aus kriminologischer, viktimologischer und aussagepsychologischer Sicht.* München: Beck.

Raskin, D. C. & Esplin, P. W. (1991). Assessment of children's statements of sexual abuse. In J. Doris (Ed.), *The suggestibility of children's recollections: Implications for eyewitness testimony* (pp. 153—164). Washington, DC: American Psychological Association.

Sanders, W. B. (1980). *Rape and Woman's Identity.* Beverly Hills: Sage.

Schuchard-Ficher, C., Backhaus, K., Humme, U., Lohrberg, W., Plinke, W. & Schreiner, W. (1982). *Multivariate Analyseverfahren.* Berlin: Springer.

Steller, M. (1989). Recent developments in statement analysis. In J. C. Yuille (Ed.), *Credibility Assessment* (pp. 135—154). Dordrecht: Kluwer.

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.

Trankell, A. (1971). *Der Realitätsgehalt von Zeugenaussagen.* Göttingen: Hogrefe.

Undeutsch, U. (1967). Beurteilung der Glaubhaftigkeit von Aussagen. In U. Undeutsch (Ed.),

Handbuch der Psychologie, Bd. 11, Forensische Psychologie (pp. 26—181). Göttingen: Hogrefe.

Undeutsch, U. (1989). The development of statement reality analysis. In J. C. Yuille (Ed.), *Credibility Assessment* (pp. 101—119). Dordrecht: Kluwer.

Wells, G. L. & Loftus, E. F. (1991). Commentary: Is this 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.

Anschrift der Verfasserinnen: Prof. Dr. Barbara Krahé und Dipl.-Psych. Silva Kundrotas, Institut für Psychologie, Freie Universität Berlin, Habelschwerdter Allee 45, W-1000 Berlin 33.