Forensische Glaubhaftigkeitsbeurteilung: Experimenteller Vergleich inhaltsorientierter und psychophysiologischer Methoden

Inauguraldissertation
zur Erlangung des Akademischen Grades
eines Dr. phil.,

vorgelegt dem Fachbereich 12 Sozialwissenschaften
der Johannes Gutenberg-Universität
Mainz

von
Heinz Werner Gödert
aus Mayen

Mainz
2002

Referent: Prof. Dr. [redacted]

Korreferent: Prof. Dr. [redacted]

Tag des Prüfungskolloquiums: 27. Februar 2002

Folgenden Personen und Einrichtungen gilt mein besonderer Dank:

Herrn Professor Dr. [redacted] für die jederzeit angenehme, inspirierende und ermutigende Betreuung der Dissertation,

Herrn Professor Dr. [redacted], der mich als wissenschaftlichen Mitarbeiter einstellte, meine Arbeit für die Landesgraduiertenförderung begutachtete und somit entscheidend dazu beitrug, daß die Rahmenbedingungen für die Promotion gewährleistet waren,

der Zentralen Kommission für Graduiertenförderung der Johannes Gutenberg-Universität Mainz, die durch die Gewährung eines Stipendiums nach dem Landesgraduiertenförderungsgesetz eine zügige Abwicklung der Promotion ermöglichte,

Frau [redacted] und Herrn Professor Dr. [redacted] für die freundliche und zuvorkommende Beratung im Rahmen der Graduiertenförderung,

Frau Dipl.-Psych. [redacted] für die freundliche Unterstützung in Fragen der Graduiertenförderung und meiner Anstellung als wissenschaftlicher Mitarbeiter,

Frau Dipl.-Phys. [redacted], Frau Dr. [redacted], Herrn Dr. [redacted] und Herrn [redacted] für die fundierte technische Unterstützung,

Frau [redacted], Frau [redacted], Frau [redacted], Frau [redacted], Frau [redacted], Frau [redacted], Frau [redacted] und Herrn [redacted], die als Versuchsleiter bzw. Auswerter maßgeblich an der Datenerhebung beteiligt waren,

Herrn Professor Dr. [redacted], Herrn Dr. [redacted] und Herrn Dr. [redacted] für zahlreiche fachliche Ratschläge,

Herrn Professor Dr. [redacted] für die freundliche Bereitstellung von Trainingsmaterialien für die Auswerterschulung,

Herrn Hochschuldozent Dr. [redacted] für die große Unterstützung bei der Rekrutierung der Versuchspersonen,

Herrn [redacted] für die Durchsicht des Manuskripts

sowie den insgesamt 142 Versuchsteilnehmerinnen und -teilnehmern.

Inhaltsverzeichnis

1 Einleitung

„Immer dann, wenn Informationen entscheidungs- oder handlungsrelevant werden, die uns nicht aus eigener Wahrnehmung bekannt sind, stellt sich prinzipiell die Frage nach deren Glaubwürdigkeit.“ (Köhnken, 1990, S. 1; Hervorhebung im Original) Dieses Zitat bringt in trefflicher Weise auf den Punkt, welche enorme Bedeutung dem Konzept Glaubwürdigkeit in sämtlichen Lebensbereichen – von der Politik über Partnerschaft und Familie bis hin zur Wissenschaft – zukommt. Zugleich impliziert das Zitat die herausragende praktische Relevanz einer effizienten Glaubwürdigkeitsbeurteilung in den verschiedenen Lebensbereichen.

Umgangssprachlich bzw. im Alltag fällt der Begriff Glaubwürdigkeit zumeist im Zusammenhang mit der Frage, ob eine Information zuverlässig ist, d.h. ob sie mit den tatsächlichen Gegebenheiten übereinstimmt (vgl. Köhnken, 1990). Dieser Auffassung zufolge ist eine Information – und somit eine Aussage – immer dann unglaubwürdig, wenn sie von den Tatsachen abweicht, also auch dann, wenn die Falschinformation auf einem Irrtum beruht.

Ist die Frage nach der Kongruenz zwischen Aussagen und Tatsachen juristisch relevant, kann die „Wahrheitsfindung“ (Döhring, 1964, S. VII) zum Gegenstand eines gerichtlichen Verfahrens werden. So stellen denn auch Zeugenaussagen „das mit weitem Abstand häufigste Beweismittel“ dar, das es in der gerichtlichen Praxis zu würdigen gilt (Nack, 1982, S. 127; s. auch Schneider, 1987). Die Glaubwürdigkeitsbeurteilung obliegt im deutschen Rechtssystem grundsätzlich dem Richter, im Strafverfahren ferner der Staatsanwaltschaft und der Polizei (§ 261 Strafprozeßordnung; § 286 Zivilprozeßordnung; vgl. auch Eisenberg, 1993; Schlothauer, 1997; Rüßmann, 1997). In bestimmten Fällen (vgl. dazu Eisenberg, 1993; Schlothauer, 1997; Fischer, 1994) können vom Gericht psychologische Sachverständige hinzugezogen werden. In jedem Fall aber sollten – wie es nicht zuletzt auch von juristischer Seite gefordert wird (z.B. R. Bender & Nack, 1995; Fischer, 1994) – die mit der Vernehmung und Glaubwürdigkeitsbeurteilung betrauten Personen bei ihren Entscheidungen die Erkenntnisse der wissenschaftlichen Aussagepsychologie berücksichtigen. Die Anwendung von aussagepsychologischem Know-how zur Glaubwürdigkeitsbeurteilung ist insbesondere dann geboten, wenn keine Sachbeweise (wie z.B. Fingerabdrücke) vorliegen, die mit hoher Wahrscheinlichkeit für oder gegen die Richtigkeit von Aussagen sprechen, so etwa, wenn lediglich Aussage gegen Aussage steht.

Die **Aussagepsychologie** als **Teildisziplin der forensischen Psychologie (Rechtspsychologie)** beschäftigt sich mit der Problematik der Aussage vor Gericht und ihrer Feh-

lerquellen. Die übergeordnete praktische Zielsetzung ist die Überprüfung des **Realitätsgehalts von Aussagen** (z.B. Wegener, 1992). Dabei wird nach den zugrundeliegenden dominierenden Ursachen zwischen absichtlichen (motivationale Verursachung) und unabsichtlichen, d.h. irrtümlichen (kognitive Verursachung) Aussageverfälschungen unterschieden. Glaubwürdigkeit als Terminus der forensischen Aussagepsychologie bezieht sich ausschließlich auf den motivationalen Aspekt der Aussage („Was will eine Person aussagen?“), während für die kognitive Komponente („Was kann eine Person aussagen?“) Begriffe wie etwa „Aussagegenauigkeit“ und „Zeugentüchtigkeit“ verwendet werden (vgl. Wegener, 1992, S. 47; Steller & Volbert, 1997, S. 22).

In der forensischen Aussagepsychologie bzw. Glaubwürdigkeitsforschung wird also eine gegenüber der Umgangssprache differenziertere Auffassung von **Glaubwürdigkeit** vertreten. Die Betonung liegt hier auf dem Aspekt der Intentionalität. Glaubwürdigkeit ist somit gleichzusetzen mit Wahrheitsvorsatz bzw. Abwesenheit absichtlicher Täuschung. Mit Furedy (1986) und Köhnken (1990) läßt sich diese Sichtweise auch in kommunikationspsychologische Begriffe fassen: Glaubwürdigkeit ist demnach gewährleistet, wenn eine Person (Kommunikator) an eine andere Person (Rezipient) eine Information vermittelt, die nach Auffassung des Kommunikators zutreffend ist. Auch wenn der Kommunikator irrtümlich und somit unbeabsichtigt eine falsche Information vermittelt, liegt Glaubwürdigkeit vor. Dagegen ist Unglaubwürdigkeit gegeben, wenn ein Kommunikator einen Rezipienten absichtlich täuscht. Der Begriff **Täuschung** umfaßt grundsätzlich sämtliche Verhaltensweisen, durch die bei anderen Personen ein Eindruck erzeugt werden soll, der nach Meinung des Handelnden (Kommunikators) falsch ist (Duprat, 1903, zitiert nach Furedy, 1986, S. 684; Köhnken, 1990). Somit liegt Unglaubwürdigkeit nicht nur vor, wenn ein Kommunikator absichtlich eine falsche Information vermittelt, sondern auch, wenn er irrtümlich eine zutreffende Information weitergibt, die er subjektiv für falsch hält. Das Spektrum des Täuschungsverhaltens reicht vom nonverbalen (auch bloßes Verschweigen von Informationen) über das paraverbale (extralinguistische) bis zum verbalen Verhalten (Furedy, 1986; Köhnken, 1990). Verbale Täuschungen bezeichnet man als **Lügen**. Wie der Name schon impliziert, geht es in der forensischen Aussagepsychologie primär um die Glaubwürdigkeit verbaler Äußerungen. Dies sind in erster Linie die Aussagen von Zeugen und die Einlassungen von Beschuldigten.

Wichtig ist die sowohl in der Rechtsprechung als auch in der Aussagepsychologie getroffene Unterscheidung zwischen allgemeiner und spezieller Glaubwürdigkeit (z.B. R. Bender & Nack, 1995; Köhnken, 1990). Der Begriff **„allgemeine Glaubwürdigkeit“** bezieht sich auf die überdauernde Neigung einer Person, aufrichtig bzw. unehrlich zu sein (personenbezogene Glaubwürdigkeit im Sinne eines positiven bzw. negativen

Leumunds). Dagegen betrifft der Terminus **„spezielle Glaubwürdigkeit“** den einer konkreten Aussage zugrundeliegenden Wahrheits- bzw. Täuschungsvorsatz (aussagebezogene Glaubwürdigkeit). Um beide Konstrukte besser voneinander abzugrenzen, wird in der psychologischen Fachliteratur anstelle des Begriffs „spezielle Glaubwürdigkeit“ häufig die Bezeichnung **„Glaubhaftigkeit der Aussage“** verwendet (z.B. Greuel, Offe, Fabian, Wetzels, Fabian, Offe & Stadler, 1998; Undeutsch, 1967). Dieser Wortwahl wird auch in der vorliegenden Arbeit gefolgt.

Heute ist allgemein akzeptiert, daß Feststellungen über die allgemeine Glaubwürdigkeit einer Person keine zwingenden Schlußfolgerungen auf die Glaubhaftigkeit ihrer Aussage zu einem konkreten Sachverhalt zulassen (z.B. Köhnken, 1990; Steller & Volbert, 1997; Wegener, 1992). Zentraler Gegenstand forensisch-psychologischer „Glaubwürdigkeitsbegutachtungen“ ist daher die Analyse der Glaubhaftigkeit der Aussage, d.h. die tatbestandsbezogenen Bekundungen sowie die damit einhergehenden Verhaltensmanifestationen stehen im Mittelpunkt.

In der forensischen Psychologie wurden verschiedene Möglichkeiten zur **Glaubhaftigkeitsbeurteilung** in Betracht gezogen. Der Versuch, frei beobachtbare Ausdruckserscheinungen (**Gestik**, **Mimik**) und Merkmale des **Sprechverhaltens** (z.B. Sprechgeschwindigkeit und -fehler) als Täuschungsindikatoren nutzbar zu machen, erbrachte bislang keine hinreichend praxisrelevanten Resultate (Furedy, 1986; Vrij, 1998a). Zu praktischer Bedeutung gelangten dagegen insbesondere zwei methodische Ansätze: der inhaltsorientierte und der psychophysiologische (Steller & Volbert, 1997; vgl. auch Undeutsch, 1983b). Bei der **inhaltsorientierten Glaubhaftigkeitsbeurteilung** wird in erster Linie überprüft, inwiefern eine Aussage bestimmte inhaltliche Merkmale enthält, die für auf einer Erlebnisgrundlage basierende Schilderungen typisch sein sollen. Der Hauptanwendungsbereich dieser Methode ist die Begutachtung von Zeugenaussagen, vornehmlich der Aussagen kindlicher und jugendlicher Opferzeugen in Sittlichkeitsprozessen. Die **psychophysiologische Glaubhaftigkeitsbeurteilung** wird in erster Linie zur Begutachtung der Einlassungen von Beschuldigten eingesetzt. Anhand der während einer standardisierten Befragung oder Reizdarbietung gemessenen physiologischen Veränderungen beim Beschuldigten werden Rückschlüsse auf die Glaubhaftigkeit der Abstreitung des Tatvorwurfs gezogen.

Die **rechtliche Situation** hierzulande unterscheidet sich für beide Methoden drastisch: Während auf dem inhaltsorientierten Ansatz basierende Glaubhaftigkeitsgutachten in deutschen Gerichtssälen schon seit mehreren Jahrzehnten als Beweismittel anerkannt werden, wird der psychophysiologischen Glaubhaftigkeitsbeurteilung von der höchstrichterlichen Rechtsprechung kein ausreichender Beweiswert zugebilligt, woraus sich

ein grundsätzliches Verwertungsverbot im Bereich der Strafgerichtsbarkeit ergibt. Dies wirft die Frage auf, ob bzw. inwiefern die psychophysiologischen und inhaltsorientierten Vorgehensweisen sich in ihrem Beweiswert unterscheiden. Daher sollen in der vorliegenden Arbeit die beiden methodischen Ansätze einem direkten empirischen Validitätsvergleich unterzogen werden.

2 Hintergrund

Im folgenden werden die beiden wichtigsten Ansätze der forensischen Glaubhaftigkeitsbeurteilung, der inhaltsorientierte und der psychophysiologische, ausführlich dargestellt. Es erfolgt jeweils eine Beschreibung der methodischen Vorgehensweisen und der zugrundeliegenden theoretischen Annahmen sowie eine kritische Würdigung unter besonderer Berücksichtigung der empirischen Validitätsbefundlage.

2.1 Inhaltsorientierte Glaubhaftigkeitsbeurteilung

Die Darstellung des inhaltsorientierten Ansatzes der Glaubhaftigkeitsbeurteilung beginnt mit einer Erläuterung der zugrundeliegenden theoretischen Überlegungen, an die sich eine Beschreibung der einzelnen inhaltlichen Glaubhaftigkeitsindikatoren anschließt. Nach der Skizzierung der konkreten Vorgehensweise bei der Begutachtung und der Absteckung der praktischen Einsatzmöglichkeiten der Methode werden schließlich die Hauptkritikpunkte diskutiert, und es wird ein Überblick über den Stand der empirischen Validitätsforschung vermittelt.

2.1.1 Theoretischer Hintergrund

Die theoretische Basis der inhaltsorientierten Glaubhaftigkeitsbeurteilung bildet bis heute eine von Undeutsch im Jahre 1967 getroffene heuristische Annahme, die sog. **„Undeutsch-Hypothese“** (Begriff von Steller, 1989, S. 145): *„Aussagen über selbsterlebte faktische Begebenheiten müssen sich von Äußerungen über nicht selbsterlebte Vorgänge unterscheiden durch Unmittelbarkeit, Farbigkeit und Lebendigkeit, sachliche Richtigkeit und psychologische Stimmigkeit, Folgerichtigkeit der Abfolge, Wirklichkeitsnähe, Konkretheit, Detailreichtum, Originalität und – entsprechend der Konkretheit jedes Vorfalles und der individuellen Erlebnisweise eines jeden Beteiligten – individuelles Gepräge.“* (Undeutsch, 1967, S. 125f.; Hervorhebung durch den Verfasser).

Die in der Hypothese aufgeführten Aussagecharakteristika sollen im **intraindividuellen Vergleich** in erlebnisbasierenden Schilderungen stärker ausgeprägt sein als in erfundenen, was von Undeutsch auch als **„bessere Qualität“** (1967, S. 125) wahrheitsgemäßer Schilderungen bezeichnet wird. Undeutsch leitete diese Hypothese zum einen aus „der alltäglichen Erfahrung und dem gesunden Menschenverstand“ ab (1967, S. 126) und berief sich zum anderen auf ähnliche Überlegungen des schwedischen Psychologen

Trankell (1963, zitiert nach Undeutsch, 1967, S. 126) und des Leipziger Landgerichtsdirektors Leonhardt (1934) sowie auf sprachwissenschaftliche Befunde über formale Merkmale wahrer und erlogener Aussagen (z.B. Kainz, 1955).

Die „Undeutsch-Hypothese“ korrespondiert mit der sog. **kognitiven Theorie des Lügens**. Dieser Konzeption zufolge werden höhere Anforderungen an das Informationsverarbeitungssystem einer Person gestellt, wenn sie eine Aussage über ein komplexes Ereignis erfinden muß als wenn sie einen wahrheitsgemäßen Sachverhalt aus der Erinnerung rekonstruiert (vgl. z.B. Köhnken, 1986, 1990; Zuckerman, DePaulo & Rosenthal, 1981). Beim Erfinden einer Aussage soll nämlich ein Großteil der kognitiven Ressourcen für kreative Prozesse (Ersinnen der Aussage) und Kontrollprozesse (Vermeidung von Widersprüchen, Unterdrückung vermeintlicher nonverbaler und extralinguistischer Lügensignale) gebunden werden. Dies sollte letztlich zur Konsequenz haben, daß nicht nur für die syntaktische, sondern auch für die inhaltliche Ausgestaltung der Aussage weniger Kapazität zur Verfügung steht, so daß konfabulierte Schilderungen – ganz im Sinne der „Undeutsch-Hypothese“ – im intraindividuellen Vergleich eine schlechtere inhaltliche Qualität aufweisen sollten als Aussagen über reale Erlebnisse.

Etwaige inhaltlich-qualitative Unterschiede zwischen erlebnisbasierenden und erfundenen Aussagen lassen sich jedoch nicht nur aus den oben dargestellten kognitionspsychologischen Überlegungen herleiten, sondern sind auch aufgrund von **motivationalen Faktoren** zu erwarten (vgl. Köhnken, Schimossek, Aschermann & Höfer, 1995). Bezugnehmend auf das Konzept des „impression management“ (Tedeschi & Norman, 1985), wird angenommen, daß es sich beim Lügen um ein zielgerichtetes Verhalten handelt und daß gerade eine lügende Person dementsprechend stark motiviert ist, einen ehrlichen Eindruck zu hinterlassen, um ihre Ziele zu erreichen. Daher wird sie versuchen, jegliche (verbalen) Verhaltensweisen zu vermeiden, die gemäß verbreiteter Stereotypen als Begleiterscheinungen von Lügen gelten und somit vom jeweiligen Kommunikationspartner als Anzeichen von Unaufrichtigkeit interpretiert werden könnten, so z.B. das Zugeben von Erinnerungslücken.

2.1.2 Inhaltliche Kriterien zur Beurteilung der Glaubhaftigkeit von Aussagen

Um den in der „Undeutsch-Hypothese“ postulierten qualitativen Unterschied zwischen erlebnisbasierenden und erfundenen Aussagen für die praktische Glaubhaftigkeitsbeurteilung nutzbar machen zu können, mußten **Qualitätsmerkmale** eruiert werden, die sich gehäuft bzw. in starker Ausprägung in erlebnisbezogenen Aussagen, jedoch nie oder nur selten bzw. lediglich in schwacher Ausprägung in erfundenen Schilderungen

finden und die folglich als Indikatoren der Glaubhaftigkeit von Aussagen dienen können. Die Identifizierung solcher Indikatoren war jedoch für lange Zeit nicht etwa Gegenstand wissenschaftlicher Grundlagenforschung; vielmehr waren es in der Praxis tätige Gerichtspsychologen und Juristen, die auf der Grundlage kasuistischen Materials **Kennzeichen erlebnisbasierender Aussagen** zusammentrugen (z.B. Kasielke, 1965; Leonhardt, 1930, 1934; vgl. dazu auch Köhnken, 1990).

Eine **Systematisierung der gesammelten Kennzeichen** glaubhafter Aussagen ist ebenfalls sowohl Psychologen als auch Juristen zu verdanken. Auf psychologischer Seite sind hier v.a. **Undeutsch** (1967, 1982, 1983b, 1984), **Arntzen** (1970, 1983a, 1993), **Trankell** (1971) sowie **Szewczyk** und **Littmann** (Szewczyk, 1973; Littmann & Szewczyk, 1983; s. auch Dettenborn, Fröhlich & Szewczyk, 1984) zu nennen. Juristischerseits haben sich vornehmlich **R. Bender**, **Röder** und **Nack** (R. Bender & Nack, 1995; R. Bender, Röder & Nack, 1981), **Prüfer** (1986) sowie **H.-U. Bender** (1987) um eine Systematisierung der Merkmale glaubhafter Bekundungen verdient gemacht. An dieser Stelle sollen die diversen Kriterienkataloge nicht im einzelnen beschrieben werden. Für einen Überblick und eine kritische Gegenüberstellung der wichtigsten Kriteriologien sei auf die Übersichtsarbeiten von H.-U. Bender (1987), Köhnken (1990) und Eisenberg (1993) verwiesen.

Während sich der überwiegende Anteil der eruierten Glaubhaftigkeitsmerkmale auf den Aussageinhalt bezieht (z.B. Detailreichtum), betreffen einige Kriterien auch die Vorgeschichte der aktuellen Aussage (Aussagemotivation und -entwicklung) sowie das aussagebegleitende Verhalten (z.B. emotionaler Ausdruck) (vgl. auch Köhnken, 1990). Die von psychologischer Seite erarbeiteten Kriteriologien enthalten ausschließlich **unipolare Kriterien der Glaubhaftigkeit**: Ist ein Kriterium erfüllt, so spricht dies für eine Erlebnisbezogenheit der Aussage. Ist dagegen ein Kriterium nicht erfüllt, so darf dies nicht gleichsam als Täuschungssymptom interpretiert werden. Eine andere Sichtweise findet sich dagegen in den von juristischer Seite erarbeiteten Kriteriensystemen. Nicht nur, daß hier mitunter von „Defizientensymptomen“ (H.-U. Bender, 1987, S. 70) die Rede ist, wenn die positiv formulierten Glaubhaftigkeitskriterien in einer Aussage nicht anzutreffen sind. Darüber hinaus werden sogar **direkte Indikatoren erfundener Aussagen** („Lügensignale“ [R. Bender, 1982, S. 122; R. Bender & Nack, 1995, S. 152] bzw. „Phantasiekriterien“ [H.-U. Bender, 1987, S. 92ff.]) postuliert, bei welchen es sich sowohl um negative Ausprägungen der Glaubhaftigkeitskriterien (z.B. „Kargheit“ als Gegenpol von Detailreichtum [R. Bender & Nack, 1995, S. 152]) als auch um unipolare Täuschungsmerkmale ohne entsprechende positive (im Sinne von glaubhaft) Ausprägung handeln kann (z.B. „Freud’sches Signal“ [R. Bender & Nack, 1995, S. 152]). Auf die empirische Befundlage zu den positiv formulierten Glaubhaftigkeitskriterien wird in

Abschnitt 2.1.6 eingegangen. Bezüglich der juristischerseits postulierten „Täuschungsindikatoren“ sei bereits an dieser Stelle betont, daß sie empirisch nicht ausreichend überprüft sind. Insbesondere weil ein gehäuftes Auftreten dieser Merkmale auch in erlebnisbasierenden Bekundungen nicht ausgeschlossen werden kann, ist von der praktischen Anwendung dieser Kriterien dringend abzusehen (Arntzen, 1993; Eisenberg, 1993).

Für die Kennzeichen erlebnisbasierender Bekundungen kursieren in der Literatur unterschiedliche Bezeichnungen. Geläufige Begriffe wie **„Realitätskriterien“** (Trankell, 1971) oder **„Realkennzeichen“** (z.B. Wolf & Steller, 1997) sind insofern mißverständlich, als sie implizieren, daß von den Kriterien auf den objektiven Realitätsgehalt einer Aussage geschlossen werden könne. Wie jedoch bereits in Abschnitt 1 erläutert wurde, zielt die forensische Glaubhaftigkeitsbeurteilung im engeren Sinne nicht auf die Rekonstruktion der objektiven Realität ab, sondern es geht primär darum zu beurteilen, ob eine Schilderung eine subjektive Erlebnisgrundlage in der „Wachwirklichkeit“ des Aussagenden besitzt (Stadler, 1997, S. 61; vgl. auch Greuel et al., 1998; Hengesch, 1989). Der ebenfalls häufig verwendete Begriff **„Glaubwürdigkeitskriterien“** (z.B. Arntzen, 1993) weist dagegen eine ungünstige Konnotation zu dem persönlichkeitsorientierten Konzept der allgemeinen Glaubwürdigkeit (vgl. Abschnitt 1) auf. Um die genannten Probleme zu vermeiden, wird im folgenden der Begriff **„Glaubhaftigkeitskriterien“** verwendet.

Insbesondere die von Arntzen, Undeutsch sowie Szewczyk und Littmann aufgestellten Kriterienkataloge weisen deutliche Überschneidungen auf. Dies nahmen Steller und Köhnken (1989) zum Anlaß, auf der Grundlage der genannten Kriteriensammlungen eine integrative Kriteriologie (s. Tabelle 1) zu entwickeln, die v.a. der empirischen Validitätsüberprüfung der inhaltsorientierten Glaubhaftigkeitsbeurteilung dienen sollte (Steller & Volbert, 1997). Diese Kategorisierung, die unter der Bezeichnung **„Kriterienorientierte Aussage- bzw. Inhaltsanalyse“** (englisch: „Criteria-Based Content Analysis“) internationale Beachtung fand, enthält ausschließlich Kriterien, die sich auf den Aussageinhalt beziehen. Im folgenden werden die inhaltlichen Glaubhaftigkeitskriterien im einzelnen definiert. Eine ausführliche Beschreibung würde den Rahmen dieser Arbeit sprengen, so daß hier nur die wesentlichen Bestimmungsstücke der Kriterien genannt werden können. Für eine umfassende Erläuterung der einzelnen Kriterien sei auf die Arbeiten der Primärautoren (Arntzen, Undeutsch bzw. Szewczyk & Littmann) verwiesen. Entsprechende detaillierte bibliographische Angaben zu den einzelnen Kriterien finden sich bei Steller und Köhnken (1989), ferner bei Steller, Wellershaus und Wolf (1992) sowie bei Steller und Volbert (1997). Ausführliche Erläuterungen der Kri-

terien finden sich auch bei Greuel et al. (1998), ferner bei Krause (1997), Petersen (1997) sowie Scheinberger (1993).

Tabelle 1. Glaubhaftigkeitskriterien in der Kategorisierung von Steller und Köhnken (1989) mit Kurzbezeichnungen

Allgemeine Merkmale	
1. Logische Konsistenz	(*Konsistenz*)
2. Ungeordnet sprunghafte Darstellung	(*Unordnung*)
3. Quantitativer Detailreichtum	(*Details*)
Spezielle Inhalte	
4. Raum-zeitliche Verknüpfungen	(*Verknüpfungen*)
5. Interaktionsschilderungen	(*Interaktionen*)
6. Wiedergabe von Gesprächen	(*Gespräche*)
7. Schilderung von Komplikationen im Handlungsverlauf	(*Komplikationen*)
Inhaltliche Besonderheiten	
8. Schilderung ausgefallener Einzelheiten	(*Ausgefallenes*)
9. Schilderung nebensächlicher Einzelheiten	(*Nebensächliches*)
10. Phänomengemäße Schilderung unverstandener Handlungselemente	(*Unverstandenes*)
11. Indirekt handlungsbezogene Schilderungen	(*Indirektes*)
12. Schilderung eigener psychischer Vorgänge	(*Eigenseelisches*)
13. Schilderung psychischer Vorgänge des Angeschuldigten	(*Fremdseelisches*)
Motivationsbezogene Inhalte	
14. Spontane Verbesserungen der eigenen Aussage	(*Verbesserungen*)
15. Eingeständnis von Erinnerungslücken	(*Erinnerungslücken*)
16. Einwände gegen die Richtigkeit der eigenen Aussage	(*Selbsteinwände*)
17. Selbstbelastungen	(*Eigenbelastung*)
18. Entlastung des Angeschuldigten	(*Fremdentlastung*)
Deliktspezifische Inhalte	
19. Deliktspezifische Aussageelemente	(*Deliktspezifisches*)

Wie Tabelle 1 zu entnehmen ist, umfaßt die *Kriterienorientierte Inhaltsanalyse* von Steller und Köhnken (1989) insgesamt 19 Glaubhaftigkeitskriterien, die in fünf Kategorien unterteilt sind. In der Kategorie **„Allgemeine Merkmale“** werden Kriterien zusammengefaßt, die sich auf eine Aussage in ihrer Gesamtheit beziehen und die deshalb ohne Rekurs auf Einzelheiten des Aussageinhalts beurteilt werden können. Das Kriterium **„Logische Konsistenz“** bezieht sich auf die innere Stimmigkeit und Folgerichtigkeit bzw. auf die Widerspruchsfreiheit der Aussage, d.h. darauf, ob alle berichteten Einzelheiten in einen passenden und folgerichtigen Zusammenhang gebracht werden können. Dem zweiten Kriterium, **„Ungeordnet sprunghafte Darstellung“**, liegt die An-

nahme zugrunde, daß erfundene Schilderungen eher durch eine kontinuierliche, strukturierte, meist chronologische Darstellungsweise gekennzeichnet sind. Eine unstrukturierte Darstellung hingegen, die inverse Verlaufsstrukturen im Sinne rückwärts aufgerollter Interaktionsschilderungen und eine Vielzahl über die Gesamtaussage verstreuter Einzelangaben enthält, wird als Glaubhaftigkeitskriterium angesehen, allerdings nur, sofern die unstrukturierte Schilderung dennoch zu einem geschlossenen, logisch konsistenten Bild rekonstruiert werden kann. Ist letzteres nicht der Fall, so bleibt neben Kriterium 2 auch Kriterium 1, „Logische Konsistenz", unerfüllt. Mit dem Kriterium **„Quantitativer Detailreichtum"** wird die Anzahl präzise geschilderter Einzelheiten bewertet. Hier geht es also darum, ob sich genaue Zeit- und Ortsangaben finden, ob einzelne Personen, Gegenstände und Ereignisse minuziös beschrieben werden und ob Handlungen und Dialoge genau Schritt für Schritt wiedergegeben werden.

Die vier in der Kategorie **„Spezielle Inhalte"** zusammengefaßten Kriterien betreffen Details, die so speziell sind, daß ihr Vorkommen in erfundenen Berichten im allgemeinen nicht zu erwarten ist (Köhnken, 1990). Das Kriterium **„Raum-zeitliche Verknüpfungen"** bezieht sich darauf, ob die Kernhandlung der Aussage mit bestimmten örtlichen oder zeitlichen Gegebenheiten, bestimmten eigenen Gewohnheiten oder Gewohnheiten von Personen aus dem sozialen Umfeld der aussagenden Person verflochten ist. Das Kriterium **„Interaktionsschilderungen"** hebt darauf ab, ob im Bericht Ketten sich gegenseitig bedingender und sich aufeinander beziehender Handlungen (wechselseitige Aktionen und Reaktionen) wiedergegeben werden. Mit dem Kriterium **„Wiedergabe von Gesprächen"** wird überprüft, ob in der Schilderung konkrete Inhalte stattgefundener Gespräche oder einzelne Äußerungen beteiligter Personen wiedergegeben werden. Dabei muß klar erkennbar sein, von wem die einzelnen Äußerungen stammen. Das Kriterium **„Schilderung von Komplikationen im Handlungsverlauf"** ist erfüllt, wenn von unvorhergesehenen Schwierigkeiten berichtet wird, d.h. wenn es in der Schilderung durch plötzlich erscheinende Personen oder andere Umstände für den Täter nötig wird, die Handlung zu unterbrechen oder abzubrechen. Die Tat selbst gilt nicht als Komplikation.

Die sechs Kriterien der Kategorie **„Inhaltliche Besonderheiten"** beziehen sich auf solche Aussageteile, die die Konkretheit und Lebendigkeit der Bekundung in besonderem Maße erhöhen (Steller & Köhnken, 1989) bzw. die man aufgrund ihrer hohen Komplexität oder wegen ihres Abweichens von allgemeinen Schemata nicht in erfundenen Schilderungen erwarten würde (Köhnken, 1990).[1] Mit dem Kriterium **„Schilderung**

[1] Insofern würde Kriterium 7, „Schilderung von Komplikationen im Handlungsverlauf", eigentlich besser in diese Kategorie als in die Kategorie „Spezielle Inhalte" passen (vgl. Köhnken, 1990, S. 106).

ausgefallener Einzelheiten“ wird überprüft, ob in der Aussage ungewöhnliche oder seltene bzw. einzigartige Einzelheiten auftreten, die zugleich aber nicht prinzipiell unrealistisch sind. Mit **„Schilderung nebensächlicher Einzelheiten“** ist gemeint, daß Einzelheiten erzählt werden, die mit dem inkriminierten Tathergang in keinem sachlogischen Zusammenhang stehen und für das Kerngeschehen irrelevant sind, die jedoch bei der Beschreibung des Handlungsablaufs beiläufig erwähnt werden. Das Kriterium **„Phänomengemäße Schilderung unverstandener Handlungselemente“** ist dann erfüllt, wenn ein Sachverhalt zutreffend im Sinne seiner äußeren Anmutungsqualität beschrieben wird und gleichzeitig deutlich wird, daß der Aussagende ebendiesen Sachverhalt nicht in seiner Bedeutung durchschaut hat. Der Sachverhalt wird also zutreffend beschrieben, aber falsch interpretiert. Der Aussagende kann, muß jedoch nicht sein Unverständnis bezüglich des Sachverhalts zum Ausdruck bringen. Ein Beispiel für dieses Kriterium ist die Beschreibung von Orgasmuszuständen durch kindliche Opfer von Sittlichkeitsdelikten. Als **„Indirekt handlungsbezogene Schilderungen“** im Sinne eines Glaubhaftigkeitskriteriums gelten Berichte über Ereignisse, die dem inkriminierten Tatgeschehen zwar ähneln, sich jedoch zu anderer Zeit und mit anderen Personen zugetragen haben. Das Kriterium ist beispielsweise erfüllt, wenn das mutmaßliche Opfer einer inzestuösen Beziehung über eine verbale Auseinandersetzung mit dem Täter berichtet, in welcher dieser dem Opfer frühere sexuelle Erfahrungen mit anderen Partnern vorwarf. Werden von der aussagenden Person eigene, im Zusammenhang mit dem Kerngeschehen aufgetretene Gedanken oder Gefühle bzw. deren motorische (z.B. Zittern) oder physiologische (z.B. Herzklopfen) Begleiterscheinungen differenziert beschrieben, so ist das Kriterium **„Schilderung eigener psychischer Vorgänge“** erfüllt. Das Kriterium **„Schilderung psychischer Vorgänge des Angeschuldigten“** ist dann erfüllt, wenn vermutete Gedanken oder Gefühle bzw. entsprechende beobachtete motorische oder physiologische Begleiterscheinungen des Täters oder auch anderer beteiligter Personen berichtet werden.

Die vierte Kategorie von Kriterien setzt sich aus Aussageinhalten zusammen, die Rückschlüsse auf die Motivation des Aussagenden zulassen. Bei **„Motivationsbezogenen Inhalten“** handelt es sich um spezielle Details, die dem Ziel einer glaubwürdigen und kompetenten Selbstdarstellung vordergründig abträglich sind und die demzufolge von durchschnittlich begabten Personen beim Lügen im allgemeinen vermieden werden (Köhnken, 1990). Mit dem Kriterium **„Spontane Verbesserungen der eigenen Aussage“** wird überprüft, ob der Inhalt der Aussage unaufgefordert präzisiert oder berichtigt wird. Das Kriterium **„Eingeständnis von Erinnerungslücken“** ist erfüllt, wenn die aussagende Person zugibt, sich an bestimmte Einzelheiten nicht mehr bzw. nur noch unsicher erinnern zu können. Mit **„Einwänden gegen die Richtigkeit der eigenen Aussage“** ist gemeint, daß die aussagende Person die Glaubhaftigkeit ihrer Aussage

oder generell die Glaubwürdigkeit der eigenen Person in Zweifel zieht (z.B. mit der Äußerung: „Auch wenn mir das sowieso kein Mensch glaubt ...“). Zudem ist das Kriterium erfüllt, wenn die aussagende Person explizit die Möglichkeiten eigener Fehlwahrnehmungen, Verwechslungen oder Mißverständnisse erwägt. Berichtet die aussagende Person in selbstkritischer Weise Unvorteilhaftes über sich bzw. über ihr eigenes Verhalten (auch vermeintliches Fehlverhalten gegenüber dem Beschuldigten), so liegt Kriterium 17, **„Selbstbelastungen“**, vor. Kriterium 18, **„Entlastung des Angeschuldigten“**, ist realisiert, wenn sich in der Aussage Angaben finden, die den Beschuldigten entlasten oder aufwerten, bzw. wenn erkennbar ist, daß die aussagende Person trotz sich bietender Möglichkeit keine Mehrbelastung des Beschuldigten vornimmt.

In der Kategorie **„Deliktspezifische Inhalte“** werden schließlich solche Aussageelemente betrachtet, die in typischer Weise mit dem behaupteten Delikt in Verbindung stehen. Das Kriterium **„Deliktspezifische Aussageelemente“** ist bislang ausschließlich für das Gebiet der Sexualdelikte definiert (was daran liegt, daß die Begutachtung von Zeugenaussagen in Sittlichkeitsprozessen das primäre praktische Einsatzgebiet der inhaltsorientierten Glaubhaftigkeitsbeurteilung ist, vgl. Abschnitt 2.1.4). In Aussagen zu Sexualdelikten ist das Kriterium erfüllt, wenn die Beschreibung des Tathergangs sich mit empirisch-kriminologischen Erkenntnissen über charakteristische Begehungsformen von Sexualverbrechen deckt, und zwar besonders dann, wenn der geschilderte Tathergang vom Alltagswissen über Sexualverbrechen abweicht.

Um Platz zu sparen, werden im folgenden anstelle der vollständigen Bezeichnungen nur noch die in Tabelle 1 aufgeführten Kurzbezeichnungen für die einzelnen Glaubhaftigkeitskriterien verwendet.

2.1.3 Diagnostische Vorgehensweise

Das Prozedere bei inhaltsorientierten Glaubhaftigkeitsbegutachtungen in der forensischen Praxis umfaßt neben dem zentralen Element, der Analyse der inhaltlichen Aussagequalität, auch noch eine Persönlichkeitsanalyse des Aussagenden, eine Analyse seiner potentiellen Falschbezichtigungsmotive sowie eine Rekonstruktion der kommunikativen Bedingungen der Aussageentstehung (Greuel et al., 1998; Steller & Volbert, 1997, Undeutsch, 1967).

Primäres Ziel der **Persönlichkeitsanalyse** ist die Feststellung der intellektuellen (v.a. Phantasiebegabung) und sprachlichen Kompetenz des Aussagenden. Dabei kommen die üblichen Verfahren der psychologischen Diagnostik zur Anwendung (biographische

Analyse, Tests, Beobachtung, Exploration; vgl. Steller & Volbert, 1997; Greuel et al., 1998).

Die **Motivanalyse** soll klären, ob seitens der aussagenden Person Gründe für eine absichtliche Falschbekundung vorliegen. Hier sind in erster Linie das Verhältnis zwischen der aussagenden Person und dem Beschuldigten (z.B. Abneigung, Neid, oder Rachemotive gegenüber dem Beschuldigten) sowie mögliche positive bzw. negative Konsequenzen in Betracht zu ziehen, welche sich aus der Aussage für die aussagende Person bzw. den Beschuldigten oder etwaige beteiligte Drittpersonen ergeben. Liegen Motive für eine Falschaussage bzw. Falschbezichtigung vor, so darf die Aussage jedoch keineswegs quasi automatisch als erfunden eingestuft werden, da sie dennoch auf einem Erlebnishintergrund basieren kann (vgl. Steller & Volbert, 1997; Greuel et al., 1998).

Hinweise auf mögliche Aussageverfälschungen sowohl absichtlicher als auch nichtintentionaler Art ergeben sich aus der genauen **Rekonstruktion der Entstehungsgeschichte der Aussage**. Hier sind v.a. der situative Kontext und der Adressat der Erstaussage, das Verhalten und die begleitenden Emotionen der aussagenden Person zum Zeitpunkt der Erstbeschuldigung sowie potentielle suggestive Einflüsse auf den Aussagenden zu berücksichtigen (vgl. Greuel et al., 1998; Steller & Volbert, 1997).

Kern der inhaltsorientierten Glaubhaftigkeitsbegutachtung ist jedoch die **Bestimmung der inhaltlichen Qualität der in Frage stehenden Aussage** (Arntzen, 1993; Undeutsch, 1967; Steller & Köhnken, 1989). Die Inhaltsanalyse erfolgt, indem der Gutachter jedes einzelne Glaubhaftigkeitskriterium daraufhin überprüft, ob es in der Schilderung so stark ausgeprägt ist, daß es auf einen Erlebnisbezug der Aussage verweist (Greuel et al., 1998). In der empirischen Forschung, in welcher zumeist die oben vorgestellte *Kriterienorientierte Inhaltsanalyse* (Steller & Köhnken, 1989) zur Anwendung kommt, werden die Ausprägungsgrade der einzelnen Kriterien in aller Regel anhand mehrstufiger Ratingskalen quantifiziert. Aber auch hier stellt die Bestimmung des Ausprägungsgrades – ebenso wie in der forensischen Begutachtungspraxis – letztlich einen komplexen Einschätzungsvorgang dar, in den sowohl qualitative als auch quantitative Elemente einfließen. So hängt die Ausprägung eines Kriteriums zum einen von der Häufigkeit ab, mit der das Kriterium in der Aussage vorkommt (quantitativer Aspekt), zum anderen aber auch davon, wie deutlich das Kriterium an den betreffenden Textstellen erfüllt ist (qualitativer Aspekt) (Greuel et al., 1998; Steller, Wellershaus & Wolf, 1992). Die Beurteilung der Ausprägungsgrade der Glaubhaftigkeitskriterien sollte in jedem Fall auf der Grundlage eines Aussagetranskripts erfolgen. Die Verwendung von Video- oder Tonaufzeichnungen zur Inhaltsanalyse ist hingegen nicht sinnvoll, da

die Gefahr einer Kontaminierung der inhaltlichen Aussageanalyse durch Merkmale des aussagebegleitenden extralinguistischen oder nonverbalen Ausdrucksverhaltens besteht.

Da sich der in der „Undeutsch-Hypothese“ postulierte qualitative Unterschied zwischen erlebnisbezogenen und erfundenen Aussagen v.a. im intraindividuellen Vergleich manifestieren soll, empfiehlt es sich, vom Pb neben der begutachtungsrelevanten Aussage zum inkriminierten Geschehen auch noch **Berichte über zweifelsfrei selbsterlebte Ereignisse sowie reine Phantasiegeschichten** zu erheben. Diese gesichert erlebnisbezogenen und erfundenen Aussagen können einer vergleichenden Qualitätsanalyse unterzogen werden, um so das individualspezifische Kriteriengepräge beider Aussageformen zu ermitteln. Hieran kann dann die inhaltliche Qualitätsstruktur der begutachtungsrelevanten Aussage relativiert werden (Arntzen, 1993; Offe & Offe, 1994; Greuel et al., 1998).

Falls mehrere, zeitlich getrennte Aussagen zu dem inkriminierten Sachverhalt vorliegen, erfolgt außerdem auch noch eine **Konstanzprüfung**. Dabei wird analysiert, inwiefern die zu verschiedenen Zeitpunkten abgegebenen Schilderungen Übereinstimmungen, Widersprüche, Ergänzungen oder Auslassungen aufweisen, aus denen sich unter Berücksichtigung gedächtnispsychologischer Gesetzmäßigkeiten Hinweise auf die Erlebnisbezogenheit der Aussage ergeben (Arntzen, 1993; Steller & Volbert, 1997).

Die Inhaltsanalyse ermöglicht also anhand der Feststellung der Ausprägungsgrade der einzelnen Kriterien eine Beurteilung der inhaltlichen Qualität der tatbestandsbezogenen Aussage (im Vergleich zu den erlebnisbasierenden bzw. erfundenen Vergleichsschilderungen). Um von der festgestellten Inhaltsqualität Schlußfolgerungen auf die Glaubhaftigkeit der Bekundung ziehen zu können, muß die inhaltliche Aussagequalität zur intellektuellen und sprachlichen Kompetenz des Aussagenden, seinen bereichsspezifischen Erfahrungen und Kenntnissen sowie den in Frage kommenden Aussagemotiven in Beziehung gesetzt werden (Steller & Volbert, 1997). Letztlich schätzt der Gutachter auf der Grundlage der Ergebnisse aus Inhalts-, Persönlichkeits- und Motivanalyse ein, „wie wahrscheinlich es ist, daß die aussagende Person mit den festgestellten Fähigkeiten, Eigenschaften, Erfahrungen und Motiven auch ohne Erlebnishintergrund eine Aussage produzieren könnte, welche die durch die Inhaltsanalyse festgestellte qualitative Struktur aufweist“ (Steller & Volbert, 1997, S. 25). Eine Aussage wird also als glaubhaft eingestuft, wenn sie eine hohe inhaltliche Qualität im Sinne der „Undeutsch-Hypothese“ aufweist und wenn diese hohe Qualität nicht allein mit der intellektuellen und sprachlichen Begabung des Aussagenden oder seinen deliktspezifischen Vorkenntnissen zu erklären ist. Kann dagegen eine hohe inhaltliche Qualität neben einer vermeintlichen Erlebnisgrundlage auch auf andere Faktoren zurückgeführt werden oder ergeben sich im

Rahmen der Konstanzprüfung erhebliche Inkonsistenzen zwischen den zeitlich getrennten Aussagen, die nicht mit normalen Gedächtnisunsicherheiten zu erklären sind, so wird die Glaubhaftigkeit der Aussage nicht gestützt.

Es sei noch erwähnt, daß sich in der internationalen Fachliteratur für die inhaltsorientierte Glaubhaftigkeitsbeurteilung die Bezeichnung **„statement validity assessment“** bzw. **„statement validity analysis“** (SVA; z.B. Steller, 1989; Vrij & Akehurst, 1998) etabliert hat. SVA wird als drei Elemente umfassend beschrieben. Diese Elemente sind erstens ein strukturiertes Interview zur Gewinnung der Aussage, zweitens die *Kriterienorientierte Inhaltsanalyse* als Kernstück der diagnostischen Prozedur und drittens die sog. „validity checklist“, ein Fragenkatalog zur Relativierung der inhaltlichen Aussagequalität an etwaigen Besonderheiten des Verhaltens der aussagenden Person, des Explorationsgesprächs, der Aussagemotivation sowie der allgemeinen Beweislage (vgl. z.B. Steller & Boychuk, 1992; Raskin & Esplin, 1991b; Raskin & Yuille, 1989; Yuille, 1988). Ein konkretes Beispiel für die Begutachtungsprozedur wird bei Steller und Boychuk (1992) beschrieben.

2.1.4 Anwendungsbereich

Der Hauptanwendungsbereich der inhaltsorientierten Glaubhaftigkeitsbeurteilung ist die **Begutachtung mutmaßlicher Zeugen bzw. Opfer von Straftaten gegen das sexuelle Selbstbestimmungsrecht** (sexueller Mißbrauch, Vergewaltigung, sexuelle Nötigung), wobei es sich meist um Kinder und Jugendliche weiblichen Geschlechts handelt (vgl. z.B. Arntzen, 1993; Undeutsch, 1967; Wolf & Steller, 1997). Die weitgehende Einschränkung auf den Bereich der Sittlichkeitsprozesse liegt jedoch nicht etwa in der Logik der Methode begründet, sondern ist auf die rechtlichen Rahmenbedingungen in Deutschland und die daraus resultierende selektive Beauftragungspraxis von Richtern und Staatsanwälten zurückzuführen (Steller & Köhnken, 1989, S. 233). Grundsätzlich sind Sachverständige immer dann hinzuzuziehen, wenn die Sachkunde des Gerichts nicht ausreicht, um einen rechtlich relevanten Sachverhalt festzustellen oder zu beurteilen (Greuel et al., 1998). Im Hinblick auf die Beurteilung der Glaubhaftigkeit von Aussagen stellte der Bundesgerichtshof (BGH) in einer Grundsatzentscheidung vom 14.12.1954 (Aktenzeichen: 5 StR 416/54) fest, daß die Glaubhaftigkeit von Kinderaussagen schwerer zu beurteilen sei als die von Aussagen erwachsener Zeugen und daß medizinischen und psychologischen Sachverständigen diesbezüglich bessere Erkenntnismittel zur Verfügung stünden als dem Gericht in der Hauptverhandlung. Insbesondere in Sexualfällen, in denen die Aussagen Minderjähriger die primären oder ausschließlichen Beweismittel darstellen oder sich nicht mit anderen Beweismitteln dek-

ken, sei daher die Hinzuziehung psychiatrischer oder psychologischer Sachverständiger geboten (BGH, 1955, S. 82ff.; vgl. auch Undeutsch, 1956, 1989). Dieses höchstrichterliche Urteil stellte die Weichen für das bis heute dominierende praktische Einsatzgebiet der inhaltsorientierten Glaubhaftigkeitsbeurteilung. Auch in der gegenwärtigen Rechtspraxis gilt die Hinzuziehung aussagepsychologischer Sachverständiger dann als indiziert, wenn „Zeugen die einzigen Belastungszeugen sind, ohne daß zusätzlich objektive Sachbeweise vorliegen oder [...] Zeugen gleichzeitig als Geschädigte in Frage kommen („Opferzeugen"), so daß das Wirksamwerden potentieller Belastungsmotive nicht von vorneherein ausgeschlossen werden kann" (Greuel et al., 1998, S. 284). Beide Bedingungen sind besonders häufig in Fällen mutmaßlicher Sexualdelikte erfüllt.

Aus psychologischer Perspektive **steht der Anwendung** der inhaltsorientierten Glaubhaftigkeitsbeurteilung auch **bei anderen Altersgruppen und Delikttypen jedoch prinzipiell nichts entgegen** (z.B. Greuel et al., 1998; Fabian, Greuel & Stadler, 1996; Fabian & Wetzels, 1991; Wetzels, 1990). So formuliert etwa Arntzen (1987, 1993) seine Glaubhaftigkeitskriterien explizit im Hinblick auf sämtliche Altersgruppen und Justizsparten. Auch die Analyse der Einlassungen von Beschuldigten ist grundsätzlich möglich (Steller & Köhnken, 1989) und wird bei Yuille und Cutshall (1989) bzw. Porter und Yuille (1995) ausführlich diskutiert. Trankell (1971) beschreibt diesbezüglich Beispiele aus der Praxis. Nicht zuletzt auch von rechtswissenschaftlicher Seite wird mitunter eine Ausdehnung des Anwendungsbereichs der inhaltsorientierten Glaubhaftigkeitsbeurteilung auf rechtliche Problemstellungen unterschiedlichster Provenienz gefordert und im Rahmen der juristischen Glaubwürdigkeitslehre bereits realisiert (z.B. H.-U. Bender, 1987; R. Bender & Nack, 1995; Prüfer, 1986; Rüßmann, 1997).

2.1.5 Problematik

Der wohl problematischste Aspekt der inhaltsorientierten Glaubhaftigkeitsbeurteilung betrifft den **Prozeß der diagnostischen Inferenz**, d.h. die Art und Weise, wie die Daten aus Inhalts-, Persönlichkeits- und Motivanalyse zu einem Wahrscheinlichkeitsurteil hinsichtlich des Erlebnisbezugs der Aussage zu integrieren sind. Nach der vorherrschenden Meinung in der forensischen Aussagepsychologie darf diese Datenintegration nicht nach formalisierten oder gar quantifizierbaren Entscheidungsregeln erfolgen, sondern hat sich an den individuellen Besonderheiten des konkreten Einzelfalls auszurichten (Greuel et al., 1998) und ist letztlich **klinisch-intuitiver Natur** (Steller, 1989; Steller & Köhnken, 1989). Gleichwohl finden sich in der Literatur einige Ansätze für allgemeingültige quantitative Entscheidungsrichtlinien. So stellt Arntzen (1993) auf der Grundlage seiner praktischen Erfahrung als Gutachter die Faustregel auf, daß im Durch-

schnitt mindestens drei deutlich ausgeprägte Glaubhaftigkeitskriterien im Verbund vorliegen müssen, um eine Aussage als glaubhaft zu qualifizieren. Einem solchen „Merkmalskomplex“ billigt er „vollen Beweiswert“ (S. 22f.) zu. Nach Trankell (1971) ist sogar schon das Vorliegen eines einzigen deutlich ausgeprägten Kriteriums ausreichend. Wenngleich Arntzen und Trankell quantitative Angaben bezüglich der erforderlichen Anzahl erfüllter Glaubhaftigkeitskriterien machen, so beruht doch gerade in den Kriteriologien dieser beiden Autoren die Entscheidung, ab wann ein Kriterium deutlich genug ausgeprägt ist, auf einem äußerst komplexen und kaum in Entscheidungsalgorithmen zu fassenden Urteilsprozeß. Angesichts dieser Operationalisierungsprobleme verwundert es kaum, daß bislang noch keine kontrollierten Untersuchungen zur Überprüfung der genannten Entscheidungsregeln publiziert wurden. Zudem ist zu betonen, daß die Kriteriologien von Arntzen und Trankell neben rein inhaltlichen auch noch andere Glaubhaftigkeitskriterien enthalten, so z.B. die aussagebegleitende Gefühlsbeteiligung. Insofern sind die genannten Entscheidungsregeln nicht unmittelbar auf die *Kriterienorientierte Inhaltsanalyse* übertragbar, die ja nur inhaltliche Kriterien aufweist (wenngleich auch deren Resultat im Rahmen der Gesamtbeurteilung letztlich an nichtinhaltlichen Aussageaspekten relativiert wird; s. Abschnitt 2.1.3).

Auch im Rahmen der empirischen Forschung zur *Kriterienorientierten Inhaltsanalyse* wurden mitunter quantitative Richtlinien für die diagnostische Urteilsbildung vorgeschlagen bzw. erprobt. Yuille (1990, zitiert nach Zaparniuk, Yuille & Taylor, 1995, S. 345) schlägt beispielsweise vor, Aussagen dann als glaubhaft einzustufen, wenn die Kriterien 1 bis 5 sowie zwei beliebige weitere Kriterien der *Kriterienorientierten Inhaltsanalyse* erfüllt seien. Landry und Brigham (1992) führten im Rahmen eines Auswertertrainings eine Entscheidungsregel an, derzufolge bei Vorliegen von mehr als fünf Glaubhaftigkeitskriterien mit hoher Wahrscheinlichkeit auf eine Erlebnisgrundlage geschlossen werden könne. Die gleiche Regel wird auch von Craig (1995, zitiert nach Vrij & Akehurst, 1998, S. 18) postuliert. Auf die empirische Bewährung dieser Entscheidungsregeln wird in Abschnitt 2.1.6.2 eingegangen.

Im Zusammenhang mit der Problematik der Datenintegration zur diagnostischen Urteilsbildung wird auch die Frage nach einer **adäquaten Gewichtung der einzelnen Kriterien** kontrovers diskutiert. So messen beispielsweise Raskin und Esplin (1991b) den Kriterien 4 und 5 (*Verknüpfungen* und *Interaktionen*) besonderes Gewicht bei, während diese Kriterien nach Steller und Köhnken (1989) eher von untergeordneter Bedeutung sind. Theoretische Begründungen für die abweichenden Auffassungen der Experten werden jedoch nicht gegeben. Letztlich ist die Frage nach der differentiellen Gewichtung der Kriterien nur zu klären, indem man in Feld- und Laborstudien überprüft, wie gut die einzelnen Kriterien zwischen erlebnisbezogenen und erfundenen Aussagen

zu differenzieren vermögen. Hierbei geht es also um die empirische Überprüfung der Validität der einzelnen Glaubhaftigkeitskriterien. Die diesbezügliche Befundlage ist Gegenstand von Abschnitt 2.1.6.1. Es kann aber bereits vorweggenommen werden, daß die vorliegende Empirie keine Schlußfolgerungen hinsichtlich einer verbindlichen Gewichtungsstruktur der Kriterien im konkreten forensischen Einzelfall zuläßt.

Bisweilen wurde auch kritisiert, daß die **operationalen Definitionen** einiger Glaubhaftigkeitskriterien zu unscharf seien, um eine ausreichende **Reliabilität** bei der Bestimmung ihres Ausprägungsgrades gewährleisten zu können (Horowitz, Lamb, Esplin, Boychuk, Krispin & Reiter-Lavery, 1997). Diesem Kritikpunkt wurde jedoch durch die Entwicklung entsprechender Auswerter-Trainingsprogramme Rechnung getragen (z.B. Krause, 1997; Petersen, 1997), so daß mittlerweile unter der Voraussetzung ausreichender Schulung der Auswerter von einer befriedigenden Reliabilität der einzelnen Kriterien ausgegangen werden kann (z.B. Höfer, Krause, Petersen, Sievers & Köhnken, 1999).

Die oben beschriebenen Kritikpunkte haben Implikationen für die empirische Validierung der inhaltsorientierten Glaubhaftigkeitsbeurteilung, deren Resultate im nächsten Abschnitt (2.1.6) genauer dargestellt werden. Grundsätzlich beinhaltet die Überprüfung der Validität zwei Stufen (Steller, 1989; Steller et al., 1992). Zum einen ist zu untersuchen, ob der in der „Undeutsch-Hypothese“ postulierte qualitative Unterschied zwischen glaubhaften und unglaubhaften Aussagen tatsächlich zutrifft, d.h. ob sich die inhaltlichen Glaubhaftigkeitskriterien in Berichten über selbsterlebte Ereignisse zahlreicher bzw. in stärkerer Ausprägung finden als in erfundenen Aussagen. Hier geht es also um die Validierung der 19 Kriterien der *Kriterienorientierten Inhaltsanalyse*, die ja das Kernstück der Beurteilung der Glaubhaftigkeit bildet. Läßt sich nachweisen, daß die Kriterien in glaubhaften Aussagen häufiger bzw. stärker ausgeprägt vorkommen als in unglaubhaften Aussagen, so kann man zur zweiten Stufe der Validierung übergehen, in welcher es gilt, die Treffsicherheit der auf Basis der Gesamtbegutachtung vorgenommenen Klassifikationen von Aussagen als glaubhaft vs. unglaubhaft zu bestimmen. Hier geht es also letztlich darum, den Prozeß der diagnostischen Datenintegration bzw. Urteilsbildung zu validieren.

Die erste Stufe, also die Validierung der *Kriterienorientierten Inhaltsanalyse*, ist angesichts der inzwischen vorliegenden präzisen operationalen Definitionen der einzelnen Kriterien leicht realisierbar. Die zweite Stufe der Validierung stößt jedoch – selbst wenn die erste Stufe positive Resultate erbringt – auf erhebliche **forschungsmethodische Schwierigkeiten**. Wie oben dargelegt wurde, handelt es sich bei der diagnostischen Urteilsbildung um einen unstandardisierten, einzelfallabhängigen, klinisch-intuitiven

Prozeß. Wenn sich jedoch nicht genau und verbindlich spezifizieren läßt, wie aus der festgestellten inhaltlichen Qualität einer Aussage unter Berücksichtigung der Ergebnisse der Persönlichkeits- und Motivanalyse ein Urteil bezüglich der Glaubhaftigkeit abgeleitet werden soll, ist es auch unmöglich, im strengen Sinne kontrollierte Untersuchungen zur Treffsicherheit der diagnostischen Methode anzulegen. Vielmehr ist man darauf angewiesen, anhand einer Stichprobe von realen oder experimentellen Begutachtungsfällen Trefferquoten zu ermitteln und sich dabei darauf zu verlassen, daß es sich bei den beteiligten Diagnostikern um „Experten" handelt, deren klinisch-intuitive Urteilsbildung „in angemessener Art und Weise" erfolgte. Was jedoch einen „Experten" bzw. die „angemessene Art und Weise" genau ausmacht, bleibt – unabhängig von der Höhe der erzielten Trefferquoten – unklar. Pragmatisch betrachtet, mag es zwar befriedigen, wenn sich auf solche Weise hohe Trefferquoten ermitteln lassen sollten; aus wissenschaftlicher Perspektive ist der bloße Nachweis einer hohen Treffsicherheit der diagnostischen Methode jedoch grundsätzlich unbefriedigend, solange nicht geklärt ist, wodurch genau sie zustandekommt. Streng genommen ergibt sich aus der **Abhängigkeit des diagnostischen Urteils von der Kompetenz des Diagnostikers** sogar die Konsequenz, daß die Treffsicherheit des Verfahrens allenfalls untersucherspezifisch bestimmt werden kann bzw. daß sich **keine generalisierbaren Validitätsanagaben** machen lassen (vgl. Ben-Shakhar & Furedy, 1990).

Im nächsten Abschnitt wird deutlich werden, daß die Befundlage zur ersten Stufe der Validierung der inhaltsorientierten Glaubhaftigkeitsbeurteilung eher heterogen ist. Somit erscheint es aus wissenschaftlicher Sicht – streng genommen – verfrüht, schon zur zweiten Validierungsstufe überzugehen. Andererseits ist es angesichts der weitverbreiteten Anwendung der Methode in der forensischen Praxis geradezu geboten, ihre Treffsicherheit zu überprüfen. Wenn die diesbezügliche empirische Befundlage erstaunlicherweise auch sehr dürftig ist (vgl. Abschnitt 2.1.6.2), ergeben sich in ihr doch Hinweise auf einen weiteren problematischen Aspekt der Methode, der an dieser Stelle vorweggenommen werden soll, zumal er von erheblicher Praxisrelevanz ist. So deuten die Forschungsergebnisse auf eine **Fehlertendenz in Richtung falsch positiver Urteile** hin, d.h. Auswerter, die Aussagen einer *Kriterienorientierten Inhaltsanalyse* unterziehen und anschließend Urteile bezüglich der Glaubhaftigkeit abgeben sollen, neigen eher dazu, erfundene Aussagen fälschlich als glaubhaft zu klassifizieren, als daß sie erlebnisbezogene Bekundungen für unglaubhaft halten (falsch negative Urteile) (vgl. Ruby & Brigham, 1997; Vrij & Akehurst, 1998). Diese Fehlertendenz tritt auch dann auf, wenn die Urteile auf quantitativen Entscheidungsrichtlinien basieren (z.B. Ruby & Brigham, 1998; vgl. Abschnitt 2.1.6.2); und selbst ausgewiesene Experten in der inhaltsorientierten Glaubhaftigkeitsbeurteilung scheinen ihr zu unterliegen (Vrij, Kneller & Mann, 2000, vgl. Abschnitt 2.1.6.2). Vor dem Hintergrund, daß es sich bei den begutachtungs-

relevanten Aussagen in der forensischen Praxis überwiegend um Anschuldigungen (sexueller Delikte) handelt, ist diese Urteilsverzerrung nicht mit dem Rechtsgrundsatz des Schutzes Unschuldiger vereinbar.

Neben den genannten Kritikpunkten sollen noch einige weitere an dieser Stelle aufgezählt werden (vgl. Köhnken, 1986). Wie in Abschnitt 2.1.1 deutlich wurde, ist die **theoretische Verankerung** der inhaltsorientierten Glaubhaftigkeitsbeurteilung relativ dürftig. Weiterhin ist die weitgehende **Beschränkung des empirischen Analysematerials auf Aussagen zu Sexualdelikten** an (überwiegend weiblichen) Kindern und Jugendlichen zu bemängeln. Insbesondere im Hinblick auf eine breitere Anwendung der Methode wäre es wünschenswert, auch solche Aussagen systematisch zu untersuchen, die von anderen Personengruppen bzw. in anderen Justizsparten vorgebracht werden. In diesem Zusammenhang ist auch zu kritisieren, daß andere Fälschungsvarianten als die der frei erfundenen Belastungsaussage in der Forschung bisher weitestgehend vernachlässigt wurden, wohingegen in der Rechtspraxis z.B. auch frei phantasierte Entlastungsaussagen von erheblicher Relevanz sind. Ferner erfährt die inhaltsorientierte Glaubhaftigkeitsbeurteilung eine wesentliche Einschränkung dadurch, daß der **Diagnostiker in hohem Maße von dem vorgefundenen Datenmaterial abhängig** ist. So läßt sich die Methode nur in solchen Fällen sinnvoll anwenden, in denen die Aussage einen gewissen Mindestumfang besitzt und sich auf einen einigermaßen komplexen Geschehensablauf bezieht.

Abschließend sei noch auf die Gefahr der **Manipulierbarkeit** der inhaltsorientierten Glaubhaftigkeitsbeurteilung hingewiesen. Diese wurde in einer in Abschnitt 2.1.6.1 näher beschriebenen experimentellen Studie von Vrij et al. (2000) untersucht. Es zeigte sich, daß Personen, die man über das Prinzip bzw. einzelne Kriterien der *Kriterienorientierten Inhaltsanalyse* informiert hatte, in ihren erfundenen Aussagen mehr Glaubhaftigkeitskriterien produzierten als uninformierte Personen. Die inhaltliche Aussagequalität der informierten Lügner reichte sogar an die von erlebnisbezogenen Schilderungen heran.

2.1.6 Empirische Validitätsbefunde

Im folgenden werden zunächst Forschungsresultate dargestellt, die die Validität der „Undeutsch-Hypothese“ bzw. der Glaubhaftigkeitskriterien betreffen. Es geht hier also um die Frage, ob die Kriterien in erlebnisbezogenen Aussagen häufiger bzw. in stärkerer Ausprägung anzutreffen sind als in erfundenen Aussagen. Anschließend werden empirische Befunde aufgeführt, die die Treffsicherheit der auf Basis der Inhaltsanalyse

erfolgenden diagnostischen Urteile betreffen. Es werden nur die Ergebnisse von Studien berichtet, in denen die Glaubhaftigkeitskriterien nach der Kategorisierung von Steller und Köhnken (1989; *Kriterienorientierte Inhaltsanalyse*) verwendet wurden. Dies geschieht einerseits, weil diese Kriteriologie sich mittlerweile als Standardbezugssytem innerhalb der empirischen Forschung etabliert hat. Zum anderen sind nur solche Forschungsresultate direkt miteinander vergleichbar, die unter Verwendung weitgehend gleicher operationaler Definitionen der Glaubhaftigkeitskriterien zustande kamen. Auf häufig zitierte frühere Studien, in denen noch nicht die *Kriterienorientierte Inhaltsanalyse* zugrunde gelegt wurde (z.B. Köhnken & Wegener, 1982; Littmann & Szewczyk, 1983; Rüth-Bemelmanns, 1984), wird hier nicht eingegangen.

2.1.6.1 Validität der inhaltlichen Glaubhaftigkeitskriterien

Zur Validität der Glaubhaftigkeitskriterien nach der Kategorisierung von Steller und Köhnken (1989) liegen sowohl Felduntersuchungen als auch experimentelle Studien vor. In Felduntersuchungen überprüft man, inwiefern die Glaubhaftigkeitskriterien in Aussagen auftreten, welche in der forensischen Praxis vorzufinden sind (z.B. Aussagen im Rahmen von polizeilichen Vernehmungen oder gerichtsgutachterlichen Explorationsgesprächen). Dabei wird der „tatsächliche“ Status der Bekundungen (erlebnisbezogen vs. erfunden) an Kriterien festgemacht, die vom Ergebnis der inhaltsorientierten Begutachtung möglichst unabhängig sein sollen (z.B. Aussagewiderruf des mutmaßlichen Zeugen oder Geständnis des Beschuldigten). In experimentellen Studien werden die zu analysierenden Aussagen unter – insbesondere im Hinblick auf den tatsächlichen Wahrheitsstatus – kontrollierten Bedingungen gewonnen. Während an Feldstudien insbesondere zu kritisieren ist, daß sich der tatsächliche Status der Aussagen kaum mit letzter Sicherheit feststellen läßt und daß eine Selektivität des Untersuchungsmaterials gegeben ist, liegt die Hauptproblematik von experimentellen Untersuchungen in ihrer vergleichsweise geringen externen Validität. Eine ausführlichere Diskussion der Vor- und Nachteile beider Forschungsansätze erfolgt in Abschnitt 3.2.2.

In Tabelle 2 sind die Ergebnisse einiger **Feldstudien** zur Validität der *Kriterienorientierten Inhaltsanalyse* (Steller & Köhnken, 1989) zusammengefaßt. Es sei angemerkt, daß es sich bei der Untersuchung von Boychuk (1991, zitiert z.B. nach Greuel et al., 1998, S. 137) um eine unveröffentlichte Dissertation handelt und daß die Untersuchung von Esplin, Houed und Raskin (1988, zitiert nach Raskin & Esplin, 1991a, S. 160ff.) lediglich in Form eines Kongreßbeitrags publiziert wurde. Dementsprechend sind von beiden Arbeiten keine Originalmanuskripte erhältlich. Die Ergebnisse dieser beiden Studien finden hier nur deshalb Berücksichtigung, weil sie in der einschlägigen Fachli-

teratur regelmäßig zitiert werden (z.B. Ruby & Brigham, 1997; Steller, Volbert & Wellershaus, 1993; Vrij & Akehurst, 1998). Die diesbezüglich in Tabelle 2 aufgeführten Angaben basieren allerdings auf Sekundärliteratur, so daß hierfür keine Gewähr gegeben werden kann.

Tabelle 2. Validitätsbefunde aus Feldstudien zur *Kriterienorientierten Inhaltsanalyse*

	Boychuk (1991) [a]	Craig et al. (1999)	Esplin et al. (1988) [b]	Krahé & Kundrotas (1992)	Lamb et al. (1997)	Hypothesenkonforme Befunde (%)	Hypothesenkonträre Befunde (%)
	K	K	K	E	K		
1. *Konsistenz*	>	?	>	>	–	75	0
2. *Unordnung*	>	?	>	–	>	75	0
3. *Details*	>	?	>	–	>	75	0
4. *Verknüpfungen*	>	?	>	–	>	75	0
5. *Interaktionen*	>	?	>	–	>	75	0
6. *Gespräche*	>	?	>	–	>	75	0
7. *Komplikationen*	>	?	>	–	–	50	0
8. *Ausgefallenes*	>	?	>	<	–	50	25
9. *Nebensächliches*	–	?	>	–	–	25	0
10. *Unverstandenes*	–	?	–	–	–	0	0
11. *Indirektes*	>	?	>	–	–	50	0
12. *Eigenseelisches*	>	?	>	>	–	75	0
13. *Fremdseelisches*	–	?	>	–	–	25	0
14. *Verbesserungen*	>	?*	>	–	–	50	0
15. *Erinnerungslücken*	–	?*	>	–		33	0
16. *Selbsteinwände*	–		–	–		0	0
17. *Eigenbelastung*	–		–	–		0	0
18. *Fremdentlastung*	–		>	–		33	0
19. *Deliktspezifisches*	>		>	–		67	0
Gesamtscore		>	>		>	100	0

Anmerkung: [a] Validitätsangaben nach Greuel et al. (1998), ohne Gewähr; [b] Validitätsangaben nach Vrij & Akehurst (1998), ohne Gewähr; „K“ = Aussagen von Kindern bzw. Jugendlichen; „E“ = Aussagen von Erwachsenen; „>“ = in glaubhaften Aussagen signifikant häufiger vorhanden bzw. stärker ausgeprägt als in unglaubhaften Aussagen; „<“ = in glaubhaften Aussagen signifikant seltener vorhanden bzw. schwächer ausgeprägt; „–“ = kein signifikanter Unterschied zwischen glaubhaften und unglaubhaften Aussagen; Leerstelle = Kriterium bzw. Gesamtscore wurde nicht untersucht; „?“ = Kriterium wurde untersucht, aber kein Ergebnis mitgeteilt; * Kriterien wurden zu einem Kriterium zusammengefaßt.

In sämtlichen Feldstudien handelte es sich beim Analysematerial um Anschuldigungen sexueller Delikte. Bei Krahé und Kundrotas (1992) waren dies von erwachsenen Frauen vorgebrachte Vergewaltigungsanzeigen; in den übrigen vier Studien dagegen stammten

die Aussagen von mutmaßlich sexuell mißbrauchten Kindern bzw. Heranwachsenden (vgl. Tabelle 2).

Um die **methodische Vorgehensweise in Felduntersuchungen** zu illustrieren, sei exemplarisch die Studie von Lamb, Sternberg, Esplin, Hershkowitz, Orbach & Hovav (1997) beschrieben. Diese ist insofern positiv hervorzuheben, als sich die Autoren in besonders sorgfältiger Weise um die Bestimmung des „objektiven" Wahrheitsstatus der Aussagen aus Realfällen bemühten. Aus einem Pool von ursprünglich 1187 Aussagen mutmaßlicher Opfer sexuellen Mißbrauchs wurden zunächst alle diejenigen Fälle eliminiert, bei denen der Beschuldigte unbekannt war oder bei denen kaum andere Beweise als die Mißbrauchsbehauptung des vermeintlichen Opfers vorlagen. Es wurden schließlich diejenigen 98 Fälle ausgewählt, in denen die Strafverfolgungsorgane in ausreichendem Maß aussageunabhängige Beweise hinsichtlich des Wahrheitsgehalts der Mißbrauchsanschuldigungen sammeln konnten. Diese 98 Fälle (Alter der Kinder: 4 bis 12 Jahre) wurden von Forschern, die keine nähere Kenntnis der Kinderaussagen bzw. der jeweiligen inhaltsanalytisch festgestellten Aussagequalität hatten, auf die Eindeutigkeit der übrigen Beweislage hin untersucht. Hierfür standen medizinische Befunde, Resultate von psychophysiologischen Glaubhaftigkeitsbegutachtungen der Beschuldigten (vgl. Abschnitt 2.2) sowie unter Eid geleistete Bekundungen anderer Zeugen, der Verdächtigen oder von ermittelnden Polizeibeamten zur Verfügung. Anhand eines speziell für die Studie entwickelten Meßinstruments („Independent Case Fact Scales") beurteilte man, inwiefern die Anschuldigungen durch die davon unabhängigen fallbezogenen Fakten gestützt wurden. Die Wahrscheinlichkeit des Zutreffens der Anschuldigungen wurde auf einer fünfstufigen Skala (von „sehr wahrscheinlich" bis „sehr unwahrscheinlich") eingeschätzt. Es wurden 76 Fälle als „wahrscheinlich" bzw. „sehr wahrscheinlich", neun Fälle als fraglich (mittlere Skalenstufe) und 13 Fälle als „unwahrscheinlich" bzw. „sehr unwahrscheinlich" eingestuft. Die von mehreren unabhängigen Ratern vorgenommene inhaltsanalytische Auswertung anhand der ersten 14 Kriterien der *Kriterienorientierten Inhaltsanalyse* ergab, daß fünf Kriterien (2. *Unordnung*; 3. *Details*; 4. *Verknüpfungen*; 5. *Interaktionen*; 6. *Gespräche*; vgl. Tabelle 2) in den als „wahrscheinlich" bzw. „sehr wahrscheinlich" eingestuften Aussagen signifikant häufiger auftraten als in den für „(sehr) unwahrscheinlich" befundenen Schilderungen. Die über die 14 Glaubhaftigkeitskriterien gebildeten Gesamtscores waren bei den „(sehr) wahrscheinlichen" Aussagen signifikant höher als bei den „(sehr) unwahrscheinlichen".

Faßt man die Resultate der fünf Feldstudien zusammen, so kann man resümieren, daß die Befundlage insgesamt für die Validität der „Undeutsch-Hypothese" bzw. der *Kriterienorientierten Inhaltsanalyse* spricht. Dies kommt am deutlichsten darin zum Ausdruck, daß in allen drei Studien, in denen über die einzelnen untersuchten Glaub-

haftigkeitskriterien hinweg ein Gesamtscore gebildet wurde, sich diesbezüglich signifikante hypothesenkonforme Unterschiede manifestierten, mit höheren Gesamtscores bei den post hoc als erlebnisbezogen eingestuften Aussagen (vgl. Tabelle 2). Die Betrachtung einzelner Kriterien ergibt, daß inbesondere die Kriterien 1 (*Konsistenz*), 2 (*Unordnung*), 3 (*Details*), 4 (*Verknüpfungen*), 5 (*Interaktionen*), 6 (*Gespräche*) und 12 (*Eigenseelisches*) durch die vorliegenden Befunde untermauert werden. Für diese Kriterien ergaben sich jeweils in drei von vier durchgeführten Validierungsstudien (75%; vgl. Tabelle 2) hypothesenkonforme Befunde, ohne daß zugleich ein erwartungswidriges Resultat vorliegt. Kriterium 19 (*Deliktspezifisches*) konnte in zwei von drei Feldstudien bestätigt werden; für die Kriterien 7 (*Komplikationen*), 11 (*Indirektes*) und 14 (*Verbesserungen*) verliefen immerhin noch jeweils die Hälfte der Validierungsversuche (zwei von vier) erfolgreich, ohne daß sich zugleich ein erwartungskonträres Ergebnis zeigte. Dagegen steht bei Kriterium 8 (*Ausgefallenes*) zwei validitätsstützenden auch ein hypothesenkonträrer Befund gegenüber (s. Tabelle 2), und bei den Kriterien 9 (*Nebensächliches*), 10 (*Unverstandenes*), 13 (*Fremdseelisches*), 15 (*Erinnerungslücken*), 16 (*Selbsteinwände*), 17 (*Eigenbelastung*) und 18 (*Fremdentlastung*) verliefen jeweils weniger als die Hälfte der Validierungsversuche erfolgreich, so daß die Validität dieser Kriterien insgesamt als problematisch anzusehen ist.

Diskussionswürdig ist die Frage, warum die einzelnen Feldstudien sich so deutlich in ihren Resultaten unterscheiden. Auf der einen Seite stehen die Untersuchungen von Boychuk (1991, zitiert nach Greuel et al., 1998, S. 137) und Esplin et al. (1988, zitiert nach Raskin & Esplin, 1991a, S. 160ff.), in denen jeweils eine große Anzahl von Einzelkriterien bestätigt werden konnte. Auf der anderen Seite stehen die vergleichsweise ernüchternden Befunde von Lamb et al. (1997) und insbesondere von Krahé und Kundrotas (1992). Die Studie von Craig, Scheibe, Raskin, Kircher und Dodd (1999) gibt keine Aufschlüsse über die Validität einzelner Kriterien, da in der Publikation keine entsprechenden Angaben gemacht werden (vgl. Tabelle 2).

Die negativen Resultate von Krahé und Kundrotas (1992) sind insbesondere aus zwei Gründen anfechtbar. Zum einen waren die Auswerter nahezu gar nicht in der Anwendung der *Kriterienorientierten Inhaltsanalyse* geschult – ihnen standen lediglich sehr knappe alltagssprachliche Erläuterungen der Glaubhaftigkeitskriterien zur Verfügung. Die mangelhafte Schulung der Rater manifestierte sich auch in entsprechend niedrigen Werten für die Interrater-Reliabilität. Hinzu kommt noch, daß das Analysematerial nicht aus wörtlichen Aussagetranskripten, sondern lediglich aus zusammenfassenden und gewichtenden polizeilichen Vernehmungsprotokollen bestand, so daß offen bleibt, in welchem Ausmaß kriterienerfüllende Aussageteile erst gar nicht in die Auswertung gelangten.

Die positiven Befunde von Esplin et al. (1988, zitiert nach Raskin & Esplin, 1991a, S. 160ff.) sind nach Argumentation von Wells und Loftus (1991; vgl. auch Vrij & Akehurst, 1998) u.a. in zweierlei Hinsicht kritisierbar. Zum einen waren die Kinder, deren Aussagen als erlebnisbezogen eingestuft wurden, im Durchschnitt älter als die Kinder, deren Aussagen für unwahr gehalten wurden (9.1 vs. 6.9 Jahre). Somit kommt der mit dem Alter variierende kognitive Entwicklungsstand als konfundierende Variable in Betracht. Zweitens hängen die gefundenen Gruppenunterschiede möglicherweise mit der Selektivität des Aussagematerials zusammen. So ist es nach Auffassung von Wells und Loftus (1991) denkbar, daß wahrheitsgemäße Anschuldigungen, die in einer wenig überzeugenden Art und Weise vorgetragen werden, von den Strafverfolgungsorganen nicht weiterverfolgt werden und somit nicht als glaubhafte Aussagen verifiziert werden können. Als Ursachen für eine geringe Überzeugungskraft von erlebnisbezogenen Anschuldigungen kommen z.B. geringe verbale Fähigkeiten, Defizite im logischen Schlußfolgern oder die Befürchtung, nicht alle peripheren Tatdetails reproduzieren zu können, in Betracht. Ebensolche Faktoren manifestieren sich jedoch auch in der inhaltlichen Aussagequalität, die ja als Indikator des Erlebnisbezugs herangezogen wird. Somit könnte die gefundene höhere inhaltliche Qualität (größere Häufigkeit bzw. Ausprägung der Glaubhaftigkeitskriterien) der als wahr eingestuften Aussagen schlichtweg darauf beruhen, daß erlebnisbezogene Aussagen mit geringer inhaltlicher Qualität von der Analyse ausgeschlossen wurden. Der letztgenannte Kritikpunkt einer möglichen Selektivität des Untersuchungsmaterials gilt jedoch nicht nur für die Studie von Esplin und Kollegen, sondern läßt sich auch auf die anderen Feldstudien generalisieren, also auch auf die Felduntersuchung von Boychuk (1991, zitiert nach Greuel et al., 1998, S. 137), in welcher ähnlich positive Validitätsbefunde erzielt wurden wie bei Esplin et al. (1988, zitiert nach Raskin & Esplin, 1991a, S. 160ff.).

Die höchste Aussagekraft kommt wohl noch der Studie von Lamb et al. (1997) zu. Nach Auffassung von Vrij und Akehurst (1998) erfüllt diese Untersuchung als einzige die von Lykken (1988) postulierten Hauptgütekriterien wissenschaftlicher Feldforschung, nämlich Repräsentativität der Stichprobe, Datenerhebung im Rahmen forensischer Ernstfälle, unabhängige Auswertung durch hinsichtlich der Bedingungszugehörigkeit der Fälle uninformierte Auswerter sowie Validierung an einem vom zu validierenden Merkmal unabhängigen Außenkriterium. Die bei Lamb et al. (1997) bestätigten Kriterien (2. *Unordnung*; 3. *Details*; 4. *Verknüpfungen*; 5. *Interaktionen*; 6. *Gespräche*) konnten auch in den Untersuchungen von Boychuk (1991, zitiert nach Greuel et al., 1998, S. 137) und Esplin et al. (1988, zitiert nach Raskin & Esplin, 1991a, S. 160ff.) in ihrer Validität untermauert werden. Zumindest für diese fünf Kriterien kann die Feldbefundlage insgesamt als validitätsstützend angesehen werden.

Tabelle 3 bietet einen Überblick über **experimentelle Befunde** zur Validität der *Kriterienorientierten Inhaltsanalyse*. Die Übersicht erhebt jedoch insofern keinen Anspruch auf Vollständigkeit, als andernorts zitierte unveröffentlichte Manuskripte (z.B. Huffman & Ceci, 1997, zitiert nach Ruby & Brigham, 1997, S. 718) wegen mangelnder Nachvollziehbarkeit bzw. fehlender Qualitätskontrolle im Rahmen von „Peer review"-Verfahren ebensowenig berücksichtigt sind wie zahlreiche Studien, die lediglich in Form von Tagungsbeiträgen publiziert wurden (z.B. Joffe & Yuille, 1992, zitiert nach Greuel et al., 1998, S. 139f.).

Tabelle 3. Validitätsbefunde aus Experimentalstudien zur *Kriterienorientierten Inhaltsanalyse*

	Köhnken et al. (1995) E	Landry & Brigham (1992) E	Porter & Yuille (1996) E	Ruby & Brigham (1998) E	Steller et al. (1992) K	Vrij et al. (2000) E	Winkel & Vrij (1995) K	Wolf und Steller (1997) E	Hypothesenkonforme Befunde (%)	Hypothesenkonträre Befunde (%)
1. *Konsistenz*	–	<	>	<	>	–	>	–	38	25
2. *Unordnung*	>		–	>	–	–	>	>	57	0
3. *Details*	>	>	>	–	>	>	>	>	88	0
4. *Verknüpfungen*		>		<	>	–	>	>	67	17
5. *Interaktionen*	–*	>		>	–	–	>	>	57	0
6. *Gespräche*	–*	>		–	–			>	40	0
7. *Komplikationen*	–	–	–	>	>	–		>	43	0
8. *Ausgefallenes*	–	>	–	>	>	>	<	>	63	13
9. *Nebensächliches*	–	>	–	>	>	–	–	>	50	0
10. *Unverstandenes*	–				>			–	33	0
11. *Indirektes*			–	<	>		>	–	40	20
12. *Eigenseelisches*		>	–	<	>			>	60	20
13. *Fremdseelisches*	–	<			–	>		>	40	20
14. *Verbesserungen*	–	>	–	>	–	>	–	>	50	0
15. *Erinnerungslücken*	>	>	>	>	–	–	–	–	50	0
16. *Selbsteinwände*	–	>		–	–	–	–	–	14	0
17. *Eigenbelastung*		–		<	<			–	0	50
18. *Fremdentlastung*	–				–			–	0	0
19. *Deliktspezifisches*										
Gesamtscore		>				>	>		100	0

Anmerkung: „K" = Aussagen von Kindern bzw. Jugendlichen; „E" = Aussagen von Erwachsenen; „>" = in glaubhaften Aussagen signifikant häufiger vorhanden bzw. stärker ausgeprägt als in unglaubhaften Aussagen; „<" = in glaubhaften Aussagen signifikant seltener vorhanden bzw. schwächer ausgeprägt; „–" = kein signifikanter Unterschied zwischen glaubhaften und unglaubhaften Aussagen; Leerstelle = Kriterium bzw. Gesamtscore wurde nicht untersucht; * Kriterien wurden zu einem Kriterium zusammengefaßt.

Es muß zudem betont werden, daß zwischen den in Tabelle 3 aufgeführten Studien z.T. **erhebliche methodische Unterschiede** bestehen. Neben der in Tabelle 3 kenntlich gemachten Differenzierung nach dem Alter der Pbn (Kinder vs. Erwachsene) sollen einige weitere Unterscheidungsmerkmale an dieser Stelle zumindest genannt bzw. an Beispielen veranschaulicht werden. So waren, was die Anzahl ausgewerteter Aussagen angeht, etwa bei Landry und Brigham (1992) lediglich sechs erlebnisbezogene und sechs erfundene Aussagen Gegenstand der Analyse, während Steller, Wellershaus und Wolf (1992) immerhin 88 wahre mit 88 unwahren Aussagen verglichen. Ferner verwendeten Steller et al. (1992) ebenso wie Ruby und Brigham (1998) ein intraindividuelles Design, d.h. sämtliche Pbn gaben sowohl einen wahren als auch einen erfundenen Bericht ab. In allen anderen Untersuchungen beruhten die Validitätsbefunde auf dem Vergleich unabhängiger Gruppen. Auch bezüglich der Auswertung des Aussagematerials unterscheiden sich die einzelnen Studien mitunter deutlich. Hier ist zum einen der Trainingsstand der Auswerter bezüglich der Handhabung der *Kriterienorientierten Inhaltsanalyse* zu berücksichtigen. Während z.B. die Rater bei Ruby und Brigham (1998) nur eine 45-minütige Einweisung erhielten, erstreckte sich die Auswerterschulung bei Porter und Yuille (1996) immerhin über drei Tage. Das Aussagematerial von Vrij, Kneller und Mann (2000) wurde u.a. von einem Experten in der *Kriterienorientierten Inhaltsanalyse* ausgewertet. Zum anderen wurden bei der Auswertung unterschiedliche Ratingskalen verwendet. Während z.B. die Auswerter bei Vrij et al. (2000) lediglich zu entscheiden hatten, ob die Kriterien vorhanden seien oder nicht, beurteilte man etwa bei Porter und Yuille (1996) die Ausprägungsgrade anhand vierstufiger Skalen. Noch differenzierter wurde bei Wolf und Steller (1997) vorgegangen. Hier verwendete man zwar nur eine dreistufige Skala, allerdings wurden für die Kriterien 4 bis 18 jeweils sämtliche kriterienerfüllenden Textstellen getrennt skaliert und die entsprechenden Skalenwerte über die gesamte Aussage aufsummiert. Zu erwähnen ist auch, daß die in Tabelle 3 aufgelisteten Ergebnisse z.T. auf unterschiedlichen statistischen Analysemethoden basieren (z.B. U-Test mit Bonferroni-Korrektur bei Wolf & Steller [1997] vs. t-Test ohne Adjustierung des Alpha-Fehlers bei Steller et al. [1992]).

Während die genannten Abweichungen zwischen den experimentellen Studien eher untergeordnete methodische Aspekte betreffen, ist ein weiteres Unterscheidungsmerkmal mehr fundamentaler Natur und soll daher im folgenden genauer erläutert werden. So lassen sich, was die Art und Weise der Gewinnung erlebnisbezogener Aussagen betrifft, innerhalb der Experimentalstudien **drei grundsätzlich verschiedene Paradigmen der Datengewinnung** ausmachen. Ein Ansatz besteht darin, den Pbn einen Film über ein bestimmtes Ereignis vorzuführen und sie anschließend aufzufordern, einen möglichst umfassenden mündlichen Bericht über das filmisch inszenierte Geschehen abzugeben (**Film-Paradigma**: Köhnken et al., 1995; Vrij et al., 2000). Beim zweiten

Ansatz läßt man die Pbn autobiographische Begebenheiten schildern (**Autobiographisches Paradigma**: Steller et al., 1992; Wolf & Steller, 1997; Landry & Brigham, 1992; Ruby & Brigham, 1998; Winkel & Vrij, 1995). Beim dritten Ansatz schließlich wird das konkrete aussagerelevante Ereignis im Labor simuliert, wobei man die Pbn in das Geschehen involviert (**Scheinverbrechen-Paradigma**: Porter & Yuille, 1996). In allen drei experimentellen Paradigmen werden den so gewonnenen erlebnisbezogenen Aussagen Schilderungen gegenübergestellt, die sich jeweils auf die gleichen Themen beziehen, jedoch (weitgehend) frei erfunden sind. Um einen genaueren Eindruck von den verschiedenen experimentellen Vorgehensweisen zu vermitteln, werden im folgenden exemplarisch die Untersuchungen von Porter und Yuille (1996), Steller et al. (1992) sowie Vrij et al. (2000) beschrieben.

Beispiel für das Film-Paradigma: Vrij et al. (2000) zeigten in einem Laborexperiment 15 erwachsenen Pbn einen Film, in dem ein Diebstahl dargestellt wurde. Die Pbn hatten anschließend die Aufgabe, einen möglichst umfassenden Bericht über den Diebstahl abzugeben (Bedingung „truthful"). Weiteren 15 Pbn wurden nur grobe schriftliche Informationen über das filmisch inszenierte Delikt gegeben. Sie wurden instruiert, gegenüber einer Interviewerin einen erfundenen Bericht über den Diebstahl abzugeben, und zwar so überzeugend, als ob sie den Videofilm selber gesehen hätten (Bedingung „uninformed deception"). In allen Versuchsbedingungen wurden die Aussagen im Rahmen standardisierter Interviews aufgenommen. Die transkribierten Berichte wurden von zwei unabhängigen Auswertern auf das Vorhandensein bzw. Fehlen von zwölf Glaubhaftigkeitskriterien hin analysiert. Die Kriterien 6 (*Gespräche*), 10 (*Unverstandenes*), 11 (*Indirektes*), 12 (*Eigenseelisches*), 17 (*Eigenbelastung*), 18 (*Fremdentlastung*) und 19 (*Deliktspezifisches*) wurden nicht in die Auswertung einbezogen, weil sie den Autoren im Hinblick auf den Aussagegegenstand nicht angemessen erschienen. Es zeigte sich, daß vier Kriterien in der Bedingung „truthful" signifikant häufiger auftraten als in der Bedingung „uninformed deception" (3. *Details*; 8. *Ausgefallenes*; 13. *Fremdseelisches*; 14. *Verbesserungen*). Kein Kriterium verhielt sich erwartungskonträr. Auch die über die zwölf erhobenen Kriterien hinweg gebildeten Gesamtscores waren in der Gruppe „truthful" durchschnittlich signifikant höher als in der Bedingung „uninformed deception" (vgl. Tabelle 3).[2] An dem hier gewählten Filmparadigma läßt sich mit Steller et al. (1992, S. 157) generell kritisieren, daß die unabhängige Variable „Sachverhalt erlebt"

[2] Neben den beiden genannten Gruppen gab es noch eine dritte Versuchsbedingung, in welcher die Pbn ebenfalls unter Bereitstellung grober Informationen einen Bericht über den Diebstahl erfinden sollten. Allerdings wurden diese 15 Pbn über neun Kriterien bzw. die Logik der *Kriterienorientierten Inhaltsanalyse* in Kenntnis gesetzt, bevor sie ihre Aussagen ablegten (Bedingung „informed deception"). Durch den Vergleich dieser Gruppe mit den beiden anderen Versuchsbedingungen sollte überprüft werden, ob hinsichtlich der *Kriterienorientierten Inhaltsanalyse* aufgeklärte Personen eher in der Lage sind, ihren erfundenen Aussagen eine hohe inhaltliche Qualität im Sinne der „Undeutsch-Hypothese" zu verleihen bzw. einen Erlebnisbezug vorzutäuschen. Diese Hypothese wurde bestätigt.

vs. „Sachverhalt nicht erlebt“ durch die unabhängige Variable „Sachverhalt beobachtet“ vs. „Sachverhalt nicht beobachtet“ ersetzt ist. Insofern eignet sich das Paradigma nur eingeschränkt zur Überprüfung der „Undeutsch-Hypothese“, welche ja gerade auf die Erlebnisbezogenheit von Aussagen abstellt.

Beispiel für das autobiographische Paradigma: Steller et al. (1992) überprüften die „Undeutsch-Hypothese“ anhand experimentell gewonnener Aussagen von 88 Erst- bzw. Viertklässlern. Da die inhaltsorientierte Glaubhaftigkeitsbeurteilung in der forensischen Praxis vornehmlich bei Kinderaussagen über Sittlichkeitsdelikte Anwendung findet, sollten im Rahmen des experimentellen Designs wesentliche psychologische Bedingungen des forensisch relevanten Sachverhalts „sexueller Mißbrauch“ nachgestellt werden, ohne die Grenzen ethischer Zumutbarkeit für die Pbn zu überschreiten. Als wesentliche psychologische Charakteristika sexueller Mißbrauchserfahrungen werden von den Autoren „die Eigenbeteiligung des Aussagenden am Geschehen [...], die vorwiegend negative emotionale Tönung des Geschehens und der weitgehende Kontrollverlust über die Situation auf seiten des Betroffenen“ genannt (S. 161). Um eine Übertragbarkeit der Untersuchungsergebnisse auf den Bereich der Aussagen zu sexuellem Mißbrauch zu gewährleisten, wurden experimentelle Aussagethemen gewählt, für die ebenfalls die genannten Charakteristika typisch sind. Die kindlichen Pbn sollten im Rahmen eines Erzählwettbewerbs je einen erlebnisbezogenen und einen erfundenen Bericht über folgende zur Auswahl stehenden Themen abgeben: „eine Injektion bekommen“, „eine Operation als Patient erfahren“, „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 Pbn hatten eine Woche Zeit, um sich vorzubereiten. Die Berichte wurden in semistandardisierten Einzelinterviews erhoben. Als Außenkriterium für den Wahrheitsgehalt einer Geschichte dienten Angaben der Eltern. Bei der Auswertung der Aussagetranskripte durch drei unabhängige Rater wurden die Ausprägungsgrade für jedes einzelne Glaubhaftigkeitskriterium auf einer vierstufigen Skala eingeschätzt. Vergleiche der Mittelwerte bei erlebnisbezogenen und erfundenen Aussagen zeigten signifikante Unterschiede bei neun Kriterien (1. *Konsistenz*; 3. *Details*; 4. *Verknüpfungen*; 7. *Komplikationen*; 8. *Ausgefallenes*; 9. *Nebensächliches*; 10. *Unverstandenes*; 11. *Indirektes*; 12. *Eigenseelisches*; vgl. Tabelle 3). Dagegen ergaben sich keine signifikanten hypothesenkonformen Unterschiede bei den Kriterien aus der Kategorie „Motivationsbezogene Inhalte“; Kriterium 17 (*Eigenbelastung*) war sogar in den erfundenen Berichten signifikant stärker ausgeprägt. Die Autoren führen dies auf die Untersuchungssituation (Erzählwettbewerb) zurück, in welcher die falschaussagenden Pbn die Erwähnung der motivationsbezogenen Inhalte vermeintlich nicht mit einem potentiellen Verlust von Glaubhaftigkeit assoziierten. Neben der im Vergleich zum Filmparadigma erhöhten

externen Validität dieses Laborexperiments ist positiv hervorzuheben, daß durch die intraindividuelle Bedingungsvariation (jedes Kind erzählt eine erlebnisbezogene und eine erfundene Geschichte) eine strenge Parallelisierung beider Aussagebedingungen im Hinblick auf aussagerelevante Persönlichkeitsmerkmale gewährleistet wurde.

Beispiel für das Scheinverbrechen-Paradigma: Das Experiment von Porter und Yuille (1996) soll etwas ausführlicher beschrieben werden, weil es der in der vorliegenden Studie gewählten Vorgehensweise (s. Abschnitt 4.3) am ähnlichsten ist. Den insgesamt 60 erwachsenen Pbn wurde zunächst mitgeteilt, daß mit dem Experiment angeblich die Eignung eines am Psychologischen Institut neu eingestellten Sicherheitsbediensteten überprüft werden solle. Die Pbn wurden auf vier Gruppen (zu jeweils 15 Pbn) aufgeteilt. Die Pbn dreier Gruppen wurden, jeweils einzeln, beauftragt, einen Gelddiebstahl zu begehen. Sie hatten zehn Minuten Zeit, ein bestimmtes Büro im Institut aufzusuchen, dort eine Mappe mit einer 100 Dollar-Note zu suchen, das Geld zu entwenden, gegebenenfalls das Büro wieder aufzuräumen und anschließend zum Ausgangsort zurückzukehren. Dabei sollten sie so vorsichtig wie möglich vorgehen, um nicht von dem angeblichen Sicherheitsbediensteten ertappt zu werden. Die Pbn der verbleibenden Gruppe instruierte man zu einer ähnlichen, jedoch unverfänglichen Aktion. Sie sollten ebenfalls innerhalb von zehn Minuten die Mappe ausfindig machen und mitnehmen. Allerdings befand sich in dieser Bedingung kein Geld in der Mappe. Stattdessen wurde den Pbn mitgeteilt, daß sie die Mappe für einen Professor abholen sollten und anschließend weitere Anweisungen erhalten würden. Nach Ausführung dieser Handlungen wurde sämtlichen Pbn mitgeteilt, es solle auch noch getestet werden, wie gut die Angestellten des angeblichen Sicherheitsunternehmens in der Lage seien, Verdächtige zu verhören und deren Schuld bzw. Unschuld zu ermitteln. Daher würden sie auch noch von einem weiteren Angehörigen des vermeintlichen Sicherheitsunternehmens vernommen werden. Dieser sei darüber informiert, daß er wahrheitsgemäße und falsche Geständnisse ebenso wie zutreffende und wahrheitswidrige Tatabstreitungen zu hören bekommen werde. Die Gruppe von Pbn, die die unverfängliche Handlung ausgeführt hatten, wurde instruiert, ihr wahrheitsgemäßes Alibi zu schildern (Bedingung „truthful alibi“). Von den drei Gruppen, die den Diebstahl begangen hatten, sollte eine eine Aussage machen, die weitgehend den Tatsachen entsprach. Allerdings sollten diese Pbn so tun, als ob sie ebenfalls nur die unverfängliche Handlung begangen hätten, d.h. sie sollten anstelle des Diebstahls erzählen, sie hätten lediglich eine Mappe für einen bestimmten Professor abgeholt (Bedingung „partial deception“). Die zweite Gruppe von „Dieben“ sollte eine komplette Falschaussage in Form eines frei erfundenen Alibis machen (Bedingung „complete deception“). Die dritte Gruppe von „Dieben“ schließlich wurde aufgefordert, ein wahrheitsgemäßes Geständnis abzulegen (Bedingung „truthful confession“). Alle Pbn hatten 15 Minuten Zeit, um sich auf ihre Aussage vorzubereiten.

Ihnen wurde eine finanzielle Belohnung versprochen, falls es ihnen gelinge, den vernehmenden „Sicherheitsbediensteten“ von der Glaubhaftigkeit ihrer Bekundung zu überzeugen. Die Vernehmungen erfolgten in Form halbstandardisierter Befragungen durch geschulte Interviewer. Zwei unabhängige, geschulte Auswerter beurteilten anhand von vierstufigen Skalen die Ausprägungsgrade von zehn Glaubhaftigkeitskriterien in den Aussagetranskripten. Die Kriterien 4 (*Verknüpfungen*), 5 (*Interaktionen*), 6 (*Gespräche*), 10 (*Unverstandenes*), 13 (*Fremdseelisches*), 16 (*Selbsteinwände*), 17 (*Eigenbelastung*), 18 (*Fremdentlastung*) und 19 (*Deliktspezifisches*) wurden von der Auswertung ausgeschlossen, da sie nach Meinung der Autoren nicht auf die Diebstahlsimulation anwendbar waren. Wie Tabelle 3 zu entnehmen ist, konnten nur für drei Kriterien (1. *Konsistenz*; 3. *Details*; 15. *Erinnerungslücken*) signifikante hypothesenkonforme Unterschiede zwischen den glaubhaften Bedingungen einerseits („truthful alibi“ bzw. „truthful confession“) und den unglaubhaften Bedingungen andererseits („partial deception“ bzw. „complete deception“) gefunden werden. Signifikante hypothesenkonträre Gruppenunterschiede zeigten sich nicht. Als Erklärung für die vergleichsweise enttäuschenden Validitätsbefunde dieser Studie zogen die Autoren zum einen die mit den geringen Gruppengrößen einhergehende schwache statistische Teststärke in Betracht. Eine weitere potentielle Ursache sahen sie in der Natur des simulierten Sachverhalts bzw. in dem experimentellen Status der Pbn. So sei es hier nicht – wie in anderen Studien und in den meisten forensischen Realfällen – um die Glaubhaftigkeitsbeurteilung von mutmaßlichen Zeugen sondern von Tatverdächtigen gegangen; und dem aussagerelevanten Geschehen habe es insbesondere der für die Situation von (Opfer-) Zeugen typischen Komponente des Kontrollverlusts ermangelt.

Nachdem nun die verschiedenen methodischen Vorgehensweisen anhand konkreter Beispiele beschrieben wurden, soll eine **Gesamtbewertung der experimentellen Befundlage** versucht werden. In den drei Untersuchungen, in denen über alle analysierten Kriterien hinweg ein Gesamtscore gebildet wurde, zeigten sich diesbezüglich statistisch signifikante, hypothesenkonforme Unterschiede zwischen erlebnisbezogenen und konfabulierten Aussagen (s. Tabelle 3). Somit sprechen also auch die Ergebnisse der Experimentalstudien für die Validität der „Undeutsch-Hypothese“ bzw. der *Kriterienorientierten Inhaltsanalyse* als Gesamtsystem. Bewegt man sich jedoch auf der Ebene einzelner Kriterien, so ist die Befundlage eher heterogen. Lediglich die Befunde zu Kriterium 3 (*Details*) konnten konsistent die Validität dieses Indikators stützen (7 von 8 Studien [= 88%] erbrachten hypothesenkonforme Resultate, d.h. höherer Detailreichtum in erlebnisbezogenen Aussagen; vgl. Tabelle 3). Kriterium 4 (*Verknüpfungen*) wurde zwar immerhin in vier von sechs Studien bestätigt, allerdings liegt auch ein hypothesenkonträrer Befund vor. Bei den Kriterien 2 (*Unordnung*), 5 (*Interaktionen*), 8 (*Ausgefallenes*) und 12 (*Eigenseelisches*) verliefen immerhin jeweils mehr als die Hälfte der Validierungs-

versuche erfolgreich, wobei zu den Kriterien 8 und 12 jeweils auch ein hypothesenkonträres Resultat vorliegt. Die Kriterien 9 (*Nebensächliches*), 14 (*Verbesserungen*) und 15 (*Erinnerungslücken*) wurden nur durch die Hälfte der diesbezüglich durchgeführten Studien untermauert. Für die übrigen neun Kriterien (1. *Konsistenz*; 6. *Gespräche*; 7. *Komplikationen*; 10. *Unverstandenes*; 11. *Indirektes*; 13. *Fremdseelisches*; 16. *Selbsteinwände*; 17. *Eigenbelastung* und 18. *Fremdentlastung*) verliefen jeweils weniger als die Hälfte der Validierungsversuche erfolgreich, wobei für die Kriterien 1, 11 und 13 auch hypothesenwidrige Resultate vorliegen. Hinsichtlich Kriterium 17 stehen sogar keinem erwartungskonformen Befund zwei hypothesenwidrige Ergebnisse gegenüber. Kriterium 19 (*Deliktspezifisches*) konnte in keiner der Laborstudien überprüft werden, da es sich bei den experimentellen Aussagethemen nie um Sexualdelikte handelte.

Auch bei den experimentellen Studien besteht z.T. eine erhebliche Diskrepanz zwischen den einzelnen Untersuchungsergebnissen. Insbesondere in den Untersuchungen von Landry und Brigham (1992), Steller et al. (1992) sowie Wolf und Steller (1997) konnten vergleichsweise viele Einzelkriterien bestätigt werden. Da in diesen drei Untersuchungen das autobiographische Paradigma verwendet wurde, liegt der Schluß nahe, daß dieses Paradigma – möglicherweise aufgrund einer gegenüber den anderen Paradigmen erhöhten externen Validität – am besten zur Validierung der *Kriterienorientierten Inhaltsanalyse* geeignet ist. Dem ist jedoch entgegenzuhalten, daß in der Untersuchung von Ruby und Brigham (1998), die sich ebenfalls des autobiographischen Paradigmas bedienten, neben sieben validitätsstützenden auch fünf hypothesenkonträre Resultate erzielt wurden (s. Tabelle 3). Allerdings weist diese Studie unter den Untersuchungen mit dem autobiographischen Paradigma auch die größten methodischen Schwächen auf. So wurde hier z.B. gar nicht überprüft, ob sich die geschilderten autobiographischen Begebenheiten tatsächlich zugetragen hatten bzw. ob die angeblich erfundenen Aussagen wirklich nicht auf einem Erlebnishintergrund beruhten. Ferner war der Umfang der Aussagen mit durchschnittlich 255 Wörtern relativ gering. Die Frage, ob das autobiographische Paradigma möglicherweise besser geeignet ist als das Film-Paradigma und das Scheinverbrechen-Paradigma, muß vorerst unbeantwortet bleiben, da insbesondere zu den beiden letztgenannten Pardigmen noch zu wenige Studien vorliegen. Es ist zwar offenkundig, daß dem autobiographischen Paradigma eine höhere externe Validität zukommt als dem Filmparadigma. Jedoch erscheint es durchaus denkbar, daß sich mit dem Scheinverbrechen-Paradigma eine noch deutlichere Annäherung an forensische Aussagebedingungen erzielen läßt als mit dem autobiographischen Paradigma, sofern es gelingt, eine starke persönliche Involvierung der Pbn in das simulierte Delikt zu gewährleisten und die motivationale Ausgangslage insbesondere falschaussagender „Zeugen“ nachzustellen.

Faßt man die Erkenntnisse aus der Feld- und Experimentalforschung zusammen, so ergibt sich folgendes Resümee: Die Validität der *Kriterienorientierten Inhaltsanalyse* als Gesamtsystem und somit die Gültigkeit der „Undeutsch-Hypothese“ können prinzipiell als erwiesen betrachtet werden. In sämtlichen Feld- und Laboruntersuchungen, in denen über die jeweils berücksichtigten Einzelkriterien hinweg Gesamtscores gebildet wurden, manifestierten sich diesbezüglich hypothesenkonforme Effekte (höhere Gesamtscores bei erlebnisbezogenen Aussagen). Dagegen konnte die Validität der meisten Einzelkriterien, die ja gewissermaßen die operationalen Bestimmungsstücke der „Undeutsch-Hypothese“ bilden, nicht hinreichend gestützt werden. Überzeugend ist die Gesamtbefundlage nur bei Kriterium 3 (*Details*). Die Validität dieses Glaubhaftigkeitskriteriums konnte sowohl in Feld- als auch in Laboruntersuchungen konsistent bestätigt werden. Für alle übrigen Kriterien jedoch ist die Befundlage eher heterogen. In den experimentellen Untersuchungen konnte neben dem genannten Kriterium 3 kein weiteres Kriterium mit hoher Regelmäßigkeit bestätigt werden. Dagegen ergab die Analyse der Feldstudien, daß die diesbezügliche Befundlage zumindest die Validität von fünf Einzelkriterien untermauert (2. *Unordnung*; 3. *Details*; 4. *Verknüpfungen*; 5. *Interaktionen*; 6. *Gespräche*). Die nächstliegende Erklärung für die Diskrepanz zugunsten der Feldbefunde ist wohl, daß in vielen der bisher durchgeführten Experimente möglicherweise keine ausreichende externe Validität erzielt wurde. Dabei ist jedoch zu betonen, daß in manchen Laborstudien (z.B. Steller et al., 1992; Wolf & Steller, 1997; vgl. Tabelle 3) mehr Kriterien bestätigt wurden als in der wohl aussagekräftigsten Feldstudie (Lamb et al., 1997; vgl. Tabelle 2). Sowohl zwischen den einzelnen Experimentalbefunden als auch zwischen den einzelnen Feldbefunden bestehen z.T. erhebliche Diskrepanzen, über deren Ursachen nur spekuliert werden kann. Abgesehen von Kriterium 3 (*Details*) und – mit Einschränkungen – den Kriterien 2 (*Unordnung*), 4 (*Verknüpfungen*), 5 (*Interaktionen*) und 6 (*Gespräche*), kann die empirische Befundlage zur Validität der einzelnen Glaubhaftigkeitskriterien nicht als befriedigend eingestuft werden.

In diesem Zusammenhang ist auch zu erwähnen, daß in den bisherigen Validitätsstudien in erster Linie überprüft wurde, ob die Glaubhaftigkeitskriterien in erlebnisbezogenen Aussagen signifikant häufiger bzw. signifikant stärker ausgeprägt auftreten als in erfundenen Aussagen. Der Nachweis signifikanter hypothesenkonformer Unterschiede sagt jedoch nocht nicht sehr viel aus über den diagnostischen Nutzen der Kriterien. **Diagnostischer Nutzen** ist dann gegeben, wenn ein Kriterium in erlebnisbezogenen Aussagen sehr oft (idealerweise immer) und in sehr starker Ausprägung vorzufinden ist, aber in erfundenen Aussagen nur sehr selten (idealerweise nie) und allenfalls in schwacher Ausprägung auftritt. Demgegenüber kann sich ein signifikanter Unterschied jedoch auch schon dann ergeben, wenn beispielsweise ein Kriterium in erfundenen Aussagen oft bzw. in starker Ausprägung vorkommt, in erlebnisbezogenen Schilderungen aber

noch häufiger bzw. noch stärker ausgeprägt zu finden ist. Vor dem Hintergrund dieser Überlegung sind die in den Tabellen 2 und 3 zusammengefaßten Ergebnisse im Hinblick auf eine praktische Anwendung der Kriterien noch vorsichtiger zu interpretieren.

2.1.6.2 Treffsicherheit der inhaltsorientierten Glaubhaftigkeitsbeurteilung

In Tabelle 4 sind einige Forschungsergebnisse zur Treffsicherheit der inhaltsorientierten Glaubhaftigkeitsbeurteilung zusammengefaßt. Die Studien wurden unter den gleichen formalen Gesichtspunkten ausgewählt wie die Untersuchungen zur Gültigkeit der „Undeutsch-Hypothese“. Es handelt sich weitgehend um Studien, die auch schon in Abschnitt 2.1.6.1 berücksichtigt wurden. Neu hinzugekommen ist lediglich die Studie von Zaparniuk et al. (1995), die das in Abschnitt 2.1.6.1 beschriebene Film-Paradigma verwendeten. Die von Steller, Wellershaus und Wolf (1988, zitiert nach Steller, 1989, S. 144ff.) berichteten Trefferquoten stammen aus der gleichen Untersuchung wie die von Steller et al. (1992) berichteten Validitätsbefunde zur „Undeutsch-Hypothese“ (s. Abschnitt 2.1.6.1). Mit Ausnahme der Felduntersuchung von Krahé und Kundrotas (1992) handelt es sich ausschließlich um experimentelle Studien.

Bei der **Bestimmung der Treffsicherheit** geht es um die Frage, wie sehr die aus den Ergebnissen der *Kriterienorientierten Inhaltsanalyse* abgeleiteten diagnostischen Urteile (glaubhaft vs. unglaubhaft) mit dem objektiven Wahrheitsstatus der Aussagen (erlebnisbezogen vs. erfunden) übereinstimmen. Wie jedoch in den Abschnitten 2.1.3 und 2.1.5 erörtert wurde, existieren keine verbindlichen, standardisierten Richtlinien für die diagnostische Urteilsbildung. Daher besteht bei prospektiv angelegten empirischen Studien zur Überprüfung der Treffsicherheit grundsätzlich das Problem, auf welche Vorgehensweise man sich bei der diagnostischen Urteilsbildung festlegen soll. Diesbezüglich sind in den vorhandenen Studien **drei Ansätze** zu unterscheiden. Ein Ansatz besteht darin, daß ein Auswerter eine Aussage zunächst der *Kriterienorientierten Inhaltsanalyse* unterzieht und anschließend nach eigenem Gutdünken ein Urteil bezüglich der Glaubhaftigkeit der Aussage fällt. Die Ergebnisse der Inhaltsanalyse werden zwar bei der Urteilsbildung berücksichtigt, die Art und Weise der Datenintegration bleibt jedoch der Intuition des Auswerters überlassen (**klinisch-intuitive Urteilsbildung**, vgl. Tabelle 4). Das Urteil kann sowohl in dichotomer Weise als auch anhand feiner abgestufter Skalen (mit den Polen „äußerst glaubhaft“ und „äußerst unglaubhaft“) erfolgen und u.U. auch die Kategorie „unentscheidbar“ enthalten. Beim zweiten Ansatz erfolgt die Urteilsbildung anhand vorher exakt spezifizierter (quantitativer) Entscheidungsrichtlinien, d.h. das Urteil ergibt sich aus der Art bzw. der Anzahl der in der Aussage vorgefundenen Glaubhaftigkeitskriterien (**Urteilsbildung anhand Entscheidungsregel**, vgl. Ta-

belle 4). Beim dritten Ansatz werden die Aussagen ebenfalls anhand quantitativer Entscheidungsrichtlinien als glaubhaft vs. unglaubhaft klassifiziert. Allerdings werden die Entscheidungsregeln hier nicht a priori festgelegt, sondern post hoc mit Hilfe des statistischen Verfahrens der **Diskriminanzanalyse** (vgl. z.B. Bortz, 1999) ermittelt. Dabei wird diejenige (gewichtete) Linearkombination der Glaubhaftigkeitskriterien berechnet, die im jeweils gerade vorliegenden Datensatz die maximale Differenzierung zwischen erlebnisbezogenen und erfundenen Aussagen ermöglicht. Auf der Basis dieser Linearkombination werden Klassifikationsregeln berechnet, mit deren Hilfe die tatsächliche Bedingungszugehörigkeit der Aussagen vorhergesagt wird.

Tabelle 4. Treffer- und Fehlerquoten (in %) bei der Klassifikation erlebnisbezogener und erfundener Aussagen unter Anwendung der *Kriterienorientierten Inhaltsanalyse*

		Erlebnisbezogene Aussagen		Erfundene Aussagen	
		Treffer	Fehler	Treffer	Fehler
Klinisch-intuitive Urteilsbildung					
Landry & Brigham (1992)	E	75	25	35	65
Steller et al. (1988)	K	78	22	62	38
Vrij et al. (2000)	E	80	20	60	40
Durchschnitt		78	22	52	48
Urteilsbildung anhand Entscheidungsregel					
Ruby & Brigham (1998) (Gesamtstichpr.) [a]	E				
Regel: 5 oder mehr Kriterien		89	11	8	92
Regel: 7 oder mehr Kriterien		70	30	26	74
Zaparniuk et al. (1995)	E				
Regel: Krit. 1 bis 5 und 2 weitere [b]		80	20	77	23
Regel: Krit. 1 bis 3 und 4 weitere		68	32	77	23
Durchschnitt		77	23	47	53
Diskriminanzanalysen					
Köhnken et al. (1995)	E	88	12	62	38
Krahé & Kundrotas (1992)	E	88	12	78	22
Porter & Yuille (1996)	E	77	23	80	20
Ruby & Brigham (1998) [a]					
Aussagen farbiger Pbn	E	66	34	71	29
Aussagen weißer Pbn	E	74	26	67	33
Vrij et al. (2000) [c]	E	53	47	80	20
Wolf & Steller (1997)	E	100	0	100	0
Durchschnitt		78	22	77	23

Anmerkung: „E“ = Aussagen von Erwachsenen; „K“ = Aussagen von Kindern. [a] Bei Ruby & Brigham (1998) werden die mit Entscheidungsregeln erzielten Trefferquoten nur für die Gesamtstichprobe mitgeteilt, wohingegen die Diskriminanzanalysen getrennt für die Aussagen farbiger und weißer Pbn gerechnet wurden. [b] Diese Entscheidungsregel mußte in der Untersuchung modifiziert werden, da Kriterium 4 gar nicht erhoben wurde; die Autoren machen jedoch keine Angaben, wie die Regel abgeändert wurde. [c] Die Diskriminanzanalyse erfolgte für eine dreifach gestufte Gruppierungsvariable (Erläuterung im Text).

Die diskriminanzanalytisch ermittelten Trefferquoten sind zwar – insbesondere bei den erfundenen Aussagen – vergleichsweise hoch (s. Tabelle 4), sie besitzen jedoch aufgrund ihres Post-hoc-Charakters bzw. angesichts des weitgehenden **Fehlens von Kreuzvalidierungen** an unabhängigen Aussagestichproben praktisch keine Aussagekraft. Lediglich bei Ruby und Brigham (1998) sowie bei Köhnken et al. (1995) wurden die Diskriminanzfaktoren bzw. die diskriminanzanalytischen Klassifikationsregeln jeweils an einem Teil der Aussagestichprobe gewonnen und anschließend an den verbleibenden Aussagen kreuzvalidiert; die in Tabelle 4 aufgeführten diskriminanzanalytischen Trefferquoten dieser beiden Studien beruhen auf den Ergebnissen der jeweiligen Kreuzvalidierungen. Die diskriminanzanalytischen Trefferquoten sind auch insofern mit Zurückhaltung zu interpretieren, als die **diskriminanzanalytisch bestimmten Gewichte der einzelnen Kriterien von Studie zu Studie sehr stark abweichen** (sofern in den Publikationen überhaupt entsprechende Angaben gemacht werden). Es muß auch betont werden, daß die statistischen Vorgehensweisen in den einzelnen Untersuchungen unterschiedlich waren. Krahé und Kundrotas (1992) etwa rechneten getrennte Diskriminanzanalysen für die wahren und erfundenen Aussagen. Hierdurch wurden die resultierenden Trefferquoten (s. Tabelle 4) zwar artifiziell erhöht, logisch nachvollziehbar ist dieses Vorgehen jedoch nicht. Im Gegensatz dazu unterzogen Vrij et al. (2000) die über die Glaubhaftigkeitskriterien gebildeten Gesamtscores einer diskriminanzanalytischen Klassifikationsprozedur. Hier wurde also gar keine (optimale) Gewichtungsstruktur für die Einzelkriterien ermittelt, so daß die resultierende Trefferquote bei den erlebnisbezogenen Aussagen vergleichsweise niedrig ausfiel (s. Tabelle 4). Darüber hinaus muß zu den Trefferquoten von Vrij et al. (2000) auch angemerkt werden, daß hier die Gruppierungsvariable nicht zweifach sondern dreifach gestuft war, d.h. neben den Gruppen „truthful“ („Erlebnisbezogene Aussagen“ in Tabelle 4) und „uninformed deception“ („Erfundene Aussagen“ in Tabelle 4) gab es auch noch eine dritte Gruppe von Pbn, die eine Aussage erfinden sollten, die jedoch vorher über die Logik der *Kriterienorientierten Inhaltsanalyse* informiert worden waren (Bedingung „informed deception“; vgl. Abschnitt 2.1.6.1, insbesondere Fußnote 2). In der diskriminanzanalytischen Klassifikationsprozedur sollten alle drei experimentellen Gruppenzugehörigkeiten vorhergesagt werden. Dies dürfte die Rate an Fehlklassifikationen bei den erlebnisbezogenen (Gruppe „truthful“) und frei erfundenenen Aussagen (Bedingung „uninformed deception“) erhöht haben, verglichen mit einer diskriminanzanalytischen Klassifikationsprozedur, bei welcher ausschließlich diese beiden Gruppenzugehörigkeiten prädiziert werden sollen.

Die diskriminanzanalytisch gewonnenen Treffer- bzw. Fehlerquoten finden an dieser Stelle nur deshalb Erwähnung, weil sie auch in der einschlägigen Fachliteratur – dort aber zumeist in einer erstaunlich unkritischen Art und Weise – kolportiert werden. So

zitieren z.B. Vrij und Akehurst (1998) in ihrem Übersichtsartikel die Gesamttrefferquote aus der Untersuchung von Porter und Yuille (1996), ohne auch nur zu erwähnen, daß es sich hierbei um das Ergebnis einer diskriminanzanalytischen Klassifikationsprozedur handelt.

Wie Tabelle 4 zu entnehmen ist, wurden in den drei Studien mit klinisch-intuitiver Urteilsbildung durchschnittlich in etwa gleich hohe Trefferquoten erzielt wie in den beiden Untersuchungen, in denen unterschiedliche Entscheidungsregeln zur Anwendung kamen. Sowohl bei der klinisch-intuitiven Urteilsbildung als auch bei der Beurteilung anhand von Entscheidungsregeln bewegte sich die durchschnittliche Trefferquote in bezug auf die erfundenen Aussagen auf dem Zufallsniveau, während erlebnisbezogene Aussagen deutlich überzufällig korrekt klassifiziert wurden. Es muß allerdings betont werden, daß die Trefferquoten – insbesondere bei den erfundenen Aussagen – jeweils deutlichen Schwankungen zwischen den einzelnen Studien unterliegen (was freilich bei Verwendung unterschiedlicher Entscheidungsregeln nicht überrascht), so daß die Interpretation der Ergebnisse nur mit größten Vorbehalten erfolgen darf. Zudem ist die empirische Befundlage mit nur drei bzw. zwei Untersuchungen (vgl. Tabelle 4) vorläufig noch zu dünn, um eindeutige Schlußfolgerungen zuzulassen. Ferner ist zu den Studien von Landry und Brigham (1992) sowie Steller et al. (1988, zitiert nach Steller, 1989, S. 144ff.) einschränkend anzumerken, daß es hier neben den beiden Urteilskategorien „glaubhaft“ vs. „unglaubhaft“ jeweils auch noch eine dritte Kategorie für unentscheidbare Fälle gab. Die als unentscheidbar klassifizierten Aussagen machten in beiden Studien jeweils einen Anteil von 10% (gerundet) aus, sie wurden jedoch bei der Berechnung der in Tabelle 4 aufgeführten Trefferquoten nicht berücksichtigt, wodurch letztere erhöht wurden. Die Ergebnisse von Landry und Brigham (1992) verlieren außerdem auch dadurch an Aussagekraft, daß die Hälfte der Inhaltsanalysen nicht anhand von Transkripten sondern auf der Basis von Videoaufzeichnungen der Aussagen erfolgte. Die von Ruby und Brigham (1998) bzw. Zaparniuk et al. (1995) erprobten Entscheidungsregeln (vgl. Tabelle 4) wurden nicht theoretisch begründet, sondern beruhen auf heuristischen Empfehlungen von Experten (z.B. Yuille, 1990, zitiert nach Zaparniuk et al., 1995, S. 345) oder wurden rein explorativ untersucht.

Insbesondere in den Befunden zur klinisch-intuitiven Beurteilung zeigt sich durchgängig eine **Fehlertendenz in Richtung falsch positiver Urteile**, d.h. die Urteiler neigen in hohem Maße dazu, erfundene Aussagen als glaubhaft zu klassifizieren (s. Tabelle 4). Der Anteil falsch negativer Entscheidungen (erlebnisbezogene Aussagen werden für unglaubhaft befunden) ist dagegen deutlich niedriger, wenn auch immer noch von beträchtlicher Höhe.

Die Studie von Vrij et al. (2000) ist zwar hinsichtlich der experimentellen Manipulation des Erlebnisbezugs der Aussagen nur von geringer externer Validität (Film-Paradigma, s. Abschnitt 2.1.6.1), ihr kommt jedoch im Hinblick auf die Art und Weise der diagnostischen Urteilsbildung die vielleicht größte externe Validität zu. Hier wurden die Inhaltsanalysen und diagnostischen Urteile nämlich nicht, wie in anderen Studien, von mehr oder weniger kurzfristig geschulten Auswertern ohne forensische Erfahrung vorgenommen, sondern von einem ausgewiesenen Experten in der *Kriterienorientierten Inhaltsanalyse*. Allerdings standen diesem Experten, ebenso wie den Auswertern in den anderen Studien, lediglich die (transkribierten) Aussagen zur Verfügung, d.h. die Resultate der Inhaltsanalysen konnten im Rahmen der diagnostischen Datenintegration nicht an den Ergebnissen etwaiger Persönlichkeits- und Motivanalysen (vgl. Abschnitt 2.1.3) relativiert werden.[3] Insofern läßt sich gegen die Trefferquoten von Vrij et al. (2000) ebenso wie gegen die der übrigen in Tabelle 4 aufgeführten Studien der Einwand erheben, daß sie möglicherweise die wahre Treffsicherheit der inhaltsorientierten Glaubhaftigkeitsbeurteilung unterschätzen, weil **nicht alle für die Urteilsbildung relevanten Informationen zur Verfügung** standen. Kontrollierte Untersuchungen, in denen den Urteilern neben den Aussagen bzw. den inhaltsanalytischen Ergebnissen auch noch sämtliche begutachtungsrelevanten Daten aus gezielten Persönlichkeits- und Motivanalysen zur Verfügung gestanden hätten, liegen bis dato leider nicht vor. Die Beseitigung dieses Defizits sollte eines der primären Ziele zukünftiger aussagepsychologischer Forschung sein.

Zur Beurteilung des diagnostischen Werts der inhaltsorientierten Glaubhaftigkeitsbeurteilung gehört auch der Nachweis, daß Aussagebeurteilungen anhand der *Kriterienorientierten Inhaltsanalyse* zuverlässiger zwischen erlebnisbasierenden und erfundenen Aussagen differenzieren als **Beurteilungen ohne Rückgriff auf inhaltliche Glaubhaftigkeitskriterien**. Allerdings gab es nur in den Untersuchungen von Krahé und Kundrotas (1992), Landry und Brigham (1992), Ruby und Brigham (1998) sowie Steller et al. (1988, zitiert nach Steller, 1989, S. 144ff.) jeweils Kontrollbedingungen, in denen die Aussagen auch von naiven Personen (keine Verwendung der *Kriterienorientierten Inhaltsanalyse*) hinsichtlich ihrer Glaubhaftigkeit beurteilt wurden. Bei Krahé und Kundrotas (1992) erzielten die Auswerter, die auf die inhaltlichen Glaubhaftigkeitskriterien zurückgreifen konnten, eine tendenziell höhere Gesamttrefferquote als die naiven Beurteiler (74% vs. 63%). Allerdings machen die Autoren keine Angaben zu den Anteilen falsch positiver und falsch negativer Entscheidungen. (Dies ist auch der Grund, weshalb die Untersuchung in Tabelle 4 nicht unter der Kategorie „Klinisch-intuitive

[3] Was die Aussagemotive in experimentellen Studien betrifft, läßt sich dem jedoch entgegenhalten, daß hier die motivationale Ausgangslage der Pbn prinzipiell durch angemessene Instruktionen innerhalb der Versuchsbedingungen konstant gehalten werden kann.

Urteilsbildung“ berücksichtigt ist.) Bei Landry und Brigham (1992) wurden zwei Auswertergruppen, die ein Training in der *Kriterienorientierten Inhaltsanalyse* erhalten hatten, zwei Gruppen von naiven Beurteilern gegenübergestellt. Jeweils eine der trainierten bzw. naiven Auswertergruppen bekam Videoaufzeichnungen der zu beurteilenden Aussagen zu sehen, während den beiden verbleibenden trainierten bzw. naiven Gruppen nur Aussagetranskripte zur Verfügung standen. Die in der *Kriterienorientierten Inhaltsanalyse* geschulten Auswerter erzielten mit 55% (vgl. Tabelle 4) eine höhere Gesamttrefferquote als die naiven Urteiler (47%). Ferner stellte sich jedoch heraus, daß lediglich die Gruppe von geschulten Auswertern, die die Aussagevideos sahen, eine Gesamttrefferquote signifikant über dem Zufallsniveau erreichte (58%; „Unentscheidbar“-Urteile nicht berücksichtigt). Dagegen lagen die geschulten Auswerter, die nur die Transkripte lasen, in ihrer Treffsicherheit nicht signifikant über der Zufallswahrscheinlichkeit (53%; „Unentscheidbar“-Urteile nicht berücksichtigt). Da die *Kriterienorientierte Inhaltsanalyse* standardmäßig jedoch ausschließlich auf der Grundlage von Aussagetranskripten zu erfolgen hat, weil Kontaminierungen durch das aussagebegleitende nonverbale und paraverbale Ausdrucksverhalten vermieden werden sollen (vgl. Abschnitt 2.1.3), können die Ergebnisse von Landry und Brigham (1992) nicht als eindeutige Evidenz für die Überlegenheit der inhaltsorientierten gegenüber der naiven Glaubhaftigkeitsbeurteilung betrachtet werden. Auch bei Ruby und Brigham (1998) ergaben sich keine eindeutigen Hinweise auf eine Überlegenheit der inhaltsorientierten Glaubhaftigkeitsbeurteilung. Sowohl die in der *Kriterienorientierten Inhaltsanalyse* geschulten Auswerter als auch die naiven Beurteiler stuften die erfundenen Aussagen auf einer siebenstufigen Skala im Durchschnitt als glaubhafter ein als die erlebnisbezogenen Aussagen (4.5 vs. 4.05 bzw. 4.5 vs. 4.025). Beide Urteilergruppen lagen also in ihrer diagnostischen Leistung unter dem Zufallsniveau. Im Gegensatz zu den Befunden von Landry und Brigham (1992) sowie Ruby und Brigham (1998) spricht das Ergebnis von Steller et al. (1988, zitiert nach Steller, 1989, S. 144ff.) eindeutig für eine Überlegenheit der inhaltsorientierten gegenüber der naiven Glaubhaftigkeitsbeurteilung. Neben drei in der *Kriterienorientierten Inhaltsanalyse* geschulten Auswertern gab es auch noch 25 naive Beurteiler, die genau wie die geschulten Auswerter auf einer fünfstufigen Skala die Glaubhaftigkeit der experimentell gewonnenen Aussagen einstufen sollten (die mittlere Skalenstufe repräsentierte das Urteil „unentscheidbar“ und wurde von den naiven Beurteilern in 4.5% der Fälle gewählt). Bei der Klassifikation der Aussagen als glaubhaft vs. unglaubhaft waren die geschulten Auswerter den naiven Beurteilern signifikant überlegen. Während die geschulten Auswerter die erlebnisbasierenden Aussagen zu 78% zutreffend als glaubhaft und die erfundenen Schilderungen zu 62% korrekt als unglaubhaft diagnostizierten (s. Tabelle 4), kamen die naiven Beurteiler nur auf 68% valide positive und auf 47% valide negative Entscheidungen. Die Gesamttrefferquote (vermutlich ohne Berücksichtigung der unentscheidbaren Fälle) betrug bei den ge-

schulten Auswertern 72%, gegenüber 60% bei den naiven Beurteilern. Diese Gesamttrefferquoten decken sich nahezu mit denjenigen, die von Krahé und Kundrotas (1992) berichtet wurden (s.o.). Für die vier genannten Studien gilt aber ebenso der oben angesprochene Kritikpunkt, daß die inhaltsanalytisch geschulten Urteiler nicht die Möglichkeit hatten, die festgestellte inhaltliche Aussagequalität an Befunden etwaiger Persönlichkeits- und Motivanalysen zu relativieren.

Zusammenfassend kann festgehalten werden, daß die empirische Befundlage eher auf eine geringe Treffsicherheit der inhaltsorientierten Glaubhaftigkeitsbeurteilung hindeutet. Während die Trefferquoten in bezug auf die erlebnisbezogenen Aussagen immerhin noch recht deutlich über dem Zufallsniveau liegen, wurden die erfundenen Aussagen im Durchschnitt annähernd so oft falsch wie richtig beurteilt. Die erkennbare Urteilsverzerrung zugunsten falsch positiver Entscheidungen ist vor allem angesichts der Tatsache sehr bedenklich, daß die Methode in der Praxis vor allem zur Begutachtung von Anschuldigungen (sexueller Delikte) eingesetzt wird und somit möglicherweise zu einer Benachteiligung Unschuldiger führt. Selbst eine Überlegenheit der inhaltsorientierten gegenüber der naiven Glaubhaftigkeitsbeurteilung konnte in den wenigen diesbezüglich durchgeführten Untersuchungen nicht konsistent nachgewiesen werden. Die alles in allem eher ernüchternden empirischen Befunde zur Treffsicherheit der inhaltsorientierten Glaubhaftigkeitsbeurteilung sind jedoch insofern angreifbar, als die Diagnosen in sämtlichen Studien ausschließlich anhand der *Kriterienorientierten Inhaltsanalyse* erfolgten bzw. keine Relativierung der festgestellten inhaltlichen Aussagequalität an aussagerelevanten Persönlichkeitseigenschaften und Motiven erfolgte, wie es grundsätzlich erforderlich wäre. Zukünftige Forschung sollte ihr Augenmerk darauf richten, welche Treffsicherheit erzielt wird, wenn den Auswertern sämtliche für die diagnostische Urteilsbildung relevanten Informationen zur Verfügung stehen.

2.2 Psychophysiologische Glaubhaftigkeitsbeurteilung

Unter dem Oberbegriff „psychophysiologische Glaubhaftigkeitsbeurteilung“ werden eine Reihe diagnostischer Vorgehensweisen zusammengefaßt, bei denen man über die **systematische Erfassung und Auswertung unwillkürlicher physiologischer Begleiterscheinungen psychischer Prozesse** Rückschlüsse auf den Wahrheitsgehalt von Aussagen bzw. die Tatbeteiligung an kriminellen Vergehen zieht. Letztlich intendiert man hiermit zu **überprüfen, ob die Abstreitung der Täterschaft durch einen Beschuldig-**

ten glaubhaft ist oder nicht.[4] Vor allem umgangssprachlich wird für die psychophysiologischen Methoden häufig der Terminus **„Lügendetektion“** verwendet. Diese Begriffswahl ist jedoch insofern irreführend, als es keine wissenschaftlichen Hinweise auf ein qualitativ spezifisches körperliches Reaktionsmuster gibt, das in einem inter- oder intraindividuell konsistenten Zusammenhang mit Lügen auftritt und somit entdeckbar wäre (z.B. Podlesny & Raskin, 1977). Vielmehr werden die diagnostischen Schlußfolgerungen über die Glaubhaftigkeit nur mittelbar aus dem quantitativen Vergleich physiologischer Reaktionsstärken auf systematisch variierte Fragen bzw. Stimuli gezogen (s.u.). Um die ungünstigen Konnotationen des Begriffs „Lügendetektion“ zu vermeiden, wird in der deutschsprachigen Fachterminologie meist auf Umschreibungen wie etwa **„psychophysiologische Täterschaftsermittlung“** (Undeutsch, 1983a) oder **„psychophysiologische Aussagebeurteilung“** (Steller, 1987) zurückgegriffen. Teilweise wird auch – in Anlehnung an das im Rahmen der Methode eingesetzte Aufzeichnungsgerät (Mehrkanalschreiber bzw. Polygraph) – der Ausdruck **„Polygraphie“** (bzw. „Polygraphenmethode“, „Polygraphtechnik“ u.ä.; z.B. Eisenberg, 1993) verwendet. Der in der vorliegenden Arbeit gewählte Begriff **„psychophysiologische Glaubhaftigkeitsbeurteilung“** geht zurück auf Greuel et al. (1998).

Von zentraler Bedeutung bei den verschiedenen psychophysiologischen Verfahren sind die unterschiedlichen **Befragungs- bzw. Reizdarbietungstechniken**. Das gemeinsame **Grundprinzip** aller Verfahren besteht darin, sogenannte relevante, d.h. auf die inkriminierte Tat bezogene Reize (Fragen oder Items), sowie nicht tatbezogene Vergleichsreize (Fragen oder Items) in einer bestimmten Kombination darzubieten. Während der Befragung bzw. Stimuluspräsentation werden bestimmte physiologische Begleiterscheinungen beim Pb apparativ erhoben. Bei den erfaßten körperlichen Reaktionen handelt es sich in aller Regel um willentlich nur schwer beeinflußbare Indikatoren autonomer Erregung. Die in der Praxis gebräuchlichsten Maße sind die elektrische Hautleitfähigkeit, thorakale und abdominale Atembewegungen sowie die kardiovaskulären Größen relativer arterieller Blutdruck und Fingerpuls. In wissenschaftlichen Laboruntersuchungen erfaßt man häufig noch zusätzliche physiologische Maße, wie z.B. das Elektrokardiogramm und vereinzelt Blutvolumen des Fingers, Lidschlagfrequenz, Pupillenweite sowie evozierte Potentiale der hirnelektrischen Aktivität (vgl. z.B. Berning, 1992; Podlesny & Raskin, 1977; Schandry, 1989).

[4] Neben der Begutachtung von Tatverdächtigen wird die psychophysiologische Glaubhaftigkeitsbeurteilung mitunter auch eingesetzt, um Zeugenaussagen zu überprüfen (vgl. z.B. Lykken, 1998; Steller, 1987). Dieser vergleichsweise unbedeutende forensische Einsatzbereich wird in der vorliegenden Darstellung ebensowenig berücksichtigt wie die v.a. in den USA weitverbreiteten kommerziellen Anwendungen im privatwirtschaftlichen Bereich und öffentlichen Dienst. Bei letzteren geht es um die Einstellungsauslese und routinemäßige Testungen der Belegschaft; sie dienen also nicht der Aufklärung eines konkreten Tatvorwurfs, sondern zielen auf die „Ehrenhaftigkeit“ eines Arbeitnehmers im allgemeinen, seine politische Gesinnung, Loyalität o.ä. ab (vgl. z.B. Furedy & Heslegrave, 1991; Lykken, 1998).

Grob vereinfacht, beruhen sämtliche Verfahren auf der **Annahme**, daß Täter sich mit den tatbezogenen Reizen (und deren Beantwortung) gefühlsmäßig bzw. gedanklich stärker auseinandersetzen als mit den nicht tatbezogenen Vergleichsreizen (und deren Beantwortung). Die erhöhte emotionale bzw. kognitive Auseinandersetzung soll mit einer verstärkten physiologischen Aktivierung einhergehen. Dagegen wird für tatunbeteiligte Pbn angenommen, daß sie sich mit den nicht tatbezogenen Vergleichsreizen gefühlsmäßig bzw. gedanklich genauso sehr bzw. stärker auseinandersetzen als mit den tatbezogenen Stimuli und folglich bei letzteren nicht stärker reagieren als bei den Vergleichsreizen. Dementsprechend soll es möglich sein, aus dem intraindividuellen Vergleich der physiologischen Reaktionsstärken bei den tatbezogenen und nicht tatbezogenen Reizen diagnostische Schlußfolgerungen hinsichtlich der Täterschaft des Pb und somit der Glaubhaftigkeit der Abstreitung des Tatvorwurfs zu ziehen.

Hinsichtlich der Befragungstechniken lassen sich im wesentlichen zwei Gruppen von Vorgehensweisen gegenüberstellen. Man unterscheidet zwischen **direkten und indirekten Verfahren** (vgl. auch Tent, 1967). Bei den direkten Verfahren wird der Pb offen darauf angesprochen, ob er das aufzuklärende Delikt begangen habe. Somit kann die Glaubhaftigkeit der Verneinung des Tatvorwurfs direkt überprüft werden. Dahingegen überprüft man bei den indirekten Verfahren, ob der Pb tatspezifische Kenntnisse besitzt, die nur ein Tatbeteiligter haben kann. Aus einem positiven Testergebnis kann somit indirekt abgeleitet werden, daß die Abstreitung der Tatbeteiligung durch den Pb unglaubhaft ist. Die wichtigsten direkten und indirekten Verfahren der psychophysiologischen Glaubhaftigkeitsbeurteilung sind der **„Kontrollfragentest“** (englisch: „Control Question Test“) bzw. der **„Tatwissentest“** („Guilty Knowledge Test“ oder „Concealed Information Test“). Eine Modifikation des Tatwissentests, der sog. **„Guilty Actions Test“**, kann als eine Mischform aus direkten und indirekten Verfahren angesehen werden. Im folgenden werden die drei genannten Methoden der psychophysiologischen Glaubhaftigkeitsbeurteilung näher erläutert.

2.2.1 Der *Kontrollfragentest* (*KFT*)

Das in der – insbesondere US-amerikanischen – Praxis weitaus am häufigsten eingesetzte Verfahren der psychophysiologischen Glaubhaftigkeitsbeurteilung ist der *KFT*. Diese Methode wurde ursprünglich abseits der akademischen Psychologie entwickelt. Als ihr Erfinder gilt **John Reid** (Reid, 1947), ein Rechtsanwalt und ehemaliges Mitglied der Chicagoer Polizei. Die wichtigsten Beiträge zur methodischen Verfeinerung des *KFT* leistete **Cleve Backster** (z.B. Backster, 1962, 1963), ein Schüler Reids, der ebenfalls keine psychologische Grundausbildung genossen hatte. Da die Methode in keiner

akademisch-psychologischen Tradition steht[5], sind die in der Bezeichnung „Kontrollfragentest“ enthaltenen Begriffe „Kontrolle“ und „Test“ nicht im wissenschaftlichen Sinne zu verstehen. So beinhaltet das Verfahren, wie weiter unten zu zeigen sein wird, weder einen Vergleich zwischen einer Experimentalbedingung und einer sich davon nur in der interessierenden Variable unterscheidenden Kontrollbedingung, noch weist es den psychologische Tests auszeichnenden Grad an Standardisierbarkeit und Objektivität auf (vgl. z.B. Ben-Shakhar & Furedy, 1990). Um eine nachträgliche wissenschaftliche Untermauerung bzw. Evaluation des *KFT* hat sich später in erster Linie der US-amerikanische Psychologe **David Raskin** bemüht (z.B. Raskin, 1989; Podlesny & Raskin, 1977, 1978).

2.2.1.1 Diagnostische Vorgehensweise

In Tabelle 5 ist eine typische Fragensequenz des *KFT* wiedergegeben. Wie der Tabelle zu entnehmen ist, beinhaltet die Befragungstechnik **drei Fragetypen**: irrelevante (neutrale), relevante (tatbezogene) und Kontrollfragen. Die in Tabelle 5 als **irrelevant** gekennzeichneten Fragen sollen vom Pb mit „ja“. beantwortet werden. Die entsprechenden physiologischen Reaktionen des Pb gehen nicht in die Auswertung ein. Während die Fragen „I1“ und „I4“ (s. Tabelle 5) wenig stimulierend sind und der anfänglichen bzw. zwischenzeitlichen Beruhigung des Pb („reduce general nervous tension“, Matté, 1996, S. 325) dienen sollen, soll die Frage „I2“ zu den tatbezogenen Fragen überleiten sowie dem Pb die Gelegenheit geben, etwaige zu Beginn der Untersuchung aufgestaute Anspannung bzw. Angst abzubauen („dissipation of general nervous tension or undue anxiety prior to the asking of the primary relevant questions“, Matté, 1996, S. 325). Die Frage „I3“ soll die Gelegenheit geben, etwaige Befürchtungen einer unfairen Befragung zu äußern bzw. auszuräumen.

Die **relevanten Fragen** beziehen sich direkt auf die Tatbegehung (s. Tabelle 5) und sollen vom Pb verneint werden. Jeder relevanten Frage wird eine, vom Pb ebenfalls zu verneinende, **Kontrollfrage** anbei gestellt. Diese steht zwar in keinem direkten Zusammenhang mit dem konkreten Tatvorwurf, sie thematisiert jedoch ähnliche, emotional belastende Inhalte (s. Tabelle 5). Die Formulierung der Kontrollfragen wird innerhalb eines **Vortestinterviews** erarbeitet, das der eigentlichen Untersuchung vorangestellt ist. Im Verlauf dieses Vortestinterviews vertritt der Untersucher gegenüber dem

[5] Gleichwohl sehen sich die Befürworter des *KFT* (z.B. Reid & Inbau, 1977; Matté, 1996) in der Tradition des berühmten Psychologie-Professors Hugo Münsterberg, der u.a. die Registrierung physiologischer Veränderungen von Angeklagten und Zeugen als Hilfsmittel bei der Beurteilung von deren Glaubhaftigkeit angeregt hatte (vgl. Steller, 1987).

Pb die Ansicht, daß einer Person, welche die in den Kontrollfragen thematisierten Verfehlungen begangen habe, durchaus auch das in den relevanten Fragen angesprochene konkrete Verbrechen zugetraut werden könne. Dem Pb wird also gewissermaßen suggeriert, daß die Kontrollfragen auf seine allgemeine Glaubwürdigkeit (im Sinne eines zeitlich stabilen Persönlichkeitsmerkmals, vgl. Abschnitt 1) abzielten und daß die allgemeine Glaubwürdigkeit für die Beurteilung der speziellen Glaubhaftigkeit der Täterschaftsabstreitung von mitentscheidender Relevanz sei. Auf diese Weise soll der Pb dazu angestiftet werden, die Kontrollfragen zu verneinen, um den Verdacht gegen sich nicht zusätzlich zu erhärten. Zugleich wird dem Pb aber auch die Befürchtung induziert, daß wahrheitswidrige Verneinungen der Kontrollfragen erhöhte physiologische Reaktionen nach sich ziehen und somit vom Untersucher als Lügen entlarvt werden können, was wiederum zu einer zusätzlichen Verdachtserhärtung führen würde.

Tabelle 5. Typische Fragensequenz eines *KFT* (aus Undeutsch, 1983b, S. 405)

Fragetyp	Wortlaut der Fragen
I1	Ist Ihr Nachname ...?
I2	Bezüglich ... (Angabe der aufzuklärenden Angelegenheit) – haben Sie die Absicht, alle diesbezüglichen Fragen wahrheitsgemäß zu beantworten?
I3	Glauben Sie mir, daß ich Ihnen nur Fragen stellen werden, die wir zuvor vereinbart haben?
K1	Haben Sie vor Ihrem 19. Lebensjahr jemals irgendwelches Geld gestohlen?
R1	Haben Sie den betreffenden Ring genommen?
K2	Haben Sie während der Schulzeit irgendeinen Gegenstand von Wert entwendet?
R2	Haben Sie den betreffenden Ring aus der Schublade entwendet?
I4	Heißen Sie mit Vornamen ...?
K3	Haben Sie jemals einen Menschen, der berechtigt war, die Wahrheit zu erfahren, belogen, um sich unangenehme Konsequenzen fernzuhalten?
R3	Waren Sie irgendwie an dem Diebstahl des betreffenden Ringes beteiligt?

Anmerkung: „I“ = irrelevante Frage; „K“ = Kontrollfrage; „R“ = relevante Frage.

Gemäß der **zugrundeliegenden Annahme des *KFT*** fühlen sich unschuldige Pbn durch die Kontrollfragen in einem stärkeren Maße bedroht als durch die relevanten Fragen. Während sie nämlich die relevanten Fragen wahrheitsgemäß verneinen können, sind sie bei der Verneinung der Kontrollfragen entweder gezwungen zu lügen, oder sie sind sich hinsichtlich des Wahrheitsgehalts der Verneinung zumindest unsicher. Diese Unsicherheit wird zusätzlich durch die absichtlich vage Formulierung der Kontrollfragen provoziert. Die erhöhte emotionale Erregung bzw. kognitive Auseinandersetzung, die durch die Kontrollfragen ausgelöst wird, soll mit stärkeren physiologischen Reaktionen einhergehen, als es bei den relevanten Fragen der Fall ist. Im Gegensatz dazu sind für die schuldigen Pbn (Täter) die relevanten Fragen bedeutsamer bzw. bedrohlicher als die Kontrollfragen, da sie bei den relevanten Fragen den konkreten Tatvorwurf wahrheits-

widrig verneinen und gleichzeitig – gemäß der vorher vermittelten „Testlogik“ – die Aufdeckung dieser Lügen befürchten müssen. Dementsprechend sollen die schuldigen Pbn auf die relevanten Fragen physiologisch stärker reagieren als auf die Kontrollfragen.

Im Rahmen einer Fragensequenz werden **verschiedene Paare von relevanten und Kontrollfragen** dargeboten (meist 3, vgl. Tabelle 5). Die vollständige Fragensequenz wird normalerweise **mindestens dreimal präsentiert**. Während der Befragung werden im allgemeinen respiratorische (thorakale und abdominale Atembewegungen), elektrodermale (Hautleitfähigkeit oder -widerstand), und kardiovaskuläre Maße (v.a. Fingerpuls und relativer arterieller Blutdruck) registriert (vgl. Ben-Shakhar & Furedy, 1990).

Nach der Befragungsprozedur werden die Reaktionsstärken in den physiologischen Aufzeichnungen ausgewertet. Sind die Reaktionen auf die Kontrollfragen deutlich stärker ausgeprägt als auf die relevanten Fragen, dann wird die Abstreitung der Täterschaft als glaubhaft beurteilt. Ergibt sich ein entgegengesetztes Reaktionsmuster, mit deutlich stärkeren Reaktionen auf die relevanten Fragen, so wird die Verneinung der Tatbegehung als unglaubhaft eingestuft. Falls keine eindeutigen Reaktionsunterschiede zwischen den relevanten und Kontrollfragen zu erkennen sind, gilt das Ergebnis als unentscheidbar. Bezüglich der **Quantifizierung der physiologischen Reaktionen** und der darauf basierenden diagnostischen Urteilsbildung lassen sich zwei Ansätze unterscheiden: die global-intuitive und die numerische Auswertungsmethode.

Bei der **global-intuitiven Auswertungsmethode** (s. dazu Reid & Arther, 1953; Reid & Inbau, 1977) stützt der Untersucher sein Urteil auf den allgemeinen Eindruck, den er während der Testung hinsichtlich der physiologischen Reaktionen des Pb gewonnen hat. Die Reaktionen werden also nicht nach standardisierten Richtlinien quantifiziert. Bei der diagnostischen Urteilsbildung werden neben dem Gesamteindruck vom physiologischen Reaktionsmuster des Pb zusätzlich auch noch die fallspezifischen Hintergrundfakten sowie das Verhalten des Pb während der Untersuchung mit einbezogen. Die Integration der Daten aus den drei genannten Bereichen erfolgt allerdings in unspezifizierter Weise (Vrij, 1998b).

Im Gegensatz dazu erfolgt bei der **numerischen Auswertungsmethode** (z.B. Raskin, 1982, 1986) die Quantifizierung der Reaktionsstärken und die darauf basierende Urteilsbildung nach festgelegten Regeln. Für jedes dargebotene Paar von relevanten und Kontrollfragen (vgl. Tabelle 5) wird – getrennt nach den verschiedenen physiologischen Reaktionssystemen (Hautleitfähigkeit, Atmung etc.) – jeweils ein numerischer Wert von –3 bis +3 vergeben. Die Höhe des Wertes hängt davon ab, in welchem Größenverhältnis

die physiologische Reaktion auf die relevante Frage und die physiologische Reaktion auf die zugehörige Kontrollfrage stehen. Wenn z.B. die Amplituden der Hautleitfähigkeitsreaktionen um weniger als das Doppelte voneinander abweichen, werden null Punkte vergeben. Bei einem Größenverhältnis der Hautleitfähigkeitsreaktionen von mindestens 1:2, 1:3 oder 1:4 resultieren Werte von ±1, ±2 bzw. ±3. Die numerische Auswertung der respiratorischen und kardiovaskulären Reaktionen folgt dem gleichen Prinzip, allerdings werden hier andere Größenverhältnisse zwischen den physiologischen Reaktionen auf die relevanten und Kontrollfragen zugrunde gelegt (vgl. O'Toole, Yuille, Patrick & Iacono, 1994). Das Vorzeichen der Scores ist durch die Richtung der Reaktionsunterschiede festgelegt. Ein negatives Vorzeichen indiziert eine stärkere Reaktion auf die relevante Frage, ein positiver Wert bedeutet stärkere Reaktion auf die Kontrollfrage. Positive Werte weisen also auf eine glaubhafte Verneinung der relevanten Fragen hin. Die Punktwerte werden schließlich über alle Paare von zusammengehörigen relevanten und Kontrollfragen, über sämtliche Befragungsdurchgänge und über alle erfaßten physiologischen Reaktionssysteme hinweg zu einem Gesamtwert aufaddiert. Je nachdem, ob der numerische Gesamtscore eines Pb einen bestimmten positiven Cutoff-Wert überschreitet, einen bestimmten negativen Cutoff-Wert unterschreitet oder im Bereich der beiden Grenzwerte liegt, wird die Abstreitung der Täterschaft als glaubhaft, unglaubhaft bzw. unentscheidbar klassifiziert. In den meisten empirischen Studien und weitgehend auch in der Praxis werden die beiden Cutoffs auf +5 bzw. -5 oder auch auf +6 bzw. -6 veranschlagt, da sich mit diesen Entscheidungsregeln in Feld- und Laborstudien hohe Trefferquoten erzielen ließen (vgl. auch Steller, 1987). Die numerische Auswertungsmethode wurde nicht zuletzt mit dem Ziel entwickelt, eine Beeinflussung der diagnostischen Urteilsbildung durch außerhalb der physiologischen Messungen liegende Informationen (wie bei der global-intuitiven Auswertungsmethode) zu minimieren (Vrij, 1998b) und somit die Auswertungs- und Interpretationsobjektivität des diagnostischen Verfahrens zu erhöhen.

2.2.1.2 Anmerkungen zum Prozedere in der forensischen Praxis

Im vorangegangenen Abschnitt kam bereits zur Sprache, daß beim KFT neben der eigentlichen Befragungs- bzw. physiologischen Meßprozedur auch noch ein **Vortestinterview** stattfindet, dessen primäres Ziel in der Formulierung adäquater Kontrollfragen liegt. Daneben erfüllt das Vortestinterview auch noch andere Funktionen. So wird der Pb grob über die Funktionsweise des vegetativen Nervensystems und die technischen Grundlagen der physiologischen Meßapparatur informiert. Damit zusammenhängend wird auch die (z.T. vorgetäuschte) Logik des diagnostischen Verfahrens sowie seine Treffsicherheit (mitunter anhand fingierter Zahlenangaben) erläutert. Dies geschieht

letztlich mit der Absicht, den Pb von der (angeblichen) Elaboriertheit bzw. wissenschaftlichen Fundiertheit der diagnostischen Methode zu überzeugen, wodurch tatunbeteiligte Pbn beruhigt und tatbeteiligte Pbn in eine gespannte Erwartung versetzt werden sollen.

Dem Vortestinterview schließt sich in aller Regel auch noch ein sog. **„Stimulationstest“** an, durch den die Treffsicherheit der psychophysiologischen Glaubhaftigkeitsbeurteilung und die Kompetenz des Untersuchers demonstriert werden sollen (vgl. Steller, 1987). Als Stimulationstests werden Karten- oder Zahlentests verwendet. Ein Kartentest kann beispielsweise so ablaufen, daß der Pb aufgefordert wird, aus einer Reihe von verdeckten Spielkarten eine aufzudecken und sich den Inhalt (z.B. Bube) zu merken. Der Untersucher bekommt (mitunter nur scheinbar) nicht mit, um welche Karte es sich dabei handelt. Der Pb soll nun versuchen, dem Untersucher gegenüber das Wissen um den Inhalt der Spielkarte zu verheimlichen. Der Untersucher versucht, die verheimlichte Information aufzudecken, indem er den Pb systematisch befragt und zugleich mit Hilfe der Meßapparatur dessen physiologische Reaktionen registriert. Die Befragung erfolgt im Stile des weiter unten (Abschnitt 2.2.2) beschriebenen Tatwissentests. Der Untersucher nennt sukzessive die in Frage kommenden Alternativen (z.B.: „Handelt es sich bei der aufgedeckten Karte um den König ... die Dame ... den Buben ...“ etc.). Der Pb soll alle Alternativen verneinen. Anschließend beurteilt der Untersucher anhand der Kurvenausschläge in den physiologischen Aufzeichnungen, welche Karte der Pb versucht hat zu verheimlichen. Um sicherzustellen, daß der Untersucher in seinem Urteil nicht falsch liegt bzw. daß der Stimulationstest die intendierte Wirkung auf den Pb nicht verfehlt, werden die Durchführung bzw. das Ergebnis des Tests häufig fingiert, so etwa indem alle dem Pb zur Auswahl stehenden Spielkarten identischen und dem Untersucher von vorneherein bekannten Inhalts sind (z.B. nur Buben). Bei Zahlentests verfährt man in analoger Weise, nur daß es sich hier bei der verheimlichten Information um eine bestimmte Zahl handelt.

Neben dem Vortestinterview, dem Stimulationstest und der eigentlichen Befragungsprozedur spielt, v.a. in der US-amerikanischen Strafverfolgungspraxis, auch noch eine vierte Untersuchungsphase eine wichtige Rolle – die sog. **„Nachtestbehandlung“** (vgl. Steller, 1987, S. 14). Damit ist in erster Linie gemeint, daß im Falle eines positiven Testergebnisses (physiologische Reaktionen deuten auf unglaubhafte Abstreitung der Täterschaft hin) eine abschließende Vernehmung durchgeführt wird, in welcher der Untersucher versucht, den Pb unter Hinweis auf den Testbefund zu einem Geständnis zu bewegen.

Die obigen Erläuterungen zur praktischen Durchführung psychophysiologischer Glaubhaftigkeitsbegutachtungen mit dem *KFT* beschränken sich auf die wichtigsten Aspekte. Für ausführliche praxisbezogene Anleitungen sei auf die Standardwerke von Reid und Inbau (1977) sowie Matté (1996) verwiesen.

2.2.1.3 Problematik

An dieser Stelle können die Kritikpunkte am *KFT* nicht umfassend dargestellt werden, da dies den Rahmen der vorliegenden Arbeit sprengen würde. Hier sollen lediglich die größten Schwachstellen aufgezeigt werden. Für eine ausführliche kritische Würdigung des *KFT* bzw. eine vergleichende Gegenüberstellung der direkten und indirekten Verfahren der psychophysiologischen Glaubhaftigkeitsbeurteilung sei exemplarisch auf Ben-Shakhar und Furedy (1990), Furedy (1993, 1996), Furedy und Heslegrave (1991), Lykken (1998) sowie Raskin (1989) verwiesen.

Ähnlich wie die inhaltsorientierte Glaubhaftigkeitsbeurteilung (vgl. Abschnitt 2.1) wurde auch der *KFT* nicht etwa im Rahmen wissenschaftlicher Forschung, sondern in der forensischen Praxis entwickelt. Während es sich jedoch bei den wichtigsten Pionieren (Undeutsch, Arntzen) sowie den Anwendern der inhaltsorientierten Glaubhaftigkeitsbeurteilung um akademisch ausgebildete Psychologen handelt, war sowohl bei den Entwicklern (Reid, Backster) als auch beim Großteil der Anwender des *KFT* von Anfang an keine enge Anbindung an die akademische Psychologie gegeben (vgl. Lykken, 1998). Dadurch ist es möglicherweise auch zu erklären, daß die inhaltsorientierte Glaubhaftigkeitsbeurteilung immerhin von einer psychologischen Grundannahme („Undeutsch-Hypothese“, s. Abschnitt 2.1.1) abgeleitet wurde (deren systematische Überprüfung in Form kontrollierter Untersuchungen freilich lange auf sich warten ließ), wohingegen die Konzeption und Weiterentwicklung des *KFT* völlig **atheoretisch** erfolgten. Hier wurde die diagnostische Vorgehensweise nicht aus überprüfbaren, geschweige denn empirisch abgesicherten, psychologischen bzw. psychophysiologischen Theorien hergeleitet, sondern sie beruht auf **stark simplifizierenden, naiv-wissenschaftlichen Vorstellungen von psychischen Vorgängen und ihren Zusammenhängen mit körperlichen Prozessen** (Lykken, 1998). Diese naiven Annahmen wurden jedoch von den Befürwortern des *KFT* kaum spezifiziert, sondern lassen sich weitgehend nur post hoc aus der praktischen Vorgehensweise rekonstruieren.

Eine zentrale Grundannahme ist in der Befragungstechnik und Auswertungsmethode des *KFT* unmittelbar offensichtlich. Man geht davon aus, daß alle schuldigen Pbn die relevanten Fragen als bedrohlicher erachten und dementsprechend stärker darauf reagie-

ren als auf die Kontrollfragen. Dagegen sollen alle unschuldigen Pbn in den relevanten Fragen eine geringere Bedrohung sehen und dementsprechend schwächer reagieren als bei den Kontrollfragen. Lykkens (1998, S. 119ff.) Analyse zufolge verbergen sich hinter der diagnostischen Prozedur des *KFT* noch drei weitere implizite Annahmen: (1) Bei den relevanten Fragen gehen unglaubhafte Antworten eines beliebigen Pb immer mit stärkeren Reaktionen einher als glaubhafte Antworten derselben Person. (2) Ein sachkundiger Untersucher formuliert die Kontrollfragen immer so, daß der Pb sie während der Befragung unglaubhaft beantwortet oder hinsichtlich des Wahrheitsgehalts seiner Antworten zumindest unsicher bzw. besorgt ist. (3) Ein beliebiger Pb reagiert, sofern er schuldig ist, auf eine relevante Frage stärker als auf die dazugehörige Kontrollfrage. Ist derselbe Pb jedoch unschuldig, so fällt seine Reaktion bei der relevanten Frage schwächer aus als bei der Kontrollfrage.

Analog zur Validierung der inhaltsorientierten Glaubhaftigkeitsbeurteilung (vgl. Abschnitt 2.1.6) ist auch die Validierung des *KFT* grundsätzlich als zweistufiger Prozeß anzusehen. Wissenschaftlichen Prinzipien gemäß müßte der Bestimmung der Treffsicherheit der Methode eigentlich die Verifizierung ihrer theoretischen Grundannahmen vorausgehen. Da der *KFT* jedoch nur auf impliziten Hypothesen basiert (s.o.), wurde die erste Validierungsstufe hier völlig übergangen. Die oben genannten (impliziten) Grundannahmen des *KFT* wurden weder empirisch bestätigt, noch stehen sie im Einklang mit anderweitig gesicherten psychologischen bzw. psychophysiologischen Erkenntnissen, wonach die Zusammenhänge zwischen psychischen Prozessen und physiologischen Veränderungen sehr komplex und variabel sind (vgl. Schandry, 1989; Vossel & Zimmer, 1998). So ist nicht davon auszugehen, daß alle Pbn die unterschiedlichen Fragetypen in der postulierten Art und Weise als mehr oder weniger bedrohlich bzw. bedeutsam bewerten. Ebenso wenig ist die Annahme gerechtfertigt, die Stärke der physiologischen Reaktionen hänge in direkter, konstanter Weise von der empfundenen Bedrohlichkeit ab (Rill & Vossel, 1998).

Abgesehen von den genannten impliziten Annahmen hat man sich jedoch auch mehr oder weniger fundierter lern-, emotions-, und motivationspsychologischer Konzepte bedient, um die postulierten differentiellen Reaktionsmuster tatbeteiligter und tatunbeteiligter Pbn im *KFT* zu erklären. Eines davon, das sog. **„lie-guilt-arousal link“-Konzept** (vgl. Furedy & Ben-Shakhar, 1990, S. 12), besagt, daß Lügen (vermutlich sozialisationsbedingt) mit Bestrafung und somit mit Schuld assoziiert wird. Das Schuldgefühl wiederum soll mit einer erhöhten körperlichen Aktivierung einhergehen. Dies habe zur Folge, daß man bei unglaubhaften Antworten im *KFT* stärkere physiologische Reaktionen zeige als bei glaubhaften Antworten. Demgegenüber wird in der **„Theorie der Angst vor Bestrafung“** („lie-threat-arousal link“, Furedy & Ben-Shakhar, 1990, S. 12)

davon ausgegangen, daß unglaubhafte Antworten im *KFT* mit Besorgnis bezüglich der negativen Konsequenzen einhergehen, die bei Aufdeckung der Lügen drohen. Diese Besorgnis soll sich in einer erhöhten physiologischen Aktivierung bei unglaubhaften Antworten manifestieren (z.B. Raskin, 1979; Raskin & Kircher, 1989). Die **„Theorie der konditionierten Reaktion“** beinhaltet, daß die während der Verübung eines Verbrechens auftretende erhöhte autonome Erregung des Täters über klassische Konditionierung an die mit dem Delikt zusammenhängenden Reize und somit auch an die relevanten (tatbezogenen) Fragen des *KFT* gekoppelt wird, so daß Täter hier stärker reagieren als Tatunbeteiligte (vgl. Davis, 1961). Der **„Konflikttheorie“** zufolge gerät man als Täter bei der Beantwortung der relevanten Fragen in einen Konflikt zwischen zwei unvereinbaren Reaktionstendenzen, nämlich lügen vs. glaubhaft antworten. Dieser Konflikt soll eine emotionale Reaktion auslösen, die sich in einer erhöhten körperlichen Aktivierung manifestiere (vgl. Ben-Shakhar & Furedy, 1990). Es ist jedoch zu betonen, daß die genannten psychologischen Konzepte nicht speziell auf den *KFT* abstellen, sondern auch zur Erklärung der Reaktionsweise in anderen psychophysiologischen Verfahren der Glaubhaftigkeitsbeurteilung, wie etwa dem weiter unten beschriebenen Tatwissentest (s. Abschnitt 2.2.2), herangezogen werden können. Experimentelle Studien, in denen die einzelnen Modelle systematisch überprüft wurden, haben jedoch ergeben, daß keiner der postulierten Kausalmechanismen die differentiellen Reaktionsmuster von tatbeteiligten und tatunbeteiligten Pbn bei der psychophysiologischen Glaubhaftigkeitsbeurteilung hinreichend erklären kann. Dies gilt für den *KFT* ebenso wie für die anderen psychophysiologischen Verfahren (s. dazu zusammenfassend Ben-Shakhar & Furedy, 1990; Steller, 1987).

Blendet man die defizitäre theoretische Fundierung des *KFT* einmal völlig aus, so ist jedoch auch in praktischer Hinsicht ein fundamentaler Schwachpunkt des Verfahrens hervorzuheben, nämlich die **Abhängigkeit des Untersuchungsergebnisses von der Kompetenz des Untersuchers**. Wie oben dargestellt wurde, müssen die Kontrollfragen so beschaffen sein, daß sie für einen Tatunbeteiligten bedrohlicher sind und dementsprechend stärkere physiologische Reaktionen auslösen als die relevanten Fragen. Umgekehrt sollen sie aber beim Täter schwächere Reaktionen hervorrufen als die tatbezogenen Fragen. Wie die Kontrollfragen lauten müssen, damit sie diese Bedingungen erfüllen, ist von der Persönlichkeit des Pb und dem inkriminierten Tatbestand abhängig. Es ist die Aufgabe des Untersuchers, im Rahmen des Vortestinterviews mit dem Pb adäquate Formulierungen zu erarbeiten. Dies setzt jedoch eine außerordentliche psychologische Kompetenz des Untersuchers voraus. Er muß zum einen befähigt sein, in reliabler Weise festzustellen, wie bedrohlich der Pb verschiedene Versionen von Kontrollfragen empfindet, wie unsicher er sich hinsichtlich des Wahrheitsgehalts seiner Antworten auf diese verschiedenen Versionen ist oder gar, ob er auf die Kontrollfragen

lügt (Lykken, 1979). (Man beachte, daß die gesamte psychophysiologische Prozedur sich eigentlich erübrigen würde, wenn der Untersucher auch ohne Hilfsmittel Lügen identifizieren könnte, wie dies in bezug auf die Kontrollfragen im Vortestinterview vorausgesetzt wird.) Weiterhin muß der Untersucher in der Lage sein, den Pb in mehrfacher Hinsicht erfolgreich zu täuschen. Einerseits soll der Pb in den (Irr-) Glauben versetzt werden, daß das Eingeständnis der in den Kontrollfragen thematisierten Verfehlungen ihn auch im Hinblick auf den konkreten Tatvorwurf zusätzlich belaste. Andererseits soll er (fälschlicherweise) annehmen, unglaubhafte Verneinungen der Kontrollfragen seien anhand der registrierten körperlichen Veränderungen identifizierbar bzw. starke Reaktionen bei den Kontrollfragen hätten negative Auswirkungen auf das Untersuchungsergebnis (das Gegenteil trifft zu). Ferner muß der Untersucher den Pb in manipulativer Weise davon überzeugen, daß der *KFT* nahezu unfehlbar sei (vgl. Abschnitt 2.2.1.2). Es ist kaum anzunehmen, geschweige denn erwiesen, daß selbst ein gut ausgebildeter Untersucher all diese hochdiffizilen Aufgaben im Rahmen des Vortestinterviews erfolgreich bewältigen kann.

Mag ein Untersucher auch noch so viel Geschick hinsichtlich der Manipulation des Pb und der Kontrollfragen haben, so **kann er im Einzelfall doch überhaupt nicht feststellen, ob die im Vortestinterview erarbeiteten Kontrollfragen die postulierten Bedingungen erfüllen**. Dies ergibt sich aus rein logischen Erwägungen (vgl. dazu Lykken, 1998, S. 122): Um wirklich sicher sein zu können, daß die Kontrollfragen bei einem gegebenen Pb angemessen sind, müßte der Untersucher sowohl wissen, wie stark der Pb als Schuldiger auf die relevanten Fragen reagieren würde als auch wie stark er als Tatunbeteiligter auf die relevanten Fragen reagieren würde. Nur in diesem Fall könnte der Untersucher nämlich verifizieren, daß die Reaktionen des Pb auf die Kontrollfragen einen angemessenen Vergleichsstandard darstellen, d.h. daß sie kleiner sind als die Reaktionen desselben Pb bei unglaubhaft beantworteten relevanten Fragen und größer als die Reaktionen desselben Pb bei glaubhaft beantworteten relevanten Fragen. Der Untersucher kennt jedoch nicht den wahren Status des Pb (Täter vs. Tatunbeteiligter), geschweige denn dessen hypothetische Reaktionsstärken bei glaubhafter vs. unglaubhafter Beantwortung der relevanten Fragen. (Würde er den wahren Status des Pb kennen, wäre die psychophysiologische Glaubhaftigkeitsbeurteilung gänzlich überflüssig. Würde er die hypothetischen Reaktionsstärken des Pb bei glaubhafter vs. unglaubhafter Beantwortung der relevanten Fragen kennen, wäre die Formulierung der Kontrollfragen überflüssig, da man die konkreten Reaktionsstärken bei den relevanten Fragen direkt mit den hypothetischen Reaktionsstärken bei glaubhafter vs. unglaubhafter Beantwortung dieser Fragen vergleichen könnte.)

Es bleibt also festzuhalten, daß eine adäquate Formulierung der Kontrollfragen äußerst unwahrscheinlich ist, da sie einerseits enorme und noch nicht genau spezifizierte psychologische Fertigkeiten seitens des Untersuchers voraussetzt, und da sich andererseits im Einzelfall überhaupt nicht feststellen läßt, welche Kontrollfragen einen angemessenen Vergleichsstandard für die relevanten Fragen repräsentieren. Je nach Art der Fehlfomulierung der Kontrollfragen können sich unterschiedliche Konsequenzen für den Untersuchungsausgang ergeben. Sind die Kontrollfragen im Vergleich zu den relevanten Fragen zu sehr emotional belastend, steigt das Risiko, daß sie auch bei einem schuldigen Pb gleich starke oder stärkere Reaktionen hervorrufen als die relevanten Fragen. Im umgekehrten Fall (zu schwache Formulierung der Kontrollfragen) besteht erhöhte Gefahr, daß ein unschuldiger Pb als unglaubhaft eingestuft wird. Wie in Abschnitt 2.2.1.4 deutlich werden wird, deuten die empirisch ermittelten Trefferquoten des *KFT* darauf hin, daß die letztgenannte Art der Fehlformulierung von Kontrollfragen wahrscheinlicher ist. Damit ist auch schon ein weiterer kritischer Aspekt der Methode angesprochen. Die empirische Befundlage verweist auf eine **Fehlertendenz in Richtung falsch positiver Entscheidungen** (Unschuldige werden als Täter klassifiziert).

Neben der vermeintlichen Fehlformulierung der Kontrollfragen kommen für diese diagnostische Urteilsverzerrung natürlich auch noch andere Gründe in Frage. So ist es z.B. naheliegend, **daß auch unschuldige Pbn die relevanten Fragen als solche erkennen** und dementsprechend starke physiologische Reaktionen zeigen. Normalerweise versucht man starker Erregung unschuldiger Pbn bei den relevanten Fragen entgegenzuwirken, indem man die Pbn im Rahmen des Vortestinterviews und des Stimulationstests von der „Unfehlbarkeit“ des *KFT* überzeugt, so daß ein unschuldiger Pb nicht befürchten muß, seine Verneinungen der relevanten Fragen könnten fälschlich als Lügen interpretiert werden. Ob es allerdings gelingt, dem Pb Glauben an die „Unfehlbarkeit“ des *KFT* zu induzieren, hängt wiederum weitestgehend von der psychologischen Kompetenz des Untersuchers ab und ist im konkreten Einzelfall nicht überprüfbar. Vor dem Hintergrund, daß der *KFT* in erster Linie zur Begutachtung von Beschuldigten eingesetzt wird, ist die diagnostische Urteilsverzerrung in Richtung falsch positiver Entscheidungen nicht mit dem Rechtsgrundsatz „in dubio pro reo“ vereinbar.

Die fehlende Standardisierbarkeit des *KFT* und die damit verbundene Abhängigkeit des Untersuchungsergebnisses von der Kompetenz des Untersuchers bringen die gleichen **Einschränkungen für die Bestimmung der diagnostischen Treffsicherheit** mit sich, die auch in bezug auf die inhaltsorientierte Glaubhaftigkeitsbeurteilung gelten (vgl. Abschnitt 2.1.5). **Allgemeingültige**, d.h. nicht an bestimmte Untersucher gebundene, **Validitätsangaben** sind – streng genommen – gar **nicht gestattet** (Ben-Shakhar & Furedy, 1990). Wenn in Abschnitt 2.2.1.4 dennoch empirisch ermittelte Trefferquoten des *KFT*

berichtet werden, so deshalb, weil der *KFT* schon seit langem auf breiter Front Anwendung findet und somit eine kritische Würdigung seiner Treffsicherheit aus pragmatischen Erwägungen unbedingt geboten ist.

2.2.1.4 Treffsicherheit

Zur kriterienbezogenen Validität des *KFTs* liegen sowohl zahlreiche **Feld- als auch experimentelle Studien** vor. In Feldstudien wird überprüft, inwiefern in forensischen Realfällen durchgeführte *KFTs* bzw. die dabei aufgezeichneten physiologischen Messungen eine zutreffende Klassifizierung der Pbn (schuldig vs. unschuldig) bzw. der Täterschaftsabstreitungen (glaubhaft vs. unglaubhaft) ermöglichen. Die Problematik der Feldforschung liegt zum einen in einer hohen Selektivität des Untersuchungsmaterials. Zum anderen kann der tatsächliche Status des Pb bzw. der Wahrheitsgehalt der Täterschaftsabstreitung nicht mit letzter Sicherheit bestimmt werden. Stattdessen zieht man als Validierungskriterien Gerichtsurteile, Entscheidungen von Expertengruppen oder Geständnisse von Beschuldigten heran; alle diese Kriterien sind jedoch mit Fehlern behaftet. Experimentelle Studien bieten demgegenüber den Vorteil, daß der tatsächliche Status des Pb bzw. der Wahrheitsgehalt der Täterschaftsabstreitung einer perfekten Kontrolle des Forschers unterliegen und daß auch die Randbedingungen, unter denen der *KFT* stattfindet, streng kontrollierbar sind. Realisiert wird dies in aller Regel, indem man die Pbn einer Experimentalgruppe („Schuldige“[6] bzw. „Täter“) auffordert, ein Verbrechen zu simulieren (Scheinverbrechen-Paradigma) und anschließend in einem *KFT* die Täterschaft abzustreiten, während die Pbn der Kontrollgruppe („Unschuldige“ bzw. „Tatunbeteiligte“) lediglich von dem simulierten Delikt in Kenntnis gesetzt werden und sich ebenfalls einem *KFT* unterziehen müssen. Allerdings leiden Experimentalstudien, verglichen mit Felduntersuchungen, unter einem Mangel an externer Validität. Eine ausführlichere Erörterung der Vor- und Nachteile von Feld- und Experimentalstudien auf dem Gebiet der forensischen Glaubhaftigkeitsforschung erfolgt in Abschnitt 3.2.2.

In den Tabellen 6 und 7 sind die Resultate einiger Feld- bzw. Experimentalstudien zusammengefaßt. Die Bestandsaufnahme enthält Untersuchungen, die auch in anderen

[6] In der englischsprachigen Literatur werden die Bezeichnungen „guilty“ und „innocent“ nicht nur zur Klassifikation von Tätern bzw. Nichttätern in forensischen Realfällen verwendet, sondern auch zur Kennzeichnung der Bedingungszugehörigkeit von Pbn in Scheinverbrechen-Experimenten. Auch in der vorliegenden Arbeit werden experimentelle Pbn je nach Gruppenzugehörigkeit plakativ als „schuldig“ bzw. „unschuldig“ bezeichnet. Es muß jedoch betont werden, daß diese Bezeichungen im Kontext experimenteller Simulationen nicht im moralischen oder juristisch-normativen Sinne zu verstehen sind, sondern ausschließlich der Vermeidung umständlicher Umschreibungen dienen.

Übersichtsarbeiten zur psychophysiologischen Glaubhaftigkeitsbeurteilung (Ben-Shakhar & Furedy, 1990; Berning; 1992, 1993; Steller, 1987) berücksichtigt sind. Da der Umfang der vorliegenden Arbeit ansonsten ausufern würde, soll hier keine Beschreibung der Vorgehensweisen in den einzelnen Studien erfolgen. Diesbezüglich sei insbesondere auf die oben genannten Arbeiten von Berning (1992) und Steller (1987) verwiesen.

Es muß allerdings betont werden, daß die in Tabelle 6 und Tabelle 7 zusammengefaßten Feld- bzw. Experimentalstudien sich jeweils in mehrfacher Hinsicht voneinander unterscheiden. Wie Tabelle 6 zu entnehmen ist, differieren die Feldstudien z.B. hinsichtlich der verwendeten Auswertungsmethode (numerisch vs. global-intuitiv). Weiterhin unterscheiden sich die Feldstudien in bezug auf das gewählte Validierungskriterium. So wurden die Fälle bei Barland und Raskin (1976) sowie Bersh (1969) anhand von Entscheidungen von Expertengremien „verifiziert“, wohingegen etwa bei Horvath (1977) Geständnisse der Beschuldigten als Validierungskriterium dienten. Ein weiteres Unterscheidungsmerkmal bei den Feldstudien betrifft die Repräsentativität der Stichprobe. Während etwa die Stichprobe von Patrick und Iacono (1991) die verschiedenartigsten Delikte umfaßt, handelt es sich bei den von Bersh (1969) analysierten Fällen ausschließlich um solche aus dem militärischen Bereich. In die Untersuchungen von Bersh (1969) und Horvath (1977) wurden nur solche Fälle aufgenommen, bei denen das diagnostische Urteil aus dem *KFT* nicht in die Kategorie „unentscheidbar“ fiel. Die Ergebnisse von Honts (1996) sind schon aufgrund des geringen Stichprobenumfangs wenig aussagekräftig. Ferner unterscheiden sich die Feldstudien danach, ob die Auswertung durch bezüglich der Gruppenzugehörigkeit der Pbn blinde Auswerter erfolgte (z.B. Rafky & Sussman, 1985) oder durch informierte Auswerter (wie etwa bei Honts, 1996; s. dazu Lykken, 1998, S. 134f.).

Läßt man die diagnostischen Urteile der Kategorie „unentscheidbar“ außer acht, so beträgt die durchschnittliche, in Feldstudien erzielte Treffsicherheit des *KFT* hinsichtlich der Glaubhaftigkeitsbeurteilung von schuldigen Pbn 93.5% (ungewichteter Mittelwert aller 7 Feldstudien, s. Tabelle 6). Die durchschnittliche Trefferquote bei den unschuldigen Pbn beläuft sich dagegen, bei Nichtberücksichtigung der unentscheidbaren Fälle, nur auf 73.7%. Berücksichtigt man auch die unentscheidbaren Fälle, die bei den unschuldigen Pbn mit 20.7% durchschnittlich sehr viel häufiger vorkommen als bei den schuldigen Pbn (5.8%), so liegt die durchschnittliche Trefferquote sowohl bei den schuldigen (87.5%) als auch bei den unschuldigen Pbn (62.3%) deutlich niedriger, wobei die mittlere Fehlerquote bei den unschuldigen Pbn weiterhin deutlich höher ist als bei den schuldigen Pbn (19.9% vs. 6.4%). Die Resultate der Feldstudien verweisen somit im Durchschnitt auf eine deutliche Fehlertendenz in Richtung falsch positiver Ent-

scheidungen. Diese diagnostische Urteilsverzerrung tritt in relativ konsistenter Weise in den einzelnen Feldstudien auf. In vier der sieben Feldstudien war die Fehlerquote bei den unschuldigen Pbn deutlich höher als bei den schuldigen Pbn (s. Tabelle 6). Lediglich in den Studien von Bersh (1969), Honts (1996) sowie Rafky und Sussman (1985) zeigte sich keine (deutliche) Benachteiligung der unschuldigen Pbn. Auffallend sind die großen Schwankungen der Trefferquoten zwischen den einzelnen Studien, und zwar insbesondere im Hinblick auf die unschuldigen Pbn. Bei letzteren liegen die Trefferquoten – unter Nichtberücksichtigung der unentscheidbaren Fälle – in einem Bereich von 45.5% (Barland & Raskin, 1976) bis 100% (Honts, 1996).

Tabelle 6. Mit dem *KFT* erzielte Treffer- und Fehlerquoten (in %) in Feldstudien

		Schuldige				Unschuldige		
Studie	N	richtig	falsch	unentsch.	N	richtig	falsch	unentsch.
Barland & Raskin (1976) n	47	83.0 (97.5)	2.1 (2.5)	14.9	17	29.4 (45.5)	35.3 (54.5)	35.3
Bersh (1969) g	104	85.6 (85.6)	14.4 (14.4)	–	112	89.3 (89.3)	10.7 (10.7)	–
Honts (1996) n	7	100 (100)	0 (0)	0	6	83.3 (100)	0 (0)	16.7
Honts & Raskin (1988) n	12	91.7 (100)	0 (0)	8.3	13	61.5 (80)	15.4 (20)	23.1
Horvath (1977) g	28	77.2 (77.2)	22.8 (22.8)	–	28	51.1 (51.1)	48.9 (48.9)	–
Patrick & Iacono (1991) n	52	92.3 (98.0)	1.9 (2.0)	5.8	37	29.7 (55.0)	24.3 (45.0)	45.9
Rafky & Sussman (1985) n	30	90.8 (96.5)	3.3 (3.5)	5.8	30	91.7 (94.8)	5 (5.2)	3.3
Durchschnitt (ungewichtet)		87.5 (93.5)	6.4 (6.5)	5.8		62.3 (73.7)	19.9 (26.3)	20.7

Anmerkung: „n“ = numerische Auswertung; „g“ = global-intuitive Auswertung. Werte in Klammern ergeben sich bei Nichtberücksichtigung der unentscheidbaren Fälle.

Die in Tabelle 7 angeführten experimentellen Untersuchungen weisen in bezug auf das methodische Vorgehen eine größere Homogenität auf als die Feldstudien. In sämtlichen Untersuchungen bezogen sich die *KFT*s auf simulierte Diebstähle. Mit Ausnahme der Studie von Bradley und Janisse (1981) versuchte man in allen Untersuchungen, die Pbn über finanzielle Anreize dazu zu motivieren, im *KFT* möglichst als unschuldig klassifiziert zu werden. In sämtlichen Untersuchungen wurde eine numerische Auswertung der physiologischen Reaktionen vorgenommen, meistens durch Auswerter, die keine Kenntnis von der Gruppenzugehörigkeit der Pbn besaßen. Allerdings bestehen auch

zwischen den experimentellen Studien einige Unterschiede, so v.a. in bezug auf die rekrutierten Stichproben. Während an den meisten Untersuchungen Studenten teilnahmen (z.B. Barland & Raskin, 1975; Bradley, 1988), bestanden die Stichproben bei Patrick und Iacono (1989) sowie Raskin und Hare (1978) jeweils aus Häftlingen. Bei Bradley (1988) konnten die Pbn ihre Versuchsbedingung (schuldig vs. unschuldig) selber auswählen, während die Grupppenzuteilung in allen anderen Untersuchungen per Randomisierung erfolgte. Auch bestand der *KFT* in den meisten Untersuchungen aus drei Befragungsdurchgängen; die in Tabelle 7 aufgeführten Trefferquoten von Dawson (1980) sowie Honts, Raskin und Kircher (1987) beruhen dagegen auf vier bzw. fünf Durchgängen.

Tabelle 7. Mit dem *KFT* erzielte Treffer- und Fehlerquoten (in %) in experimentellen Studien

	Schuldige				Unschuldige			
Studie	N	richtig	falsch	unentsch.	N	richtig	falsch	unentsch.
Barland & Raskin (1975)	36	63.9 (88.5)	8.3 (11.5)	27.8	36	41.7 (71.4)	16.7 (28.6)	41.7
Bradley (1988)	20	85.0 (100)	0 (0)	15.0	56	53.6 (66.7)	26.8 (33.3)	19.6
Bradley & Janisse (1981)	96	60.4 (81.7)	13.5 (18.3)	26.0	96	58.3 (86.2)	9.4 (13.8)	32.3
Dawson (1980)	12	100 (100)	0 (0)	0	12	75.0 (81.8)	16.7 (18.2)	8.3
Honts et al. (1985)	19	100 (100)	0 (0)	0	19	47.4 (64.3)	26.3 (35.7)	26.3
Honts et al. (1987)	10	80.0 (100)	0 (0)	20.0	10	70.0 (77.8)	20.0 (22.2)	10.0
Kircher & Raskin (1988)	50	60.0 (93.7)	4.0 (6.3)	36.0	50	76.0 (97.4)	2.0 (2.6)	22.0
Patrick & Iacono (1989)	24	83.3 (87.0)	12.5 (13.0)	4.2	24	41.7 (55.6)	33.3 (44.4)	25.0
Podlesny & Raskin (1978)	20	75.0 (83.3)	15.0 (16.7)	10.0	20	85.0 (94.4)	5.0 (5.6)	10.0
Raskin & Hare (1978)	24	87.5 (100)	0 (0)	12.5	24	75.0 (94.7)	4.2 (5.3)	20.8
Rovner et al. (1979) [a]	24	87.5 (100)	0 (0)	12.5	24	87.5 (91.3)	8.3 (8.7)	4.2
Durchschnitt (ungewichtet)		80.2 (94.0)	4.8 (6.0)	14.9		64.7 (80.1)	15.3 (19.4)	20.0

Anmerkung: Werte in Klammern ergeben sich bei Nichtberücksichtigung der unentscheidbaren Fälle.
[a] Zahlenangaben aus Steller (1987, S. 39).

Auch in den experimentellen Studien liegt die mittlere Trefferquote bei den schuldigen Pbn (94.0% bzw. 80.2%; ohne bzw. mit Berücksichtigung unentscheidbarer Fälle) deutlich höher als bei den unschuldigen Pbn (80.1% bzw. 64.7%, s. Tabelle 7). Auch hier zeigen sich in relativ konsistenter Weise höhere Fehlerquoten bei den unschuldigen als bei den schuldigen Pbn. Nur bei drei der elf Studien (Bradley & Janisse, 1981; Kircher & Raskin, 1988; Podlesny & Raskin, 1978) wurden die schuldigen Pbn häufiger fehlklassifiziert als die unschuldigen, während es sich bei den übrigen acht Untersuchungen umgekehrt verhielt. Somit sprechen auch die Ergebnisse der Experimentalstudien recht eindeutig für eine Fehlertendenz des *KFT* in Richtung falsch positiver Entscheidungen. Zudem weisen die Experimentalstudien ebenfalls deutliche Schwankungen in den erzielten Trefferquoten auf, insbesondere was die Glaubhaftigkeitsbeurteilung unschuldiger Pbn betrifft. Bei letzteren streuen die Trefferraten zwischen 55.6% (bei Berücksichtigung der unentscheidbaren Fälle 41.7%) bei Patrick und Iacono (1989) sowie 97.4% (bzw. 76%) bei Kircher und Raskin (1988).

Zusammengefaßt spricht die empirische Befundlage aus Feld- und Laboruntersuchungen v.a. für eine **diagnostische Urteilsverzerrung des *KFT* zuungunsten unschuldiger Pbn**. Diese werden mit einer deutlich größeren Wahrscheinlichkeit fehlklassifiziert als schuldige Pbn. Des weiteren weisen sowohl die Feld- als auch die Experimentaluntersuchungen starke Schwankungen bezüglich der erzielten Trefferquoten auf, und zwar insbesondere bei den unschuldigen Pbn. Die nächstliegenden Erklärungen für diese Validitätsmängel des *KFT* (Kontrollfragenproblematik, geringe Standardisierung, Abhängigkeit von der Kompetenz des Untersuchers) wurden in Abschnitt 2.2.1.3 bereits erörtert.

2.2.2 Der *Tatwissentest* (*TWT*)

Der *TWT* wurde (im Gegensatz zum *KFT*) in einem wissenschaftlichen Umfeld entwikkelt. Sein Erfinder ist der US-amerikanische Psychologe **David Lykken** (z.B. 1959, 1960, 1991, 1998). Praktische Anwendung findet diese Methode jedoch vornehmlich in Israel und Japan (vgl. Ben-Shakhar & Furedy, 1990).

2.2.2.1 Diagnostische Vorgehensweise

Als indirekte Methode der psychophysiologischen Glaubhaftigkeitsbeurteilung zielt der *TWT* nicht direkt auf den Wahrheitsgehalt der Täterschaftsabstreitung ab, sondern es **soll festgestellt werden, inwiefern ein Beschuldigter Kenntnisse bezüglich spezifi-**

scher, tatbezogener Detailinformationen besitzt, die nur einem Tatbeteiligten bekannt sein können. Hierzu werden wichtige Details des Tathergangs als **relevante Items** in eine Reihe gleich plausibler, nicht tatbezogener Alternativen (**irrelevante Items**) eingebettet und in Form von Multiple-Choice-Fragen dargeboten. Aus dem Vorhandensein bzw. Fehlen differentieller physiologischer Reaktionen des Pb auf die tatbezogenen vs. nicht-tatbezogenen Items werden Rückschlüsse auf sein Tatwissen und – indirekt – auf seine Täterschaft gezogen. In Tabelle 8 ist ein exemplarisches Befragungsschema eines *TWT* wiedergegeben.

Tabelle 8. Exemplarischer Fragen- und Itemkatalog eines Tatwissentests (aus Steller, 1987, S. 11)

1.	Welche Nummer hatte der Raum, in dem der Diebstahl begangen wurde? War es
	a) Raum 321, b) Raum 214, c) Raum 411*, d) Raum 206, e) Raum 129, f) Raum 217?
2.	Wo befand sich der Gegenstand, der gestohlen wurde? War er
	a) in dem Schrank, b) auf dem Bücherregal, c) in dem Aktenbock, d) auf der Fensterbank, e) in der Schublade*, f) in dem Ablagekasten?
3.	Was für ein Gegenstand war es, der gestohlen wurde? War es
	a) ein Armband, b) eine Uhr*, c) eine Brosche, d) ein Ring, e) eine Kette, f) eine Geldbörse?
4.	Welche Farbe hatte das Armband der Uhr, die gestohlen wurde? War es
	a) rot, b) golden, c) braun, d) schwarz*, e) silber, f) weiß?
5.	Welcher Buchstabe stand auf dem Briefumschlag, in dem sich die Uhr befand? War es
	a) F, b) A, c) Z, d) M, e) K, f) E*?
6.	Welche Farbe hatte der Briefumschlag, in dem die Uhr war? War es
	a) blau, b) grün, c) gelb*, d) grau, e) braun, f) rot?

Anmerkung: * relevantes Item.

Wie Tabelle 5 zu entnehmen ist, werden im Rahmen einer Untersuchung i.d.R. mehrere Multiple-Choice-Fragen gestellt, die sich auf unterschiedliche tatbezogene Sachverhalte beziehen. Im Regelfall werden zu jeder Frage sechs Antwortalternativen formuliert (1 relevantes und 5 irrelevante Items). Bei der ersten Antwortalternative handelt es sich stets um ein irrelevantes Item, das als Puffer dient und nicht in die Auswertung eingeht. Bei der Auswahl der relevanten Items muß man darauf achten, daß es sich hierbei nicht um allzu marginale Tatdetails handelt, die sich möglicherweise der Kenntnis des Täters entziehen. Die irrelevanten Antwortalternativen müssen aus dem gleichen Inhaltsbereich stammen wie die relevanten Items. Sie müssen zudem für Tatunbeteiligte ebenso plausibel sein, d.h. mit der gleichen Wahrscheinlichkeit zutreffen können, wie die relevanten Items. Alle Antwortalternativen einer Mehrfachwahlfrage sollen in etwa gleich lang sein. Die Position der relevanten Antwortalternative innerhalb der Itemsequenz ist von Frage zu Frage zu variieren. Der zeitliche Abstand zwischen den Darbietungen der

einzelnen Items muß konstant gehalten werden. In der Standardprozedur des *TWT* werden die Pbn aufgefordert, auf alle Items mit einer einfachen Verneinung zu reagieren.[7]

Auch beim *TWT* findet vor der eigentlichen Untersuchung ein **Vortestinterview** statt, um die Tatkenntnisse des Pb zu erkunden. Stellt sich dabei heraus, daß der Pb auf „unverfängliche“ Weise (z.B. über die Medien oder im Rahmen polizeilicher Verhöre) Kenntnis bestimmter Tatdetails erlangt hat, so dürfen letztere nicht als relevante Items im *TWT* verwendet werden. Im Gegensatz zum *KFT* wird jedoch im *TWT* der genaue Wortlaut der Fragen und Items nicht im vorhinein mit dem Pb besprochen. Ihm wird lediglich angekündigt, auf welche Bereiche sich die eigentlichen Testfragen beziehen werden (z.B. Art des entwendeten Gegenstands). Die Ankündigung der Fragenkomplexe erfolgt auch mit dem Ziel, die deliktbezogenen Gedächtnisinhalte tatbeteiligter Pbn vorzuaktivieren.

Die **Grundannahme** des Verfahrens ist, daß nur Tatbeteiligte die zutreffenden Sachverhalte identifizieren bzw. von den irrelevanten Antwortalternativen differenzieren können. Das Wiedererkennen der Tatdetails soll sich in erhöhten physiologischen Reaktionen, verglichen mit den Reaktionsstärken bei den irrelevanten Antwortalternativen, manifestieren. Tatunbeteiligten Pbn hingegen sind die zutreffenden Alternativen nicht bekannt. Dementsprechend sollten ihre physiologischen Reaktionsstärken bei den relevanten und irrelevanten Items einer *TWT*-Frage zufällig variieren. Da die Items des *TWT* zu relativ schwachen physiologischen Reaktionen führen, beschränkt man sich bei der praktischen Anwendung meist auf einen relativ empfindlichen psychophysiologischen Indikator, nämlich die Amplituden der Hautleitfähigkeitsreaktionen. Bei der Auswertung der Meßkurven wird überprüft, ob der Pb konsistent auf die relevanten Items stärkere physiologische Reaktionen zeigt als auf die irrelevanten Items. Ist dies der Fall, so wird daraus die Schlußfolgerung gezogen, daß der Pb über tatspezifische Kenntnisse verfügt bzw. daß sein Abstreiten der Tatbeteiligung unglaubhaft ist.

Für die **Auswertung** des *TWT* wurde ein **numerisches System** entwickelt (Lykken, 1959). Demnach wird die Amplitude der elektrodermalen Reaktion nach dem relevanten Item mit den Reaktionsamplituden nach den irrelevanten Items derselben Frage verglichen (wobei die Reaktion auf die erste Antwortalternative [Pufferitem] nicht berücksichtigt wird, s.o.). Zeigt sich innerhalb der Sequenz von Antwortalternativen die stärkste Reaktion auf das relevante Item, so werden zwei Punkte vergeben. Zieht das relevante Item die zweitstärkste Reaktion nach sich, so wird ein Punkt vergeben. Ansonsten

[7] Allerdings kann der Test auch so durchgeführt werden, daß der Pb das jeweils vorgegebene Item wiederholt, bejaht oder gar nicht verbal reagiert (vgl. Steller, 1997).

ergeben sich null Punkte. Schließlich addiert man die Punkte über alle Fragen auf. Der resultierende Summenscore bildet die Grundlage für das diagnostische Urteil. Je höher die Gesamtpunktzahl ausfällt, desto mehr deutet dies auf das Vorhandensein tatspezifischer Kenntnisse und somit auf Tatbeteiligung des Pb hin. Der Cutoff-Wert für die Diagnose „Tatwissen vorhanden" richtet sich grundsätzlich nach dem akzeptierten Fehlerrisiko. Lykkens (1959) Empfehlung zufolge schließt man auf vorhandenes Tatwissen, wenn mehr als die Hälfte der möglichen Punkte erzielt werden (z.B. elf Punkte in einem *TWT* mit zehn Fragen).

Es wurde bereits erwähnt, daß dem *TWT* in der Praxis ebenso wie dem *KFT* ein Vortestinterview vorangehen sollte. Es muß jedoch betont werden, daß das Vortestinterview beim *TWT* nicht in erster Linie der psychologischen Manipulation des Pb dient (wie dies beim *KFT* der Fall ist), sondern der Abklärung des deliktbezogenen Kenntnisstandes des Pb bzw. dem Ausschluß unangemessener, d.h. dem Pb auf unverfängliche Weise bekannt gewordener, relevanter Items aus dem Befragungsschema. Es ist auch nicht unabdingbarer Bestandteil der Logik des *TWT*, daß man den Pb vor der Untersuchung durch entsprechende (falsche) Informationen sowie (fingierte) Stimulationstests von der „Unfehlbarkeit" der psychophysiologischen Glaubhaftigkeitsbeurteilung überzeugt. Gleichwohl können auch beim *TWT* die differentiellen Reaktionen tatbeteiligter Pbn bei den relevanten vs. irrelevanten Items durch die genannten Maßnahmen verstärkt werden (vgl. z.B. Steller, 1987; Steller & Dahle, 1997).

2.2.2.2 Theoretischer Hintergrund

Im Gegensatz zum *KFT* (vgl. Abschnitt 2.2.1) basiert der *TWT* auf klar definierten und überprüfbaren theoretischen Annahmen. Ebenfalls im Gegensatz zu den vagen Grundannahmen des *KFT* bezieht sich die theoretische Basis des *TWT* ausschließlich auf kognitive Prozesse des Pb und kommt ohne emotions- bzw. motivationsbezogene Konstrukte aus.

Die **„Theorie des Tatwissens"** (Lykken, 1974; vgl. auch Furedy & Ben-Shakhar, 1990, S. 107ff.), postuliert, daß die erhöhten physiologischen Reaktionen von Tätern bei den relevanten Items des *TWT* durch das bloße Vorhandensein tatbezogener Kenntnisse erklärt werden können. Als Bindeglied zwischen Tatkenntnissen und physiologischen Reaktionen dient das Konzept der Orientierungsreaktion (OR; z.B. Berlyne, 1960; Sokolov, 1963). Die OR wird konzeptualisiert als eine angeborene, zentralnervös vermittelte Antwort höherer Organismen auf neuartige Reize bzw. auf unerwartete Änderungen der Stimulation. Neben der Unterbrechung begonnener Handlungen, der Hin-

wendung zur Reizquelle und der Einleitung von Erkundungsverhalten manifestiert sich die OR in einer Vielzahl unwillkürlicher körperlicher Veränderungen (u.a. der elektrodermalen Aktivität), die dazu dienen, den Reizkontakt, die Sensibilität der Sinnesorgane, die kortikale Informationsverarbeitung und die Reaktionsbereitschaft zu optimieren. Bei wiederholtem konsequenzlosem Kontakt mit dem gleichen Reiz klingt die OR nach und nach ab (Habituation) und tritt erst wieder auf, sobald sich die Reizsituation ändert. Die OR ist dann stärker ausgeprägt und habituiert langsamer, wenn der auslösende Reiz biologischen oder aber erfahrungsbedingten psychologischen Signalwert besitzt. Dementsprechend wird die funktionale Bedeutung der OR u.a. in der unwillkürlichen Aufmerksamkeitsregulation gesehen bzw. darin, den Organismus vor der Mißachtung potentiell bedeutsamer Umweltereignisse zu schützen (s. Fröhlich, 2000). Jedes Item des *TWT* soll beim Pb, egal ob Täter oder Tatunbeteiligter, grundsätzlich eine OR auslösen können. Allerdings besitzen die relevanten Items für den Täter – bedingt durch die Erfahrung der Tatbegehung – einen höheren psychologischen Signalwert und lösen eine stärkere OR aus als die irrelevanten Items. (Man beachte, daß die normalerweise adaptive Funktion der OR in diesem Fall negative Konsequenzen für den Organismus bewirkt.) Dagegen sind für einen tatunbeteiligten Pb die relevanten und irrelevanten Items psychologisch gleich bedeutsam und evozieren dementsprechend indifferente ORn.

Die **„Dichotomisierungstheorie“** (z.B. Ben-Shakhar, 1977; Lieblich, Kugelmass & Ben-Shakhar, 1970) bezieht sich ebenfalls auf das Konzept der OR, betont diesbezüglich aber besonders den Aspekt der Habituation. Es wird angenommen, daß Personen mit Tatwissen die Testitems ausschließlich unter dem Gesichtspunkt der Tatbezogenheit verarbeiten und dementsprechend eine dichotome Kategorisierung in tatbezogene vs. nicht tatbezogene Reize vornehmen, während sie alle anderen Unterschiede zwischen den Testitems vernachlässigen. Weiterhin wird angenommen, daß alle Testitems ORn auslösen (die bei den [wiedererkannten] relevanten Items nicht notwendigerweise stärker sein müssen als bei den irrelevanten). Wie oben beschrieben, schwächen sich ORn nach und nach ab, wenn das auslösende Ereignis mehrmals hintereinander unverändert auftritt (Habituation). Mit jedem Auftreten des unveränderten Ereignisses schreitet die Habituation fort. Die ORn im *TWT* sollen – so eine weitere Annahme der Dichotomisierungstheorie – interkategoriell unabhängig sein, d.h. die Habituation generalisiere jeweils nur auf die Items ein und derselben Kategorie (z.B. Habituation über sämtliche relevanten Items), nicht jedoch auf die Items der anderen Kategorie. Da die Anzahl der Reizdarbietungen (Testitems) in der vom Tatwissenden gebildeten Kategorie „tatbezogene Reize“ kleiner ist als in der Kategorie „nicht tatbezogene Reize“ (Verhältnis 1:5; vgl. Tabelle 8) und weil die Habituation innerhalb einer Kategorie mit jeder neuen Darbietung eines der Kategorie angehörenden Reizes fortschreitet, ergibt sich, daß die

durchschnittliche OR eines Tatbeteiligten auf die relevanten Items stärker sein sollte als seine durchschnittliche OR auf die irrelevanten Items. Tatunbeteiligte können aufgrund fehlender Tatkenntnisse keine Kategorisierung in tatbezogene vs. nicht tatbezogene Testitems vornehmen. Dementsprechend können bei ihnen auch keine distinkten Habituationsprozesse bei den relevanten und irrelevanten Items ablaufen, so daß ihre durchschnittliche OR auf die relevanten Items genauso stark ausfallen sollte wie diejenige auf die irrelevanten Items.

Die „Theorie des Tatwissens“ steht im Einklang mit experimentellen Untersuchungen, in denen es gelang, verheimlichte Informationen aufzudecken, obwohl die Pbn weder sonderlich täuschungsmotiviert waren noch verbale Reaktionen (Lügen) erfolgten (z.B. Ben-Shakhar, 1977). Die „Dichotomisierungstheorie“ wird u.a. durch Studien untermauert, in denen demonstriert werden konnte, daß die Aufdeckung verheimlichter Informationen erschwert wird, wenn der relative Anteil der relevanten Items an der Gesamtzahl der Antwortalternativen erhöht wird (z.B. Lieblich et al., 1970). Gleichwohl deuten die vorliegenden Forschungsergebnisse auch darauf hin, daß neben dem Vorhandensein bzw. Fehlen tatspezifischer Kenntnisse sowie dem Ablaufen distinkter Habituationsprozesse bei relevanten und irrelevanten Items auch noch andere Faktoren einen zusätzlichen Beitrag zur differentiellen Reaktionsweise von Tätern vs. Tatunbeteiligten im *TWT* leisten können (s. dazu zusammenfassend Ben-Shakhar & Furedy, 1990, S. 108f.; Steller 1987).

2.2.2.3 Kritische Würdigung

Positiv hervorzuheben ist, daß der *TWT* die Möglichkeit der **Standardisierung und Objektivierung** bietet und insofern, verglichen mit dem *KFT*, eher einem wissenschaftlich fundierten Testverfahren entspricht. Während beim *KFT* die unterschiedliche Bedeutsamkeit bzw. Bedrohlichkeit von relevanten und Kontrollfragen erst durch deren individuelle Formulierung im Verlaufe eines suggestiv geführten Vortest-Interviews geschaffen werden muß (s. Abschnitt 2.2.1.1), ergeben sich die Fragen und Antwortalternativen beim *TWT* allein aus den Fakten des inkriminierten Delikts und sind somit weitgehend unabhängig von der Interaktion zwischen Untersucher und Pb. Sofern die relevanten und irrelevanten Alternativen für tatunbeteiligte Personen gleich plausibel erscheinen und sich auch ansonsten nicht systematisch unterscheiden (z.B. in ihrem emotionalen Gehalt), stellen die irrelevanten Items eine adäquate Kontrolle dar, so daß konsistent stärkere Reaktionen auf die relevanten Items mit hoher Wahrscheinlichkeit auf Tatwissen zurückzuführen sind. Wie auch durch die vorliegende Empirie bestätigt wird (s. Abschnitt 2.2.2.4), ist beim *TWT* das **Risiko, Personen ohne Tatwissen irrtümlich als schuldig zu klassifizieren, sehr ge-**

ring. Es ist außerdem möglich, die akzeptierte **Irrtumswahrscheinlichkeit** über die Anzahl der gestellten Fragen und Items zu **kontrollieren**. So beträgt die Wahrscheinlichkeit, daß eine Person ohne Tatwissen bei einer Frage mit fünf Alternativen zufällig am stärksten auf das relevante Item reagiert, 1:5. Bei einem *TWT*, der aus fünf solcher Multiple-Choice-Fragen besteht, sinkt das Risiko für konsistent stärkere Reaktionen bei den relevanten Alternativen, gemäß dem Multiplikationssatz der Wahrscheinlichkeitsrechnung, bereits auf 1:3125.

Die **Problematik** des *TWT* liegt v.a. in den eingeschränkten praktischen Einsatzmöglichkeiten (vgl. Steller, 1987; Steller & Dahle, 1997; Vrij, 1998b). Die Anwendung der Methode setzt voraus, daß außer dem Täter bzw. den Tätern niemand (abgesehen von den ermittelnden Personen bzw. dem Testleiter) nähere Kenntnisse von dem inkriminierten Delikt hat. Der *TWT* ist somit **nur sinnvoll einsetzbar, solange noch keine tatbezogenen Informationen an die Öffentlichkeit gedrungen sind**. Letzteres trifft in aller Regel nur auf die Frühphase von Ermittlungen zu. Ist die genannte Voraussetzung nicht erfüllt, so besteht eine erhöhte Gefahr, daß Unschuldige als Täter klassifiziert werden (Furedy & Heslegrave, 1991).

Jedoch selbst dann, wenn die Voraussetzung erfüllt ist, daß niemand außer dem Täter Kenntnisse von den Tatumständen haben kann, ist nicht sichergestellt, daß der Täter die im *TWT* als relevante Items verwendeten Details bei der Begehung der Tat auch wirklich wahrgenommen hat bzw. sich zum Zeitpunkt der Testung noch daran erinnern kann. Wurden also bestimmte kritische Details nie im Gedächtnis des Täters enkodiert oder hat dieser sie vor dem *TWT* schon wieder vergessen, so steigt damit die Gefahr, daß er durch das Testergebnis entlastet wird (falsch negative Entscheidung).

Darüber hinaus kann in bestimmten Fällen ein Tatverdächtiger die Methode außer Kraft setzen, indem er von vorneherein zugibt, die Umstände des fraglichen Tatbestands zu kennen, zugleich jedoch jegliche kriminelle Handlung seinerseits abstreitet. Hier ist insbesondere an Fälle mutmaßlicher sexueller Gewaltanwendung zu denken, in denen der Angeschuldigte auf die Freiwilligkeit der sexuellen Handlungen seitens des vermeintlichen Opfers pocht. Wenig sinnvoll ist der *TWT* auch in solchen Fällen, in denen aus mehreren bei der Tat anwesenden bzw. tatbeteiligten Personen der Täter bzw. Hauptschuldige ermittelt werden soll.

Wie oben dargelegt wurde, kommt bei einer sachgerechten Anwendung des *TWT* in erster Linie die **Gefahr falsch negativer Entscheidungen** in Betracht, welche sich aus wahrnehmungs- und gedächtnispsychologischen Überlegungen heraus ergibt. Wie im nächsten Abschnitt deutlich werden wird, verweisen auch die vorliegenden empirischen

Befunde auf eine diagnostische Urteilsverzerrung in Richtung falsch negativer Entscheidungen. Allerdings bietet sich für diese Fehlertendenz auch noch eine andere Erklärung an. So fallen die physiologischen (d.h. in erster Linie elektrodermalen) Reaktionen im *TWT* insgesamt eher schwach aus; und es bestehen zudem erhebliche interindividuelle Unterschiede in der elektrodermalen Reagibilität. Es ist denkbar, daß Täter, die generell schwache elektrodermale Reaktionen zeigen, nicht in ausreichendem Maße differentiell auf die relevanten vs. irrelevanten Items im *TWT* reagieren und daher mit einer erhöhten Wahrscheinlichkeit unentdeckt bleiben.

2.2.2.4. Treffsicherheit

Die Treffsicherheit des *TWT* wurde bislang überwiegend in experimentellen Studien mit dem Scheinverbrechen-Paradigma überprüft. Erst in jüngerer Zeit wurden auch erste Felduntersuchungen vorgenommen. In Tabelle 9 und Tabelle 10 sind die entsprechenden empirischen Befunde zusammengestellt. Feld- und Experimentaluntersuchungen zum *TWT* sind prinzipiell mit den gleichen Vor- und Nachteilen behaftet, die auch in bezug auf den *KFT* und die inhaltsorientierte Glaubhaftigkeitsbeurteilung gelten (Feldforschung: hohe externe Validität, aber Problematik der Validierungskriterien und Selektivität des Untersuchungsmaterials; Experiment: strenge Kontrolle des Validierungskriteriums, aber eingeschränkte externe Validität). Für eine ausführliche Erörterung dieser Problematik sei erneut auf Abschnitt 3.2.2 verwiesen.

Die **experimentellen Validitätsstudien** sind von der methodischen Vorgehensweise her nicht völlig homogen. So handelte es sich z.B. bei dem simulierten Delikt bei Davidson (1968) um einen Raubmord; bei Steller (1984) um einen Diebstahl und bei Balloun und Holmes (1979) um eine Täuschung in einem Intelligenztest. Lykken (1959) ließ einen Teil seiner Pbn gar zwei Delikte simulieren (Mord und Diebstahl). In der Studie von Balloun und Holmes (1979) konnten die Pbn selber entscheiden, ob sie die Tat begingen oder nicht, während in allen anderen Untersuchungen die Zuteilung zu den Versuchsbedingungen per Randomisierung erfolgte. Des weiteren unterscheiden sich die Experimentalstudien in bezug auf das verwendete Testformat. So bestand z.B. der *TWT* bei Balloun und Holmes (1979) aus fünf Fragen mit je fünf Antwortalternativen. Dagegen beinhaltete etwa der *TWT* bei Steller (1984) sechs Fragen mit jeweils sechs Items. Auch die Art der geforderten Reaktionen auf die Antwortalternativen differierte zwischen den Studien. So forderten z.B. Davidson (1968) und Lykken (1959) von ihren Pbn gar keine Reaktionen, während etwa bei Podlesny und Raskin (1978) mit nein zu antworten war. Balloun und Holmes (1979) ließen ihre Pbn die Testitems wiederholen. Auch die Art, wie man die Pbn dazu motivierte, im *TWT* möglichst als unschuldig klas-

sifiziert zu werden, war nicht in allen Studien einheitlich. Während etwa Davidson (1968) den Pbn für den Fall eines negativen Testergebnisses eine Geldprämie in Aussicht stellte, applizierte Lykken (1959) während der Befragung in unsystematischer Weise elektrische Schläge. Die Pbn hatten vorher die fingierte Information erhalten, daß diese elektrischen Schläge nur auftreten würden, wenn die registrierten physiologischen Reaktionen auf Täterschaft hindeuteten. Zu erwähnen ist auch, daß in einigen Studien neben den elektrodermalen Veränderungen auch noch andere physiologische Maße als abhängige Variablen mit erfaßt wurden (z.B. Herzrate und Pupillenweite bei Bradley & Janisse, 1981; Atmung, Herzrate und periphere Durchblutung bei Steller, 1984). Die hier berichteten Treffer- bzw. Fehlerquoten (Tabelle 9) beruhen jedoch ausschließlich auf der Auswertung der elektrodermalen Reaktionen.

Tabelle 9. Mit dem *TWT* erzielte Treffer- und Fehlerquoten (in %) in experimentellen Studien

	Schuldige			Unschuldige	
Studie	N	richtig	falsch	N	richtig	falsch
Balloun & Holmes (1979)	18	61.1	38.9	16	87.5	12.5
Bradley & Janisse (1981)	96	59.4	40.6	96	88.5	11.5
Davidson (1968)	12	91.7	8.3	36	100	0
Giesen & Rollison (1980)	20	95.0	5.0	20	100	0
Lykken (1959)	37	88.0	12.0	12	100	0
Podlesny & Raskin (1978)	10	80.0	20.0	10	100	0
Steller (1984)	47	85.1	14.9	40	100	0
Durchschnitt (ungewichtet)		80.0	20.0		96.6	3.4

Wie Tabelle 9 zu entnehmen ist, betrugen die durchschnittlichen Trefferquoten in den Simulationsstudien 80% bei den schuldigen Pbn und 96.6% bei den unschuldigen. Somit wurden im Mittel deutlich mehr schuldige als unschuldige Pbn fehlklassifiziert. Diese Fehlertendenz „zuungunsten“ der schuldigen Pbn zieht sich konsistent durch alle Untersuchungen. Bei den schuldigen Pbn ist auch die Streuung der erzielten Trefferraten größer. Diese schwanken zwischen 59.4% (Bradley & Janisse, 1981) und 95.0% (Giesen & Rollison, 1980). Dagegen beläuft sich die niedrigste Trefferquote bei den

unschuldigen Pbn auf 87.5% (Balloun & Holmes, 1979). Bemerkenswert ist, daß in fünf der sieben angeführen Studien keine falsch positiven Entscheidungen auftraten (s. Tabelle 9).

Zur Verifizierung der in die beiden **Feldstudien** (s. Tabelle 10) aufgenommenen Fälle wurden jeweils Geständnisse von Beschuldigten herangezogen. Die in den forensischen Realfällen durchgeführten *TWTs* wiesen (auch innerhalb der Studien) kein einheitliches Format auf. Die Tests beinhalteten zwischen einer und sechs Fragen. Die jeweilige Anzahl der Antwortalternativen variierte von drei bis acht. Mitunter wurden die Mehrfachwahlfragen in bis zu vier Befragungsdurchgängen wiederholt dargeboten. In den Realfällen wurde eine modifizierte Form der numerischen Auswertung vorgenommen; entsprach die erreichte Punktzahl der Anzahl der gestellten Fragen, so galt das Ergebnis als „unentscheidbar". Wenngleich die Auswertung blind erfolgte, muß doch einschränkend angemerkt werden, daß die Testleiter jeweils die relevanten Antwortalternativen kannten, was möglicherweise zu Versuchsleitererwartungseffekten (insbesondere in Form falsch positiver Entscheidungen) geführt haben könnte. Erwähnt sei auch, daß in den *TWT*s neben den elektrodermalen Reaktionen auch noch andere physiologische Maße mit erhoben wurden. Die in Tabelle 10 angeführten Resultate basieren jedoch ausschließlich auf der Auswertung der elektrodermalen Reaktionen.

Tabelle 10. Mit dem *TWT* erzielte Treffer- und Fehlerquoten (in %) in Feldstudien

	Schuldige				Unschuldige			
Studie	N	richtig	falsch	unentsch.	N	richtig	falsch	unentsch.
Elaad (1990)	48	41.7 (50)	41.7 (50)	16.7	50	92 (97.9)	2 (2.1)	6
Elaad et al. (1992)	40	40 (53.3)	35 (46.7)	25	40	92.5 (97.4)	2.5 (2.6)	5
Durchschnitt (ungewichtet)		40.9 (51.7)	38.4 (48.4)	20.9		92.3 (97.7)	2.3 (2.4)	5.5

Anmerkung: Werte in Klammern ergeben sich bei Nichtberücksichtigung der unentscheidbaren Fälle.

Die Ergebnisse der beiden Feldstudien sind nahezu gleich. Im Durchnitt beider Studien betrugen die Trefferquoten, ohne Berücksichtigung der unentscheidbaren Fälle, 51.7% bei den schuldigen Pbn und 97.7% bei den unschuldigen Pbn. Somit zeigte sich auch in den Feldstudien eine deutliche Fehlertendenz in Richtung falsch negativer Entscheidungen. Die schuldigen Pbn wurden in etwa genauso oft falsch wie richtig klassifiziert. Die Treffsicherheit in bezug auf schuldige bewegt sich also auf dem Zufallsniveau.

Faßt man die Resultate aus den Experimental- und Feldstudien zusammen, so ist insbesondere die **hohe Treffsicherheit bei der Identifizierung tatunbeteiligter** Pbn hervorzuheben. Dagegen bleibt jedoch eine hohe Rate von Tätern unerkannt. Ist die Rate falsch negativer Urteile in den Laboruntersuchungen mit durchschnittlich 20% noch relativ niedrig, so wurden jedoch in den untersuchten Realfällen die schuldigen Pbn ebenso oft falsch wie richtig klassifiziert. Über die Gründe für die diesbezügliche Diskrepanz zwischen den Feld- und Laboruntersuchungen kann nur spekuliert werden. Die nächstliegende Erklärung ist, daß die schuldigen Pbn in den analysierten Realfällen häufig relevante Items nicht wiedererkannten, da sie sie bei der Tatbegehung nicht wahrgenommen oder zum Zeitpunkt des *TWT* bereits wieder vergessen hatten. Dagegen sind experimentelle Studien in der Regel so angelegt, daß die schuldigen Pbn die kritischen Antwortalternativen im *TWT* mit hoher Wahrscheinlichkeit wiedererkennen.

2.2.3 Der *Guilty Actions Test* (*GAT*)

Beim *GAT* handelt es sich um eine modifizierte Version des *TWT*, die von der kanadischen Forschergruppe um **Michael T. Bradley** entwickelt wurde (Bradley, MacLaren & Carle, 1996; Bradley & Rettinger, 1992; Bradley & Warfield, 1984).

2.2.3.1 Hintergrund, diagnostische Vorgehensweise und Grundannahme

Den Hintergrund für die Konzeption des *GAT* bildete einerseits **die beim *TWT* bemängelte Einschränkung der praktischen Einsatzmöglichkeiten**. Wie in Abschnitt 2.2.2.3 dargestellt wurde, ist eine sinnvolle Anwendung des *TWT* an die Voraussetzung geknüpft, daß kein Unschuldiger tatbezogene Kenntnisse haben darf, was in aller Regel nur auf die Frühphase von Ermittlungen zutrifft. Sind dagegen bereits tatbezogene Informationen an die Öffentlichkeit und somit zu unschuldigen Verdächtigen gedrungen, so wäre bei Anwendung des *TWT* von einem erhöhten Risiko falsch positiver Entscheidungen auszugehen (Unschuldige werden als Tatbeteiligte klassifiziert). Neben diesem anwendungsbezogenen Defizit sahen sich Bradley und Kollegen auch aus theoretischen Erwägungen zur Modifizierung des herkömmlichen *TWT* veranlaßt. So kritisieren sie, daß differentielle physiologische Reaktionen auf die relevanten vs. irrelevanten Items des *TWT* grundsätzlich nicht eindeutig erklärbar seien, da das **Tatwissen mit der Glaubhaftigkeit der Antworten konfundiert** sei. Anders ausgedrückt: Es ist unklar, inwiefern stärkere physiologische Reaktionen bei den relevanten Items des *TWT* auf das Wiedererkennen der Tatdetails oder die bewußt wahrheitswidrige Verneinung der Items (Lügen) zurückzuführen sind.

Mit dem *GAT* sollte nun ein Instrument geschaffen werden, welches die genannten praxis- und theoriebezogenen Schwächen des *TWT* nicht aufweist. Das modifizierte Testformat sollte also zum einen geeignet sein, die im *TWT* gegebene **Konfundierung von Tatwissen und Täuschung aufzulösen**, um dadurch ein besseres Verständnis der psychologischen Prozesse zu erlangen, die den physiologischen Reaktionen auf die Testitems zugrunde liegen (theoretischer Aspekt). Zum anderen sollte es die Möglichkeit bieten, **nicht nur zwischen Personen mit und ohne Tatwissen zu differenzieren, sondern zusätzlich auch noch zwischen Schuldigen (Tätern) und Unschuldigen mit Tatwissen**, um so das Einsatzgebiet auch auf solche Fälle auszudehnen, in denen unschuldige Verdächtige möglicherweise Tatkenntnisse erlangt haben (praktischer Aspekt).

Um diese Ziele zu erreichen, wurde das **Frageformat des *TWT* abgeändert**. Während die Fragen des herkömmlichen *TWT* ausschließlich auf die Kenntnis von Tatdetails abzielen (z.B. in einem Mordfall: „Womit wurde Herr X getötet? a) Messer, b) Hammer ..."), beziehen sich die Fragen des *GAT* sowohl auf das Tatwissen des Pb als auch auf die Täterschaft selbst („Womit *haben Sie* Herrn X getötet? a) Messer, b) Hammer ..."). Insofern handelt es sich beim *GAT* gewissermaßen um eine Mischform aus direkten und indirekten Methoden der psychophysiologischen Glaubhaftigkeitsbeurteilung. Alle Fragen bzw. Alternativen sind mit nein zu beantworten. Abgesehen von der Frageformulierung erfolgt die Durchführung des *GAT* nach den gleichen Prinzipien wie die des konventionellen *TWT*.

Die **Grundannahme des *GAT*** ist folgende: Schuldige Pbn (Täter) erkennen die zutreffenden Alternativen wieder, und sie lügen, indem sie diese wahrheitswidrig mit nein beantworten. Unschuldige Pbn mit Tatwissen erkennen die relevanten Items zwar ebenfalls wieder, im Gegensatz zu den schuldigen Pbn antworten sie jedoch wahrheitsgemäß, indem sie die relevanten Items verneinen. Unter der Prämisse, daß das wahrheitswidrige Beantworten der Items (Lügen) einen zusätzlichen Effekt auf die Stärke der physiologischen Reaktionen hat[8], sollten Unschuldige mit Tatwissen auf die relevanten Items schwächer reagieren als Schuldige. Somit sollte auch die Reaktionsstärkedifferenz zwischen relevanten und irrelevanten Items (letztere werden von allen Pbn weder wiedererkannt noch wahrheitswidrig verneint) bei den Unschuldigen mit Tatwissen kleiner sein als bei den Schuldigen. Unschuldige Pbn ohne Tatwissen sollten im *GAT* das gleiche Reaktionsmuster zeigen wie im klassischen *TWT*. Da sie weder die relevan-

[8] Ein spezifischer Effekt wahrheitswidriger Antworten konnte im Rahmen des sog. „Differentiation of Deception"-Paradigmas nachgewiesen werden. Bei wahrheitswidriger Beantwortung einfacher (auto-) biographischer oder Wissensfragen zeigen Pbn u.a. stärkere Hautleitfähigkeitsreaktionen als wenn sie wahrheitsgemäß antworten (z.B. Furedy, Davis & Gurevich, 1988; Gödert, Rill & Vossel, 2001).

ten Items identifizieren können noch bei deren Verneinung lügen, ist hier nicht zu erwarten, daß sich die physiologischen Reaktionsstärken bei den relevanten und irrelevanten Alternativen systematisch unterscheiden. Gemäß dieser Grundannahme sind die drei Gruppen von Verdächtigen (Schuldige, Unschuldige mit Tatwissen, Unschuldige ohne Tatwissen) prinzipiell anhand ihrer physiologischen Reaktionsmuster im *GAT* unterscheidbar.

Das physiologische Reaktionsmuster, welches Unschuldige mit Tatwissen im *GAT* zeigen, sollte sich folgendermaßen von ihrem Reaktionsmuster im *TWT* abheben. Während sie im *TWT* sowohl die relevanten Items wiedererkennen als auch lügen, indem sie diese verneinen, erkennen sie die relevanten Items des *GAT* lediglich wieder ohne jedoch zu lügen. Dementsprechend sollten sie – verglichen mit dem *TWT* – im *GAT* schwächere physiologische Reaktionen auf die relevanten Items zeigen. Folglich sollten hier auch ihre Reaktionsstärkeunterschiede in bezug auf die relevanten vs. irrelevanten Items kleiner ausfallen als im *TWT*.

Überträgt man die numerische Auswertungsmethode des *TWT* (vgl. Abschnitt 2.2.2.1) auf den *GAT*, so könnte man erwarten, daß Unschuldige mit Tatwissen im *GAT* alles in allem niedrigere numerische Scores erzielen als im *TWT*, da ihre – im Vergleich zum *TWT* – abgeschwächten physiologischen Reaktionen auf die relevanten Items seltener die Reaktionen auf die irrelevanten Items übertreffen. Somit sollte auch das Risiko falsch positiver Klassifikationen von Unschuldigen mit Tatwissen im *GAT* geringer sein als im *TWT*. Die Übertragung der numerischen Auswertungsmethode des *TWT* auf den *GAT* ist allerdings nicht unproblematisch, da diese Scoring-Methode der Logik des *GAT* prinzipiell nicht voll gerecht werden kann. Hierauf wird in Abschnitt 2.2.3.3 näher eingegangen. Zunächst sollen jedoch die empirische Forschung zum *GAT* und deren Resultate beschrieben werden.

2.2.3.2 Empirische Befunde

Der Arbeitskreis um Bradley hat bislang **drei Laborstudien** zum *GAT* vorgelegt (Bradley et al., 1996; Bradley & Rettinger, 1992; Bradley & Warfield, 1984). Das experimentelle Prozedere folgte jeweils dem gleichen Prinzip: Jeweils eine Gruppe von Pbn (Schuldige) mußte ein Scheinverbrechen begehen. Dabei handelte es sich um einen inszenierten Mord. Das Scheinverbrechen beinhaltete jeweils zehn kritische Details, die man später im *GAT* als relevante Items verwendete (z.B. Name des Opfers, Anzahl der abgegebenen Schüsse). Andere Pbn begingen zwar das Scheinverbrechen nicht, wurden aber in irgendeiner Form von den 10 kritischen Tatdetails in Kenntnis gesetzt (Unschul-

dige mit Tatwissen), so etwa, indem sie die Tat als Zeugen beobachteten (Bradley et al., 1996; Bradley & Warfield, 1984) oder indem sie in schriftlicher Form explizit über die Tatdetails informiert wurden (Bradley & Rettinger, 1992; Bradley & Warfield, 1984).[9] In den Untersuchungen von Bradley und Warfield (1984) sowie Bradley und Rettinger (1992) gab es neben den Gruppen der Schuldigen und Unschuldigen mit Tatwissen jeweils auch noch eine Kontrollgruppe, deren Pbn sich einem *GAT* unterzogen, ohne vorher einen Scheinmord begangen oder Kenntnis der Tatdetails erlangt zu haben (Unschuldige ohne Tatwissen).

Allen Pbn wurde eine finanzielle Belohnung in Aussicht gestellt, falls sie im *GAT* als unschuldig eingestuft würden. Der *GAT* bestand in allen drei Studien aus zehn Fragen mit je fünf Alternativen. Als physiologisches Maß wurde die Amplitude der Hautwiderstandsreaktionen erfaßt (bei Bradley & Rettinger, 1992, außerdem thorakale und abdominale Atembewegungen, auf deren Resultate hier aber nicht eingegangen wird). Die Auswertung erfolgte mit der für den herkömmlichen *TWT* entwickelten numerischen Methode (vgl. Abschnitt 2.2.2.1). Dementsprechend konnte ein Pb über alle zehn Fragen maximal einen numerischen Score von 20 erzielen. Die numerischen Scores der Pbn wurden zum einen im Hinblick auf etwaige Unterschiede zwischen den verschiedenen Versuchsbedingungen statistisch analysiert. Zum anderen wurden die Pbn anhand ihrer numerischen Scores als schuldig vs. unschuldig klassifiziert, wobei der beim konventionellen *TWT* übliche Cutoff-Wert von 10 angelegt wurde.

Wichtig ist, daß in allen drei Untersuchungen im Anschluß an den *GAT* das freie Erinnern und Wiedererkennen der kritischen Tatdetails getestet wurde. Dadurch wurde kontrolliert, ob sich Schuldige und Unschuldige mit Tatwissen in ihrer Gedächtnisleistung für die Tatdetails unterschieden. Dies war jedoch nicht der Fall, so daß eine etwaige differentielle Gedächtnisleistung als konfundierende Variable ausgeschlossen werden konnte.

---

[9] Während durch diese expliziten Informationen die Situation simuliert werden sollte, daß ein Unschuldiger über die Medien oder im Rahmen polizeilicher Verhöre Tatkenntnisse erlangt hat und sich somit – genauso wie ein Zeuge – des Tatbezugs seines Wissens voll bewußt ist, gab es bei Bradley und Warfield (1984) auch noch eine Gruppe von Unschuldigen mit Tatwissen, die sich des Tatbezugs ihrer Kenntnisse nicht bewußt waren („Innocent Associations Group", S. 685). Diese Bedingung wurde realisiert, indem man die Pbn instruierte, den Raum aufzuräumen, in dem auch (zeitversetzt) der simulierte Mord begangen wurde. In die Aufräumhandlung waren dieselben Details involviert wie bei dem simulierten Mord. So sollten die Pbn u.a. *blaue Umschläge* aus der *oberen Schublade des Schreibtischs* herausnehmen und in den *Papierkorb* werfen. Die Details „blauer Umschlag", „obere Schreibtischschublade" und „Papierkorb" dienten jedoch auch als relevante Items des *GAT*, da bei dem simulierten Delikt u.a. die „Mordwaffe" aus der *oberen Schreibtischschublade* genommen und später in den *Papierkorb* geworfen sowie dem Opfer ein *blauer Umschlag* entwendet wurde. Die Ergebnisse der „Innocent Associations Group" sind jedoch von solch geringer Praxisrelevanz, daß sie in der vorliegenden Darstellung nicht berücksichtigt werden.

Als Ergebnis der Untersuchung von Bradley und Warfield (1984) zeigte sich, daß die schuldigen Pbn im Durchschnitt höhere numerische Scores erzielten als alle anderen Pbn, also auch höhere als die unschuldigen Pbn mit Tatwissen. Dies spricht grundsätzlich für die Differenzierungsfähigkeit des *GAT* im Hinblick auf Schuldige und Unschuldige mit Tatwissen. Alle schuldigen Pbn und alle unschuldigen Pbn ohne Tatwissen wurden korrekt klassifiziert. Die unschuldigen Pbn mit Tatwissen wurden jedoch zu 25% als schuldig fehlklassifiziert. Bei Bradley und Rettinger (1992) waren die numerischen Scores der schuldigen Pbn signifikant höher als die der unschuldigen Pbn mit Tatwissen. Letztere wiederum erzielten signifikant höhere Scores als die unschuldigen Pbn ohne Tatwissen. Somit konnte auch diese Studie die Grundannahme des *GAT* prinzipiell untermauern. Auch hier wurden alle schuldigen Pbn sowie sämtliche unschuldigen Pbn ohne Tatwissen korrekt klassifiziert. Die Fehlerquote bei den unschuldigen Pbn mit Tatwissen betrug allerdings 50%. Zusammengefaßt stützen die Resultate dieser beiden Untersuchungen die Grundannahme, daß Schuldige und Unschuldige mit Tatwissen (sowie Unschuldige ohne Tatwissen) auf die relevanten *GAT*-Items unterschiedlich stark reagieren. Tatwissen erwies sich als notwendige, jedoch nicht als hinreichende Bedingung für die Klassifikation als schuldig. So wurden einerseits Personen ohne Tatwissen nie als schuldig eingestuft; andererseits wurden jedoch nicht alle (unschuldigen) Personen mit Tatwissen als schuldig beurteilt.

Die Studie von Bradley et al. (1996) ist besonders interessant, da hier der *GAT* direkt mit dem konventionellen *TWT* verglichen wurde. Es handelte sich um ein dreifaktorielles Versuchsdesign mit den Gruppenfaktoren Status des Pb (Schuldige vs. Unschuldige mit Tatwissen), Testart (*TWT* vs. *GAT*) und Antwortmodus (Verneinung vs. Itemwiederholung vs. Schweigen). Somit ergaben sich zwölf Versuchsgruppen, für die, basierend auf der Grundannahme des *GAT*, spezifische Vorhersagen getroffen werden konnten.

Man setzte voraus, daß alle Pbn im *TWT* bzw. im *GAT* die kritischen Items als solche identifizieren würden. Diese Voraussetzung konnte aufgrund der Resultate der Gedächtnistests als erfüllt betrachtet werden. Somit implizierte das experimentelle Design, daß nur drei der zwölf Versuchsgruppen auf die relevanten Items mit einer Lüge (fälschliche Verneinung) reagierten. Bei diesen drei Gruppen handelte es sich um (1) schuldige Pbn, die im *TWT* mit nein antworteten, (2) schuldige Pbn, die im *GAT* mit nein antworteten, sowie (3) unschuldige Pbn mit Tatwissen, die im *TWT* mit nein antworteten. Für diese drei Gruppen wurde erwartet, daß die Kombination aus Identifizierung *und* wahrheitswidriger Verneinung der kritischen Items zu stärkeren physiologischen Reaktionen auf die kritischen Items und somit letztlich zu höheren numerischen Scores führen würde, verglichen mit den übrigen neun experimentellen Gruppen, die die

relevanten Alternativen zwar ebenfalls wiedererkannten, aber darauf nicht mit Lügen reagierten.

Anstelle der erwarteten dreifachen Interaktion (Status des Pb-Testart-Antwortmodus) zeigte sich lediglich eine zweifache Wechselwirkung zwischen den Faktoren Status des Pb und Testart. Die Unschuldigen mit Tatwissen, die mit dem *GAT* getestet wurden, erzielten signifikant niedrigere numerische Scores als alle anderen Pbn, d.h. niedrigere Scores als die Schuldigen im *GAT*, die Schuldigen im *TWT* und die Unschuldigen mit Tatwissen im *TWT*. Der Antwortmodus übte lediglich einen Haupteffekt aus, der darauf beruhte, daß die mit nein antwortenden Pbn höhere numerische Scores erzielten als die schweigenden Pbn.

Post-hoc-Analysen zeigten jedoch, daß nur die Gruppe der Schuldigen, die mit dem *GAT* getestet wurden, nicht der experimentellen Hypothese entsprach, d.h. diese Pbn erzielten keine höheren numerischen Scores, wenn mit nein geantwortet werden sollte als wenn Itemwiederholung bzw. gar keine Antwort gefordert war. Alle anderen Gruppen zeigten im Durchschnitt die vorhergesagten Reaktionsmuster. So erzielten die Unschuldigen mit Tatwissen im *GAT* – unabhängig vom Antwortmodus – relativ niedrige Scores, was zu erwarten war, da sie ja bei keinem Antwortmodus gezwungen waren zu lügen. Sowohl die Schuldigen als auch die Unschuldigen mit Tatwissen erzielten im *TWT* höhere Scores, wenn sie mit nein antworteten, also bei den kritischen Items lügen mußten, als wenn sie die Alternativen wiederholten oder schwiegen. So gesehen sprechen die Ergebnisse insgesamt (mit Ausnahme der Gruppe Schuldige-*GAT*) dafür, daß Pbn mit Tatwissen, egal ob schuldig oder unschuldig, stärkere physiologische Reaktionen zeigen, wenn sie lügen.

Differenziert man nicht nach den drei Antwortmodi, so wurden anhand der numerischen Scores im *GAT* 21 von 30 Schuldigen (70%) und im *TWT* 20 von 30 Schuldigen (67%) korrekt klassifiziert, d.h. in bezug auf die schuldigen Pbn erwiesen sich *GAT* und *TWT* in etwa als gleich treffsicher. In bezug auf die Unschuldigen mit Tatwissen lag die Trefferquote im *GAT* bei 63% (19 von 30 Pbn), im *TWT* hingegen nur bei 37% (11 von 30). Diese Trefferquoten sind jedoch wenig aussagekräftig, da die Grundannahme des *GAT* an die Bedingung geknüpft ist, daß die Testitems verneint werden.

Betrachtet man nur die Gruppen, in denen mit nein geantwortet wurde, so lagen die Trefferquoten in bezug auf die Schuldigen im *GAT* und *TWT* bei 90% bzw. 80% (9 bzw. 8 von 10 Pbn). Die entsprechenden Trefferquoten in bezug auf die Unschuldigen mit Tatwissen beliefen sich auf 50% im *GAT* sowie auf 10% im *TWT*. Die Treffsicherheit des *GAT* bei Unschuldigen mit Tatwissen bewegte sich also auf dem Zufallsniveau und

war somit, bei isolierter Betrachtung, unbefriedigend. Allerdings erwies sich der *GAT* insofern als Verbesserung, als die Rate falsch positiver Entscheidungen gegenüber dem *TWT* wesentlich reduziert war.

Faßt man die Resultate der Studien von Bradley und Kollegen zusammen, so läßt sich folgendes **Resümee** ziehen: Die Grundannahme des *GAT* wird durch die vorliegenden Befunde prinzipiell bestätigt. Unschuldige Pbn, die Tatwissen besitzen, unterscheiden sich hinsichtlich ihrer physiologischen Reaktionsstärken im *GAT* von schuldigen Pbn; dies wird durch den Unterschied in der Höhe der numerischen Scores indiziert. Die Befunde deuten darauf hin, daß Tatwissen eine notwendige, aber keine hinreichende Bedingung ist, um im *GAT* als schuldig klassifiziert zu werden. So wurden Personen ohne Tatwissen nie als schuldig klassifiziert; aber nicht alle Personen mit Tatwissen wurden als schuldig eingestuft. Es bleibt allerdings offen, inwiefern die höheren numerischen Scores der Schuldigen mit der wahrheitswidrigen Verneinung der relevanten Items zusammenhängen und insofern einen spezifischen Lügeneffekt wiederspiegeln. So sprechen die Ergebnisse von Bradley et al. (1996) zwar dafür, daß im *TWT* Schuldige ebenso wie Unschuldige mit Tatwissen stärker auf die relevanten Items reagieren, wenn sie diese wahrheitswidrig verneinen. Allerdings konnte dieser Effekt der wahrheitswidrigen Verneinung relevanter Items bei den Schuldigen im *GAT* nicht beobachtet werden.

Im Hinblick auf die praktische Anwendung bietet der *GAT* im Vergleich zum *TWT* den Vorteil, daß das Risiko falsch positiver Entscheidungen bei Unschuldigen mit Tatwissen reduziert ist. Gleichwohl übersteigt die Treffsicherheit des *GAT* bei dieser Probandengruppe alles in allem kaum das Zufallsniveau. Es ist allerdings nicht auszuschließen, daß diese Fehleranfälligkeit auch mit der numerischen Auswertungsmethode zusammenhängt, welche eigentlich speziell für den *TWT* konzipiert wurde bzw. der Logik des *GAT* nicht voll gerecht wird. Dieser Punkt wird im nächsten Abschnitt genauer erläutert.

2.2.3.3 Kritik an der bisherigen Forschung

Bradley und Kollegen verwendeten in ihren Untersuchungen zum *GAT* ausschließlich die für den *TWT* konzipierte numerische Auswertungsmethode. Bei genauerer Betrachtung wird jedoch deutlich, daß **das numerische Auswertungsverfahren des *TWT* der Logik des *GAT* gar nicht gerecht werden kann** und somit auch nicht optimal geeignet ist, die postulierten diagnostischen Vorzüge des *GAT*, sprich die Differenzierungsfähigkeit im Hinblick auf Schuldige vs. Unschuldige mit Tatwissen, voll auszuschöpfen. Wie

in Abschnitt 2.2.3.1 erläutert wurde, beruht diese Differenzierungsfähigkeit darauf, daß der Unterschied zwischen den Reaktionsstärken auf die relevanten vs. irrelevanten *GAT*-Items bei Unschuldigen mit Tatwissen kleiner ist als bei Schuldigen. Oder anders ausgedrückt: Die Differenzierungsfähigkeit des *GAT* bezüglich Schuldiger und Unschuldiger mit Tatwissen kommt dadurch zustande, daß Unschuldige mit Tatwissen im *GAT* geringere Reaktionsstärkeunterschiede zwischen relevanten vs. irrelevanten Items zeigen als im *TWT*, was für die Schuldigen nicht gilt. Dieses Weniger an Unterschied kann jedoch prinzipiell durch das numerische Auswertungsverfahren überhaupt nicht erfaßt werden. Mit der numerischen Auswertung wird lediglich quantifiziert, welche Rangposition die Reaktionsstärke beim relevanten Item innerhalb der Reaktionsstärken in der gesamten Itemsequenz einnimmt (vgl. Abschnitt 2.2.2.1). Entscheidend ist, ob die Reaktionsstärke beim relevanten Item die größte (2 Punkte) bzw. zweitgrößte (1 Punkt) innerhalb der Itemsequenz ist. Dabei spielt es keine Rolle, ob der Unterschied zur nächstschwächeren Reaktion beträchtlich oder minimal ist. Daraus ergibt sich streng genommen, daß Unschuldige mit Tatwissen im *GAT* die gleichen numerischen Scores erzielen sollten wie Schuldige, da beide Probandengruppen bei den relevanten Items stärker reagieren als bei den irrelevanten. Eine Senkung der numerischen Scores unschuldiger Pbn mit Tatwissen bei Verwendung des *GAT* kann man nur unter der Prämisse erwarten, daß die zufällig variierenden Reaktionsstärken bei den irrelevanten Items die Reaktionen auf die relevanten Items häufiger übertreffen als im *TWT*, da im *GAT* der Abstand zur Reaktionsstärke bei den relevanten Items im Durchschnitt nicht so groß ist. Somit würde jedoch die Senkung der numerischen Scores und damit die Identifizierung Unschuldiger mit Tatwissen letztlich vom Zufall abhängen.

Eine logische Konsequenz hieraus ist auch, daß die **Grundannahme des *GAT* gar nicht in adäquater Weise mit Hilfe der numerischen Auswertungsmethode überprüft werden kann**, da diese die Größe des Reaktionsstärkeunterschieds zwischen relevanten und irrelevanten Items nicht präzise abzubilden vermag und somit auch nicht für den Nachweis geeignet ist, daß dieser Reaktionsstärkeunterschied bei Unschuldigen mit Tatwissen schwächer ausgeprägt ist als bei Schuldigen. Insofern ist es etwas verwunderlich, daß in den Arbeiten der Bradley-Gruppe dennoch ausschließlich die numerische Auswertungsmethode verwendet wurde und die Autoren noch nicht einmal auf deren Problematik hinweisen. Ein adäquateres Vorgehen wäre, die physiologischen Reaktionsstärken bei sämtlichen Items direkt zu quantifizieren und dann statistisch zu analysieren, wobei gemäß der Grundannahme des *GAT* eine Interaktion zwischen Probandengruppe (Schuldige vs. Unschuldige mit Tatwissen vs. Unschuldige ohne Tatwissen) und Itemtyp (relevant vs. irrelevant) zu erwarten wäre. Allerdings spricht es um so mehr für die *GAT*-Grundannahme, daß diese mit Hilfe der numerischen Auswertungs-

methode trotz deren angesprochener Unzulänglichkeiten im wesentlichen bestätigt werden konnte.

Daß die numerische Auswertungsmethode somit auch im Hinblick auf eine individualdiagnostische Anwendung des *GAT* unzulänglich ist, versteht sich von selbst. Eine angemessene Vorgehensweise müßte z.B. dem Diagnostiker klare Entscheidungsrichtlinien vorgeben, ab wann ein Testergebnis nicht mehr nur auf vorhandenes Tatwissen, sondern zusätzlich auch noch auf Täterschaft hindeutet, d.h. es müßte ein entsprechender Cutoff-Wert definiert werden.

Schließlich ist auch noch zu betonen, daß die **bisherigen Erkenntnisse ausschließlich im Labor gewonnen** wurden. Die Frage, ob die im *GAT* vorgefundenen physiologischen Reaktionsmuster von Schuldigen und Unschuldigen mit Tatwissen sich auch in Realsituationen unterscheiden, bleibt somit bis auf weiteres unbeantwortet.

2.2.4 Anmerkungen zur Anwendbarkeit der psychophysiologischen Methoden

Zwei wichtige Einschränkungen sämtlicher psychophysiologischer Methoden der Glaubhaftigkeitsbeurteilung wurden bisher noch nicht erwähnt. Zum einen sind die verschiedenen diagnostischen Verfahren nur sinnvoll anwendbar, wenn die Pbn **freiwillig** teilnehmen bzw. bereit sind, den Instruktionen des Untersuchungsleiters Folge zu leisten. Bei Verweigerung oder unsystematischen Störmanövern seitens der Pbn (z.B. Nichtbeachten der Fragen oder Herumzappeln) werden die Ergebnisse unbrauchbar (Steller, 1987). Zum anderen liegt ein noch gravierenderes Problem darin, daß die **Messungen auch systematisch manipulierbar** sind. Sofern einem schuldigen Pb die Logik des jeweiligen Verfahrens bekannt ist, kann er durch für den Untersucher unsichtbare mentale (z.B. Kopfrechnen) und insbesondere motorische Aktivitäten (z.B. Fuß gegen den Boden pressen) die physiologische Aktivierung bei den Kontrollfragen (*KFT*) bzw. irrelevanten Items (*TWT, GAT*) gezielt erhöhen und somit einen unentscheidbaren oder gar falsch negativen Untersuchungsbefund herbeiführen (Lykken, 1998; Vrij, 1998b).

3 Problemstellung

3.1 Ableitung der Fragestellung

Die Fragestellung der vorliegenden Untersuchung leitet sich ab aus der rechtlichen Situation der forensischen Glaubhaftigkeitsbeurteilung in Deutschland, welche sich für die inhaltsorientierten und psychophysiologischen Methoden grundlegend unterscheidet.

Die **juristische Situation der psychophysiologischen Glaubhaftigkeitsbeurteilung hierzulande** wurde für lange Zeit durch ein **Grundsatzurteil des Bundesgerichtshofs (BGH) vom 16.02.1954** geregelt (Aktenzeichen: 1 StR 578/53; vgl. BGH, 1954, S. 332ff.). Diesem Urteil zufolge verletze die Durchführung einer psychophysiologischen Glaubhaftigkeitsbegutachtung die Freiheit der Willensentschließung und -betätigung des Beschuldigten und sei insofern nicht mit den Artikeln 1 und 2 des Grundgesetzes vereinbar, in denen die Unantastbarkeit der Menschenwürde sowie das Recht auf freie Persönlichkeitsentfaltung, Leben, körperliche Unversehrtheit und persönliche Freiheit garantiert werden. Dabei ist zu betonen, daß gemäß diesem BGH-Urteil auch dann eine Beeinträchtigung der Willensfreiheit vorliegt, wenn der Beschuldigte in die Durchführung einwilligt. Aus diesem Urteil ergab sich ein grundsätzliches Verwertungsverbot psychophysiologischer Glaubhaftigkeitsgutachten im Bereich der Strafgerichtsbarkeit.

Das BGH-Urteil wurde durch einen **Beschluß des Vorprüfungsausschusses des Bundesverfassungsgerichts (BVerfG) vom 18.08.1981** (Aktenzeichen: 2 BvR 166/81; vgl. BVerfG, 1981, S. 446f.) bestätigt, und zwar bemerkenswerterweise in einem Fall, in dem ein Angeklagter die Durchführung einer psychophysiologischen Glaubhaftigkeitsbegutachtung beantragte, um so einen Unschuldsnachweis zu erbringen. Neben der Unvereinbarkeit der Untersuchungsmethode mit der Menschenwürde wurde der Beschluß auch damit begründet, daß der Beweiswert der psychophysiologischen Glaubhaftigkeitsbeurteilung zu gering sei (das BVerfG ging, ohne zwischen den verschiedenen psychophysiologischen Verfahren zu differenzieren, von einer immerhin 90-prozentigen Treffsicherheit aus). Insbesondere die Frage, ob bei einer freiwilligen Untersuchungsteilnahme von einer Beeinträchtigung der Willensfreiheit ausgegangen werden könne, sowie der paradox anmutende Sachverhalt, daß man einem Beschuldigten die einzige Entlastungsmöglichkeit verwehrt, weil man seine Persönlichkeitsrechte schützen will, führten zu heftigen juristischen Kontroversen (vgl. z.B. Steller, 1987), in deren Folge die Akzeptanz freiwilliger psychophysiologischer Glaubhaftigkeitsuntersuchungen zur Erbringung eines Entlastungsbeweises wuchs. Auf der Ebene höchstrichterlicher Rechtsprechung spiegelte sich dies erstmals in einem **Beschluß des Vorprüfungsausschus-**

ses des BVerfG vom 15.10.1997 (Aktenzeichen: 2 BvR 1211/97; vgl. Scherer, 1998) wider, in welchem die strafprozessuale Verwertbarkeit psychophysiologischer Glaubhaftigkeitsgutachten als Entlastungsbeweis nicht kategorisch aus verfassungsrechtlichen Gründen abgelehnt wurde. Undeutsch (1996, 1997) verweist bereits auf eine Reihe von Gerichten, die psychophysiologische Gutachten als Entlastungsbeweis anerkannt haben, wobei es sich allerdings meist nicht um strafrechtliche Verfahren handelte.

Der zwischenzeitliche Aufschwung der psychophysiologischen Begutachtungspraxis erfuhr jedoch durch die jüngste höchstrichterliche Rechtsprechung wieder einen Rückschlag. So entschied der **BGH** in zwei **Urteilen vom 17.12.1998** (Aktenzeichen: 1 StR 156/98; 1 StR 258/98; vgl. BGH, 2000a, S. 308ff.), daß psychophysiologische Glaubhaftigkeitsbegutachtungen, sofern sie mit dem Einverständnis des Beschuldigten erfolgen, zwar nicht gegen Verfassungsgrundsätze (Schutz der Menschenwürde etc.) verstoßen, jedoch als Beweismittel völlig ungeeignet seien. Die Urteilsbegründung bezieht sich ausschließlich auf den *KFT*[10] und den *TWT*, wobei der Schwerpunkt auf dem *KFT* liegt. Am *KFT* wird in erster Linie kritisiert, daß bereits dessen Grundannahmen wissenschaftlich zweifelhaft seien. Die diesbezügliche Argumentation des BGH korrespondiert weitgehend mit der Darstellung in Abschnitt 2.2.1.3 der vorliegenden Arbeit. Ferner wird der wissenschaftliche Wert der existierenden empirischen Validitätsbefunde zum *KFT* in Zweifel gezogen, wobei die gleichen Mängel an Labor- und Feldstudien angeführt werden, auf die auch in Abschnitt 3.2.2 der vorliegenden Arbeit hingewiesen wird. Was den *TWT* angeht, so werden dessen eingeschränkte praktische Einsatzmöglichkeiten (vgl. Abschnitt 2.2.2.3) bemängelt, aus denen sich ergebe, daß dieses Verfahren zum Zeitpunkt der Hauptverhandlung nicht mehr sinnvoll einsetzbar sei. Genau genommen wird also der Beweiswert des *TWT* nicht grundsätzlich in Frage gestellt, sondern nur auf dessen Problematik im fortgeschrittenen Ermittlungsstadium hingewiesen. Der *GAT*, der gerade auf eine Beseitigung dieses anwendungsbezogenen Defizits des *TWT* abstellt, findet in den beiden BGH-Urteilen keine Berücksichtigung.

[10] Neben dem konventionellen *KFT* wird auch eine neuere methodische Variante, der sog. „Directed Lie Control Question Test“ (DLCQT; z.B. Honts, 1994; Honts & Raskin, 1988; Horowitz, Kircher, Honts & Raskin, 1997), berücksichtigt. Dieser soll gegenüber dem klassischen *KFT* ein erhöhtes Maß an Standardisierung und Objektivität gewährleisten. Die Kontrollfragen des DLCQT thematisieren Verfehlungen, die jeder einmal begangen hat (z.B.: „Haben Sie jemals gelogen?“) und sollen vom Pb bewußt mit einer Lüge beantwortet werden. Diese Instruktion erfolgt mit der Begründung, daß die physiologischen Reaktionen bei den Lügen auf die Kontrollfragen als Vergleichsstandard für die Beurteilung der Reaktionen auf die relevanten Fragen dienen. Dem Pb wird also nahegelegt, daß es für ein entlastendes Ergebnis wichtig ist, bei den Kontrollfragen stark zu reagieren. Somit muß der intendierte Signalwert der Kontrollfragen nicht durch individuelle Formulierung in einem suggestiven Vortest-Interview induziert werden, dessen Gelingen vom manipulativen Geschick des Untersuchers abhängt. Vielmehr ist der Wortlaut der Kontrollfragen standardisierbar und unabhängig von der Interaktion zwischen Pb und Untersucher. Dem BGH-Urteil zufolge führen diese Modifikationen gegnüber dem herkömmlichen *KFT* jedoch nicht zu einer entscheidenden Erhöhung des Beweiswerts.

Dem strafrechtlichen Verwertungsverbot der psychophysiologischen Glaubhaftigkeitsbeurteilung in Deutschland steht eine vergleichsweise breite **juristische Akzeptanz der inhaltsorientierten Vorgehensweise** gegenüber. Wie in Abschnitt 2.1.4 dargestellt, werden seit einem **Grundsatzurteil des BGH aus dem Jahre 1954** inhaltsorientierte Begutachtungen regelmäßig von Gerichten angeordnet, wenn Zeugenaussagen die einzigen Beweismittel darstellen oder das Vorliegen von Falschbezichtigungsmotiven nicht auszuschließen ist, was insbesondere auf Anschuldigungen sexueller Delikte zutrifft. Die jüngste höchstrichterliche Rechtsprechung zur inhaltsorientierten Glaubhaftigkeitsbeurteilung (**BGH-Urteil vom 30.07.1999**; Aktenzeichen: 1 StR 618/98; vgl. BGH, 2000b, S. 164f.) hat lediglich Gütekriterien definiert, die an strafprozessuale Glaubhaftigkeitsgutachten anzulegen sind (insbesondere Nachvollziehbarkeit und Transparenz), ohne jedoch den Beweiswert der inhaltsorientierten Methode in ähnlicher Weise kritisch zu würdigen, wie dies bei den jüngsten BGH-Urteilen zur psychophysiologischen Glaubhaftigkeitsbeurteilung der Fall war. Während in letztgenannten BGH-Urteilen die Validität der psychophysiologischen Glaubhaftigkeitsbeurteilung – allerdings unter einseitiger Konzentration auf den *KFT* – umfassend erörtert wurde (so sind z.B. 60 von 74 Absätzen des Urteilstextes „1 StR 156/98“ dem Beweiswert der Methode gewidmet), wird im aktuellen BGH-Urteil zur inhaltsorientierten Glaubhaftigkeitsbeurteilung (1 StR 618/98) deren Beweiswert lapidar in drei Absätzen abgehandelt. Darin heißt es u.a.:

„Diese sog. Realkennzeichen können als *grundsätzlich empirisch überprüft* angesehen werden. Zwar handelt es sich um *Indikatoren mit jeweils für sich genommen nur geringer Validität*, d. h. mit durchschnittlich *nur wenig über dem Zufallsniveau liegender Bedeutung*. Eine gutachterliche Schlußfolgerung kann aber eine beträchtlich höhere Aussagekraft und damit *Indizwert* für die Glaubhaftigkeit zu beurteilender Angaben erlangen, wenn sie aus der Gesamtheit aller Indikatoren abgeleitet wird. Denn durch das Zusammenwirken der Indikatoren werden deren Fehleranteile insgesamt gesenkt. [...] Dementsprechend lagen die mit Realkennzeichen in Forschungsvorhaben erzielten Ergebnisse *regelmäßig deutlich über dem Zufallsniveau*. Allerdings bestanden dabei *teilweise nicht unerhebliche Fehlerspannen*.“ (S. 11; Hervorhebungen durch den Verfasser der vorliegenden Arbeit)

Diese optimistische Einschätzung ist angesichts der in Abschnitt 2.1.6 dargestellten empirischen Befundlage zur Validität der inhaltsorientierten Glaubhaftigkeitsbeurteilung nicht ohne weiteres nachvollziehbar. Die Behauptung, die Glaubhaftigkeitskriterien seien als grundsätzlich empirisch überprüft anzusehen, trifft zwar zu, sofern man die *Kriterienorientierte Inhaltsanalyse* als Gesamtsystem betrachtet. In Abschnitt 2.1.6.1 wurde jedoch deutlich, daß die Befunde zur Validität einzelner Kriterien sehr heterogen sind. Lediglich die Validität von Kriterium 3 (*Details*) kann als hinreichend

gesichert gelten. Ferner sprechen die Resultate der Feldforschung relativ konsistent für die Validität der Kriterien 2 (*Unordnung*), 4 (*Verknüpfungen*), 5 (*Interaktionen*) und 6 (*Gespräche*), wobei allerdings zu berücksichtigen ist, daß diese Kriterien in experimentellen Studien nicht durchgängig bestätigt werden konnten. Für die übrigen 14 Einzelkriterien ist die Annahme einer gesicherten Validität nicht mit der empirischen Befundlage vereinbar. Diesem Umstand wird zwar in der Formulierung des BGH-Urteils ansatzweise Rechnung getragen, indem von „Indikatoren mit jeweils für sich genommen nur geringer Validität“ gesprochen wird, deren Bedeutung „durchschnittlich nur wenig über dem Zufallsniveau“ liege. Dennoch wird behauptet, daß durch die integrierende Berücksichtigung aller Kriterien im Rahmen der gutachterlichen Schlußfolgerung „Indizwert“ erzielt werden könne, was durch „regelmäßig deutlich über dem Zufallsniveau“ liegende Forschungsergebnisse bestätigt werde. Es erscheint zwar nicht unplausibel, daß sich auch bei Verwendung nur wenig valider Einzelindikatoren treffsichere diagnostische Urteile erzielen lassen, sofern die Einzelindikatoren im Rahmen der diagnostischen Datenintegration angemessen gewichtet und an zusätzlichen diagnostisch relevanten Informationen relativiert werden; allerdings wurden – nach dem Kenntnisstand des Verfassers der vorliegenden Arbeit – bislang in wissenschaftlichen Fachzeitschriften keine Ergebnisse aus kontrollierten Untersuchungen publiziert, die auf eine hohe Treffsicherheit der inhaltsorientierten Glaubhaftigkeitsbeurteilung verweisen (s. Abschnitt 2.1.6.2). Auch die im BGH-Urteil angesprochenen deutlich überzufälligen Forschungsergebnisse dürften sich wohl in erster Linie auf die Beurteilung glaubhafter Aussagen beziehen. Wie in Abschnitt 2.1.6.2 dargestellt, wurden erfundene Aussagen in empirischen Studien durchschnittlich etwa genauso oft falsch wie richtig klassifiziert, wobei allerdings erneut darauf hingewiesen werden muß, daß den Beurteilern neben den Aussagen selbst keine weiteren beurteilungsrelevanten Informationen (aus Persönlichkeits- und Motivanalysen etc.) zur Verfügung standen, so daß die ermittelten Trefferquoten nur sehr eingeschränkt interpretierbar sind.

Die beschriebene **rechtliche Situation** der forensischen Glaubhaftigkeitsbeurteilung ist – zumindest aus wissenschaftlicher Perspektive – in mehrfacher Hinsicht **inkonsistent**. Während die höchstrichterliche Würdigung der psychophysiologischen Methoden auf einer umfassenden Analyse des wissenschaftlichen Erkenntnisstands beruht, wird die wissenschaftliche Fundierung des inhaltsorientierten Ansatzes kaum kritisch durchleuchtet. Während somit die Gründe für die Negation des Beweiswerts der psychophysiologischen Methoden nachvollziehbar sind, steht der dem inhaltsorientierten Ansatz zugebilligte Indizwert nur bedingt im Einklang mit den diesbezüglichen Forschungsresultaten. Obwohl die empirische Befundlage zur inhaltsorientierten Glaubhaftigkeitsbeurteilung (vgl. Abschnitt 2.1.6) nicht auf eine höhere Validität hindeutet als die Forschungsresultate zu den verschiedenen psychophysiologischen Methoden (vgl.

Abschnitt 2.2.1.4, 2.2.2.4 und 2.2.3.2)[11], wird der inhaltsorientierten Glaubhaftigkeitsbeurteilung Indizwert zugesprochen, den psychophysiologischen Verfahren jedoch nicht. Es hat gewissermaßen den Anschein, als ob an verschiedene Methoden der forensischen Glaubhaftigkeitsbeurteilung nicht die gleichen Gütemaßstäbe angelegt würden.

Aus der beschriebenen juristischen Situation ergibt sich die Forderung nach einem direkten Vergleich des Beweiswerts inhaltsorientierter und psychophysiologischer Methoden. Dementsprechend lautet die Fragestellung der vorliegenden Untersuchung, ob bzw. inwiefern sich inhaltsorientierte und psychophysiologische Methoden der Glaubhaftigkeitsbeurteilung hinsichtlich ihrer Validität unterscheiden.

3.2 Konzeption eines Forschungsparadigmas zum direkten Vergleich psychophysiologischer und inhaltsorientierter Methoden

Angesichts der beschriebenen Rechtslage der forensischen Glaubhaftigkeitsbeurteilung in Deutschland mag es verwundern, daß bislang noch kein direkter empirischer Vergleich psychophysiologischer und inhaltsorientierter Methoden vorgenommen wurde. Die Ursache dieses Forschungsdefizits dürfte hauptsächlich mit den **divergierenden praktischen Anwendungsdomainen** beider methodischer Ansätze zusammenhängen. Während die inhaltsorientierte Glaubhaftigkeitsbeurteilung nahezu ausschließlich zur Begutachtung von Zeugenaussagen in Sittlichkeitsprozessen herangezogen wird, werden mit den psychophysiologischen Methoden in erster Linie die Einlassungen von Beschuldigten überprüft.[12] Somit stehen die beiden methodischen Ansätze in der forensischen Praxis nicht in unmittelbarer Konkurrenz, wodurch eine vergleichende Analyse des diagnostischen Werts nicht auf den ersten Blick erforderlich erscheint. Aus dem Umstand, daß die inhaltsorientierte und die psychophysiologische Glaubhaftigkeitsbe-

[11] Steller (1987, S. 166) stellt in diesem Zusammenhang fest, daß die inhaltsorientierte Glaubhaftigkeitsbeurteilung „nicht annähernd so gründlich wissenschaftlich überprüft wurde wie die psychophysiologischen Methoden“. Seit dieser Feststellung Stellers hat die systematische Forschung zwar einige Fortschritte erzielt; in den publizierten Untersuchungen ging es jedoch fast ausschließlich um die Evaluation der *Kriterienorientierten Inhaltsanalyse*, die ja nur ein, wenn auch zentrales, Teilelement der diagnostischen Gesamtprozedur ist. Dagegen liegt bis heute keine aussagekräftige Studie zur kriterienbezogenen Validität der inhaltsorientierten Glaubhaftigkeitsbeurteilung als Gesamtprozedur vor, wobei „aussagekräftig“ bedeutet, daß den Auswertern sämtliche für die diagnostische Urteilsbildung relevanten Informationen (also auch aus Persönlichkeits- und Motivanalysen etc.) zur Verfügung stehen müßten (vgl. Abschnitt 2.1.6.2).

[12] Allerdings sind beide diagnostischen Ansätze nicht per definitionem für unterschiedliche Anwendungsfelder reserviert. Wie in Abschnitt 2.1.4 dargestellt wurde, ist die inhaltsorientierte Glaubhaftigkeitsbeurteilung prinzipiell auch bei der Begutachtung von Beschuldigten einsetzbar. Andererseits weist Steller (1987, S. 167f.) explizit auf die Möglichkeit der Glaubhaftigkeitsbeurteilung von Zeugenaussagen mit Hilfe des *KFT* hin. Dieses insbesondere in den USA bei der Begutachtung vermeintlicher Vergewaltigungsopfer praktizierte Vorgehen ist jedoch – abgesehen von den bereits angesprochenen Mängeln des *KFT* – äußerst problematisch (vgl. Lykken, 1998) und entbehrt bislang jeglicher empirischer Untermauerung (Steller, 1987).

urteilung faktisch bei unterschiedlichen Zielgruppen (Zeugen vs. Beschuldigte) eingesetzt werden, resultiert natürlich die Frage, wie man die beiden methodischen Ansätze dennoch einem direkten empirischen Vergleich unterziehen kann. Die Konzeption eines entsprechenden Untersuchungsdesigns wird im folgenden erläutert.

3.2.1 Beschreibung einer geeigneten empirischen Fallkonstellation

Eine ideale Konstellation für einen direkten empirischen Vergleich psychophysiologischer und inhaltsorientierter Methoden ist dann gegeben, wenn die Glaubhaftigkeit der Bekundungen ein und derselben Person mit beiden diagnostischen Ansätzen beurteilt werden kann, so daß die **Treffsicherheit beider Methoden intraindividuell vergleichbar** ist. Liegt eine Vielzahl solcher empirischer Idealfälle vor, besteht zudem die Möglichkeit, etwaige Unterschiede zwischen der inhaltsorientierten und der psychophysiologischen Glaubhaftigkeitsbeurteilung mit Hilfe statistischer Methoden zufallskritisch abzusichern.

Da der inhaltsorientierte Ansatz faktisch nur bei Zeugen, der psychophysiologische Ansatz dagegen nur bei Beschuldigten zum Einsatz kommt, ist der geforderte intraindividuelle Vergleich nur in solchen Fällen möglich, in denen die Pbn gleichzeitig sowohl als potentielle Zeugen als auch als mutmaßliche Täter in Frage kommen. Die beschriebene Konstellation dürfte in der forensischen Praxis insbesondere dann vorliegen, wenn zwei einer kriminellen Tat gleichermaßen verdächtige Personen jeweils die eigene Täterschaft abstreiten und in ihren vermeintlichen Zeugenaussagen den jeweils anderen als Täter bezichtigen.[13] In solchen Fällen kann die Glaubhaftigkeit der Aussagen beider Pbn jeweils sowohl inhaltsanalytisch als auch psychophysiologisch untersucht werden, wobei die psychophysiologische Begutachtung speziell auf die Glaubhaftigkeit der Abstreitung eigener Täterschaft und die inhaltsorientierte Begutachtung speziell auf die Glaubhaftigkeit der Beschuldigung des anderen Tatverdächtigen abzielt. Diejenige Methode, mit der sich in solchen Situationen die Glaubhaftigkeit der Aussagen zuverlässiger beurteilen läßt bzw. die zu einer verläßlicheren Aufdeckung des wahren Status der Pbn (schuldig vs. unschuldig) führt, kann prinzipiell auch als die bessere Methode der Glaubhaftigkeitsbeurteilung gelten.

Der Vergleich inhaltsorientierter und psychophysiologischer Methoden anhand der beschriebenen Fallkonstellation kann grundsätzlich **sowohl im Feld als auch im Experi-**

[13] Daß diese Konstellation von einiger praktischer Relevanz ist, bemerkt in anderem Zusammenhang bereits Undeutsch (1954, S. 148): „Besondere Probleme bestehen dort, wo die Einlassungen zweier oder mehrerer Angeklagter in Frage stehen, die sich möglicherweise gegenseitig zu Unrecht belasten [...]“.

ment erfolgen. Während man bei Feldforschung darauf angewiesen wäre, forensische Realfälle zusammenzutragen, die die beschriebenen Voraussetzungen erfüllen und in denen sowohl eine psychophysiologische als auch eine inhaltsorientierte Begutachtung der Pbn stattfand, könnte im Experiment eine Simulation der beschriebenen Fallkonstellation erfolgen. Bevor auf die konkrete Vorgehensweise in der vorliegenden Untersuchung eingegangen wird, sollen die grundsätzlichen Vor- und Nachteile der Feld- und Experimentalforschung auf dem Gebiet der forensischen Aussagepsychologie erläutert werden.

3.2.2 Exkurs: Kritische Gegenüberstellung der Feld- und Experimentalforschung zur forensischen Glaubhaftigkeitsbeurteilung

Die Frage, ob aussagepsychologische Forschung eher im Feld oder im Labor stattfinden sollte, ist seit jeher Gegenstand **kontroverser Diskussionen**. Dies gilt für die inhaltsorientierte und die psychophysiologische Glaubhaftigkeitsbeurteilung gleichermaßen. Insbesondere auf dem Gebiet der **inhaltsorientierten Glaubhaftigkeitsbeurteilung** war **experimentelle Forschung** lange Zeit verpönt, da man davon ausging, daß die artifizielle Situation im Experiment aufgrund **fehlender Lebensnähe** nicht mit den Bedingungen vergleichbar sei, unter denen „echte“ Beobachtungen und Aussagen gemacht werden (Arntzen, 1983b, 1993; Undeutsch, 1967, 1982, 1984). Arntzens (1983b, S. 524) Argumentation zufolge fehlt in der experimentellen Simulation „vor allem die *persönliche, oft erhebliche gefühlsmäßige Betroffenheit*, die sich etwa bei Vergewaltigungsdelikten und Raubüberfällen auf Beobachtungs- und Einprägungsintensität auswirkt“ (Hervorhebung im Original). Auch „*nachgespielte Szenen, die auf Filme und Videobänder aufgenommen werden*“, eigneten sich nicht als Ersatz für „*Zeugenbeobachtungen, die sich aus Eigenerlebnissen unter natürlichen komplexen Bedingungen ergeben*“ (Arntzen, 1993, S. 9, Hervorhebung im Original). Diese Ablehnung experimenteller Aussageforschung hatte zur Konsequenz, daß die Validität der inhaltsorientierten Glaubhaftigkeitsbeurteilung anfangs fast ausschließlich auf dem Wege systematischer Feldbeobachtung überprüft wurde. Erst in den achziger Jahren regte sich zunehmend Kritik an diesem einseitigen Vorgehen (Köhnken, 1986, 1987; Köhnken & Wegener, 1982, 1985), wobei insbesondere zwei Aspekte der **Feldforschung** bemängelt wurden – die Wahl der **Validierungskriterien** und die **Selektivität des Untersuchungsmaterials** (vgl. Köhnken, 1990).

Grundsätzlich hat die Überprüfung der Treffsicherheit zu erfolgen, indem die Resultate der inhaltsorientierten Glaubhaftigkeitsbeurteilung mit einem unabhängig hiervon zu erhebenden Außenkriterium verglichen werden, welches den tatsächlichen Wahrheits-

status der Aussagen (erlebnisbezogen vs. erfunden) indiziert. In der Feldforschung wurden in erster Linie **Geständnisse der Beschuldigten** als Validierungskriterium herangezogen. Die Beziehung zwischen dem Geständnis eines Beschuldigten und dem Wahrheitsgehalt einer belastenden Zeugenaussage ist jedoch aus mehreren Gründen nicht eindeutig (vgl. Köhnken & Wegener, 1985). Zum einen sollte die Wahrscheinlichkeit falscher Geständnisse nicht unterschätzt werden. Letztere können z.B. aus prozeßtaktischen Erwägungen erfolgen, so etwa um die Einstellung weiterer Ermittlungen zu bewirken oder um dem Gericht gegenüber Reue zu signalisieren und somit eine Strafmilderung zu erreichen. Weiterhin besteht die Gefahr, daß Unschuldige Geständnisse ablegen, weil sie dem Vernehmungsdruck nicht mehr standhalten. Zum anderen werden in einem erheblichen Teil der forensischen Realfälle die Geständnisse erst abgelegt, nachdem das positive Ergebnis der inhaltsorientierten Zeugenbegutachtung bekannt wurde und somit die Beweislast gegen den Beschuldigten erhöhte, so daß hier die Unabhängigkeit von Diagnose und Validierungskriterium nicht gewahrt ist. Auch an den anderen in Feldstudien verwendeten Validierungskriterien (Gerichtsurteile, Experteneinschätzungen, medizinische Befunde, Ergebnisse psychophysiologischer Begutachtungen der Beschuldigten, Aussagewiderrufe der mutmaßlichen Zeugen) ist auszusetzen, daß sie jeweils fehlerbehaftet sein können und somit nicht mit absoluter Sicherheit den tatsächlichen Wahrheitsstatus der Aussagen indizieren. So dürfte bei **Gerichtsentscheidungen** ein systematischer Fehler insofern vorliegen, als Gerichte „im Zweifel für den Angeklagten“ entscheiden, d.h. ein unbekannter Anteil der Schuldigen wird aus Mangel an Beweisen frei gesprochen. Zudem sind gerade auch die Gerichtsurteile wohl nur in den seltensten Fällen unabhängig vom Ergebnis der inhaltsorientierten Glaubhaftigkeitsbeurteilung, das ja im Gerichtsverfahren als Beweismittel herangezogen wird. Die beiden genannten Unzulänglichkeiten von Gerichtsurteilen kann man umgehen, indem man den Realitätsgehalt der Aussagen von **Experten(gruppen)** einschätzen läßt, denen für ihr Urteil sämtliche Ermittlungsergebnisse unter Ausschluß des Befundes der inhaltsorientierten Aussagebegutachtung zur Verfügung stehen und die man instruiert, bei der Beurteilung der Schuldfrage weniger konservativ vorzugehen als im Rahmen von Gerichtsentscheidungen. Dennoch ist auch hier zu berücksichtigen, daß inhaltsorientierte Begutachtungen i.d.R. angeordnet werden, wenn die übrige Beweislage mehrdeutig ist. Insofern ist nicht zu erwarten, daß Experten(gruppen) anhand der (unklaren) Beweislage den Realitätsgehalt der Aussagen präziser einschätzen als die Gerichte. **Medizinische Befunde** lassen i.d.R. keine eindeutigen Rückschlüsse auf einen stattgefundenen körperlichen Mißbrauch oder gar die Identität des Täters zu, da für die körperlichen Auffälligkeiten auch alternative Erklärungen in Betracht kommen; insbesondere das Fehlen körperlicher Auffälligkeiten indiziert nicht zwangsläufig, daß kein Mißbrauch stattgefunden hat. Daß **psychophysiologische Glaubhaftigkeitsbeurteilungen** der Beschuldigten eine gewisse Fehleranfälligkeit aufweisen, wurde in den Abschnitten 2.2.1.4,

2.2.2.4 und 2.2.3.2 erläutert. **Aussagewiderrufe von Zeugen** sind als Validierungskriterium ähnlich problematisch wie Geständnisse der Beschuldigten. So ist es – insbesondere im Zusammenhang mit Sittlichkeitsprozessen – denkbar, daß wahrheitsgemäße Behauptungen zurückgezogen werden, um der mit den Vernehmungen und dem Gerichtsverfahren verbundenen psychischen Belastung zu entgehen. Nicht zuletzt die Angst vor dem Beschuldigten, der häufig dem engeren Familienkreis angehört, kommt hier als Widerrufmotiv in Betracht.

Unter dem Gesichtspunkt der Selektivität des Untersuchungsmaterials ist insbesondere zu bemängeln, daß ein Großteil der inhaltsorientierten Glaubhaftigkeitsbegutachtungen noch in der Ermittlungsphase von der Staatsanwaltschaft angefordert wird. Werden Beschuldigungen hier als unglaubhaft eingestuft, erfolgt in aller Regel eine Einstellung des Verfahrens. Somit kann die Mehrzahl der als unglaubhaft klassifizierten Zeugenaussagen gar nicht an einem „objektiven“ Außenkriterium validiert werden; folglich sind bei Validitätsuntersuchungen im Feld die **als „glaubhaft“ diagnostizierten Aussagen überrepräsentiert** (vgl. Köhnken & Wegener, 1985). Im Zusammenhang mit der Stichprobenselektivität sei auch an die von Wells und Loftus (1991; s. Abschnitt 2.1.6.1) vorgetragene Kritik erinnert, daß möglicherweise viele **wahre Zeugenaussagen** erst gar nicht in die inhaltsorientierte Begutachtung gelangen, weil sie **nicht in überzeugender Weise vorgetragen** werden und somit eine genauere Untersuchung der Tatvorwürfe von den zuständigen Instanzen für überflüssig gehalten wird. Ferner liegt auch insofern eine Selektivität des Untersuchungsmaterials vor, als inhaltsorientierte Glaubhaftigkeitsbeurteilungen in der forensischen Praxis weitestgehend bei kindlichen Aussagen zu Sexualdelikten durchgeführt werden, so daß die Ergebnisse von Feldstudien **nicht ohne weiteres auf andere Altersgruppen und Delikttypen generalisierbar** sind.

Die Kritik an der Feldforschung ging mit einer Zunahme experimenteller Studien einher. Dem Vorwurf zu geringer Lebensnähe wurde dabei insbesondere entgegnet, daß die Generalisierbarkeit experimenteller Befunde nicht von der physischen Identität zwischen Simulation und Realsituation abhänge, sondern von der „*Vergleichbarkeit* der jeweils durch externe Faktoren ausgelösten *psychologischen Reaktionen*“ (Köhnken & Wegener, 1985, S. 108, Hervorhebungen im Original). So seien etwa die wesentlichen psychologischen Momente des Erlebens einer Straftat gegen die sexuelle Selbstbestimmung die **negative emotionale Tönung** des Geschehens sowie die **Eigenbeteiligung** und der weitgehende **Kontrollverlust** des Betroffenen. Diese psychologischen Momente ließen sich durchaus im Experiment realisieren, indem man Aussagethemen wählt, für die die genannten psychologischen Momente ebenfalls typisch sind, z.B. Ge-

burtserleben aus der Sicht der Mutter (Wolf & Steller, 1997) oder medizinische Eingriffe aus der Perspektive des Patienten (Steller et al., 1992; s. Abschnitt 2.1.6.1).

Feld- und Laboruntersuchungen zur **psychophysiologischen Glaubhaftigkeitsbeurteilung** sind grundsätzlich mit den gleichen Vor- und Nachteilen behaftet wie auch bei der inhaltsorientierten Methode. Bei den in **Felduntersuchungen** herangezogenen **Validierungskriterien** handelt es sich in erster Linie um Geständnisse, Gerichtsurteile und Experten-Entscheidungen. Die Schwachpunkte dieser Kriterien wurden bereits im Zusammenhang mit der Feldforschung zur inhaltsorientierten Glaubhaftigkeitsbeurteilung erläutert (s.o.) und gelten genauso in bezug auf die psychophysiologische Glaubhaftigkeitsbeurteilung. Zudem kann insbesondere für Gerichtsurteile und Geständnisse in aller Regel keine Unabhängigkeit vom Ergebnis der psychophysiologischen Begutachtung angenommen werden.

Auch bei der Feldforschung zur psychophysiologischen Glaubhaftigkeitsbeurteilung ist eine **Selektivität des Untersuchungsmaterials** zu bemängeln. So werden in Feldstudien nur solche Fälle aufgenommen, bei denen das Ergebnis der psychophysiologischen Begutachtung des Beschuldigten anhand von mindestens einem der genannten Validierungskriterien überprüft werden kann. Dagegen werden von vorneherein alle diejenigen Fälle aus der Analyse ausgeschlossen, bei denen ein solcher Maßstab fehlt. Abgesehen davon, daß also nicht die Gesamtmenge der forensischen Realfälle bzw. eine repräsentative Stichprobe sondern nur eine verzerrte Teilmenge analysiert wird, ist besonders problematisch, daß es sich bei dieser Teilmenge überwiegend um solche Fälle handelt, die a priori darauf angelegt sind, die Treffsicherheit der psychophysiologischen Glaubhaftigkeitsbeurteilung zu bestätigen (Lykken, 1998). In diesem Zusammenhang ist hervorzuheben, daß psychophysiologische Begutachtungen in forensischen Realfällen hauptsächlich dann vorgenommen werden, wenn die übrige Beweislage unklar ist. Infolgedessen besteht logischerweise nur eine äußerst geringe Chance, daß das psychophysiologische Begutachtungsergebnis durch unabhängige Evidenz stringent widerlegt werden kann. So werden z.B. die wenig aussichtsreichen sonstigen Ermittlungen häufig eingestellt, sobald ein Verdächtiger durch den psychophysiologischen Begutachtungsbefund belastet wird, und zwar auch dann, wenn der Beschuldigte kein Geständnis ablegt. Dies hat zur Konsequenz, daß ein etwaiger falsch positiver Befund nicht durch ein Außenkriterium (unabhängiger Beweis für die Unschuld des Verdächtigen) invalidiert werden kann und folglich auch keine Berücksichtigung in Feldstudien erfährt. In letzteren werden falsch positive Befunde eher fälschlich als „valide positive“ Befunde berücksichtigt, und zwar dann, wenn falsche Geständnisse der unschuldigen Pbn erfolgen und somit ein „Validierungskriterium“ vorhanden ist – wobei es freilich unentdeckt bleibt, daß dieses Kriterium in Wirklichkeit der „Bestätigung“ einer falschen Diagnose

dient. Auch falsch negative Befunde werden aus Felduntersuchungen systematisch mangels Validierungskriterium ausgeschlossen. So ist es beispielsweise kaum zu erwarten, daß ein Schuldiger, der durch einen fehlerhaften psychophysiologischen Untersuchungsbefund entlastet wird, anschließend noch ein Geständnis ablegt und dadurch ein Kriterium schafft, durch welches das Begutachtungsergebnis invalidiert werden könnte. Dagegen würde der gleiche falsch negative Befund fälschlich als „valide negativer" Befund in Felduntersuchungen berücksichtigt, falls irgendein unschuldiger Verdächtiger in der gleichen Angelegenheit ein falsches Geständnis ablegt und somit ein Kriterium liefert, durch welches der psychophysiologische Untersuchungsbefund – wohlgemerkt fälschlicherweise – „bestätigt" wird.

Auch bei der psychophysiologischen Glaubhaftigkeitsbeurteilung steht der beschriebenen Problematik der Feldforschung das Argument zu **geringer Lebensnähe bzw. externer Validität der Laborexperimente** gegenüber. Diesbezüglich wird z.B. kritisiert, daß in Laborexperimenten meist **keine repräsentativen**, sondern größtenteils studentische **Stichproben** untersucht werden (Steller, 1987; Lykken, 1998). Ein gravierenderes Problem dürfte jedoch darin liegen, daß sich die in forensischen Realfällen gegebene **emotionale und motivationale Ausgangslage** der Pbn nicht adäquat mit Hilfe experimenteller Simulationen rekonstruieren läßt. Ein Scheinverbrechen ist grundsätzlich nicht mit einer realen Straftat vergleichbar. Dies gilt sowohl für die Beweggründe der Tat als auch für etwaige mit der Täterschaft (bzw. Tatverdächtigung) assoziierte Emotionen, wie z.B. Schuldgefühle und Angst (zu unrecht) bestraft zu werden (Furedy, 1986; Lykken, 1998). So werden die Pbn experimenteller Studien in aller Regel den Versuchsbedingungen zugeweisen, d.h. sie entscheiden sich, im Gegensatz zu Realfällen, nicht freiwillig für oder gegen die Begehung eines „Delikts" (Lykken, 1998); und die im Falle eines positiven Begutachtungsbefundes drohenden Konsequenzen sind in Simulationsstudien vergleichsweise nichtig (z.B. Verlust einer in Aussicht gestellten Geldprämie vs. langjährige Haftstrafe).

Allerdings muß bei der kritischen Würdigung der Experimentalforschung **zwischen den verschiedenen psychophysiologischen Methoden differenziert** werden. Wie in den Abschnitten 2.2.1.3 bzw. 2.2.2.2 erläutert wurde, spielen motivationale und emotionale Faktoren nur in den Grundannahmen zum ***KFT*** eine Rolle, während die Funktionsweise des ***TWT*** in erster Linie an kognitive Prozesse gebunden ist. Folglich gilt die genannte Einschränkung der externen Validität von Feldforschung hauptsächlich für den *KFT*, wohingegen man argumentieren kann, daß sich die für den *TWT* relevanten kognitiven Abläufe grundsätzlich nicht zwischen forensichem Ernstfall und experimenteller Simulation unterscheiden. Allerdings ist die Generalisierbarkeit der meisten experimentellen Befunde zum *TWT* in anderer Hinsicht eingeschränkt. So wird in Analogstudien in aller

Regel dafür Sorge getragen, daß eine wesentliche Anwendungsvoraussetzung dieses Verfahrens erfüllt ist, nämlich daß der Täter alle oder zumindest die meisten relevanten Items bei der Tatbegehung wahrgenommen hat und sich zum Zeitpunkt der Testung noch an sie erinnert. Es ist jedoch zumindest fraglich, ob diese Bedingung in den meisten Realfällen annähernd so gut erfüllt ist wie in den Laboruntersuchungen; die erhöhte Rate falsch negativer Befunde in den wenigen vorhandenen Feldstudien zum *TWT* deutet eher auf das Gegenteil hin (s. Abschnitt 2.2.2.4). Eine eingeschränkte externe Validität experimenteller Forschung muß auch für den ***GAT*** angenommen werden. Hier gilt ebenso wie beim *TWT* das Argument, daß die Generalisierbarkeit im Hinblick auf Realsituationen eingeschränkt ist, weil dort das Tatwissen der Pbn nicht so gut kontrolliert werden kann wie im Experiment. Inwiefern jedoch die unterschiedliche motivationale und emotionale Ausgangslage in Ernstfall und Simulation die Generalisierbarkeit der experimentellen Befunde einschränkt, ist schwer einzuschätzen. Befunde im Rahmen des sog. „Differentiation of Deception"-Paradigmas (s. Fußnote 8) deuten zumindest an, daß der im *GAT* anvisierte aktivierungssteigernde Effekt der wahrheitswidrigen Verneinung relevanter Items nicht von motivationalen und emotionalen Randbedingungen abhängt (vgl. zusammenfassend Gödert et al., 2001). Allerdings ist zu betonen, daß es sich beim „Differentiation of Deception"-Paradigma um einen rein experimentellen Forschungsansatz handelt, so daß die externe Validität der hiermit gewonnenen Erkenntnisse ihrerseits diskussionswürdig ist.

Zusammenfassend ist festzuhalten, daß sowohl die Feld- als auch die Experimentalforschung auf dem Gebiet der forensischen Glaubhaftigkeitsbeurteilung jeweils spezifische Stärken und Schwächen aufweisen. Prinzipiell erscheint es daher sinnvoll, beide Forschungsstrategien simultan anzuwenden, um eine gegenseitige Kompensation der wesentlichen Nachteile zu erzielen. Allerdings besteht eine noch nicht genannte **Einschränkung der Feldforschung** darin, daß sie stark von den **rechtlichen Rahmenbedingungen** abhängt. Bezogen auf die juristische Situation in Deutschland bedeutet dies konkret, daß Feldforschung hierzulande nur im Bereich der inhaltsorientierten Glaubhaftigkeitsbeurteilung möglich ist, während eine systematische Erprobung der psychophysiologischen Methoden in forensischen Realfällen durch das geltende strafrechtliche Verwertungsverbot (s. Abschnitt 3.1) weitestgehend verhindert wird. Daraus ergibt sich auch, daß ein direkter empirischer Vergleich der psychophysiologischen und der inhaltsorientierten Glaubhaftigkeitsbeurteilung hierzulande vorerst nur im Experiment erfolgen kann.

3.2.3 Grundentwurf eines experimentellen Untersuchungsdesigns

Um die inhaltsorientierte experimentell mit der psychophysiologischen Glaubhaftigkeitsbeurteilung vergleichen zu können, ist es erforderlich, die in Abschnitt 3.2.1 beschriebene **Fallkonstellation** zu **simulieren**. Die wesentlichen Elemente einer solchen experimentellen Simulation werden im folgenden beschrieben. Wie in Abschnitt 3.2.1 angesprochen wurde, müssen die **Pbn gleichzeitig sowohl als Zeugen als auch als Täter in Frage kommen**, so daß sie sowohl inhaltsanalytisch als auch psychophysiologisch untersucht werden können und somit ein intraindividueller Vergleich beider methodischer Ansätze erfolgen kann. Darüber hinaus ist es jedoch auch erforderlich, den Wahrheitsgehalt der Aussagen bzw. den Status der Pbn zu variieren, da die Treffsicherheit beider Ansätze sowohl bezüglich glaubhafter als auch unglaubhafter Aussagen ermittelt werden soll. Das heißt man benötigt sowohl Pbn, die unglaubhafte Angaben machen als auch solche, deren Bekundungen glaubhaft sind.

Alle genannten Erfordernisse sind grundsätzlich realisierbar, indem jeweils zwei Pbn gleichzeitig an einer Verbrechenssimulation teilnehmen (vgl. Abbildung 1). Einer davon begeht das Delikt (*Täter*) und wird dabei von dem anderen Pb (*Zeuge*) beobachtet. Nach der Begehung bzw. Beobachtung des Scheinverbrechens stehen sowohl der *Täter* als auch der *Zeuge* unter Tatverdacht. Beide werden instruiert, zu dem fraglichen Sachverhalt Stellung zu nehmen, um sich selbst zu entlasten und den anderen unter Tatverdacht stehenden Pb zu beschuldigen. Der ***Zeuge*** bewerkstelligt dies, indem er nur wahrheitsgemäße Angaben macht, d.h. indem er **wahrheitsgemäß die Täterschaft abstreitet** und eine **wahrheitsgemäße Zeugenaussage** über den tatsächlichen Tathergang macht. Der ***Täter*** beteuert die eigene Unschuld[14], indem er nur wahrheitswidrige Angaben macht, d.h. indem er **wahrheitswidrig die Täterschaft abstreitet** und eine **wahrheitswidrige „Zeugenaussage“** darüber abgibt, wie er angeblich selber als Zeuge den anderen Pb bei der Tatbegehung beobachtet habe. Um die Motivation zu erhöhen, wird sowohl dem *Täter* als auch dem *Zeugen* eine **finanzielle Belohnung in Aussicht** gestellt für den Fall, daß es ihnen gelinge, in den Untersuchungen zur Aufklärung des Delikts für unschuldig befunden zu werden bzw. den Tatverdacht auf den anderen Pb abzuwenden. Die Untersuchungen bestehen bei beiden Pbn jeweils aus einer psychophysiologischen und einer inhaltsorientierten Glaubhaftigkeitsbeurteilung, welche jeweils von hinsichtlich der Bedingungszugehörigkeit der Pbn uninformierten Versuchsleitern durchgeführt werden. Mit der **psychophysiologischen Aussagebeurteilung** wird jeweils die **Glaubhaftigkeit der Abstreitung der eigenen Täterschaft** analysiert. An-

[14] Es sei erneut darauf hingewiesen, daß die Bezeichungen „schuldig“ bzw. „unschuldig“ im Kontext experimenteller Scheinverbrechen nicht im moralischen oder juristisch-normativen Sinne zu verstehen sind, sondern ausschließlich der plakativen Kennzeichnung der experimentellen Bedingungszugehörigkeit der Pbn (Begehung vs. Nichtbegehung eines simulierten Delikts) dienen.

hand der **Inhaltsanalyse** überprüft man jeweils die **Glaubhaftigkeit der vermeintlichen Zeugenaussage**.

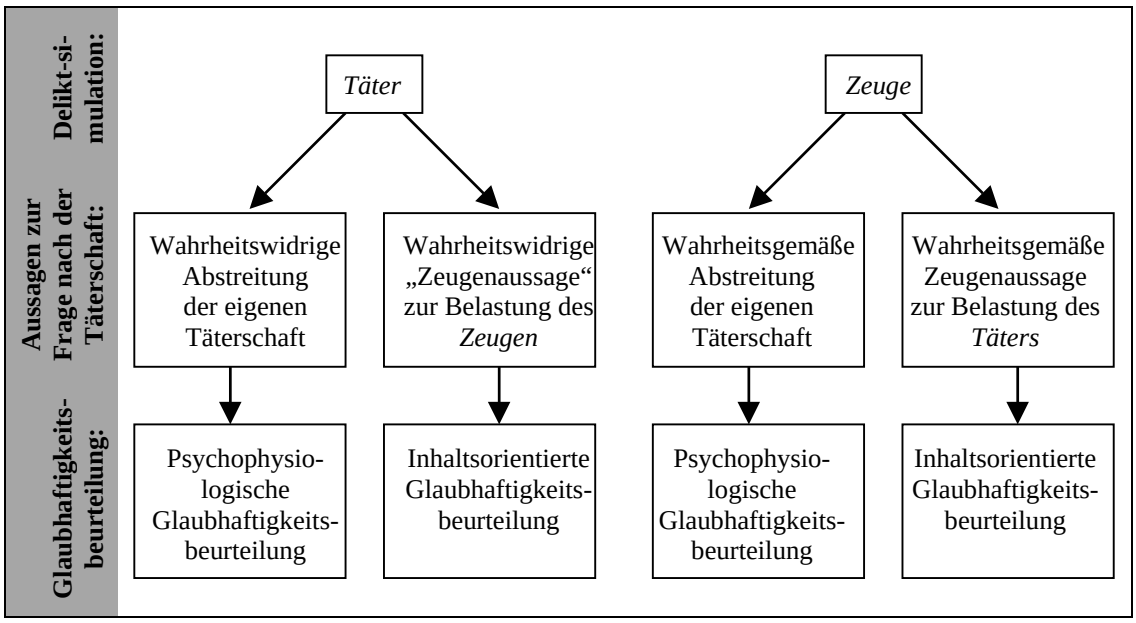


Abbildung 1. Experimentelles Paradigma zum direkten Vergleich der inhaltsorientierten und der psychophysiologischen Glaubhaftigkeitsbeurteilung

Somit ergibt sich grundsätzlich ein **zweifaktorielles Versuchsdesign**, in welchem der Wahrheitsgehalt der Aussage (bzw. Status der aussagenden Person) interindividuell und die Methode der Glaubhaftigkeitsbeurteilung intraindividuell variiert wird. Abhängige Variable ist die Treffsicherheit der Glaubhaftigkeitsbeurteilung.

3.2.4 Konkretisierung und Erweiterung des experimentellen Grunddesigns in der vorliegenden Untersuchung

Als Delikttyp für die experimentelle Verbrechenssimulation wird in der vorliegenden Studie ein **Gelddiebstahl** gewählt (s. genauer Abschnitt 4.3.3). Als psychophysiologische Begutachtungsmethode wird der ***GAT*** gewählt. Auf eine Berücksichtigung des *KFT* wird verzichtet, da dieses Verfahren, wie in Abschnitt 2.2.1.3 dargestellt wurde, auf sehr zweifelhaften Grundannahmen beruht. Demgegenüber wird der Beweiswert des *TWT* auch durch die höchstrichterliche Rechtsprechung nicht generell in Abrede gestellt, sondern nur im Hinblick auf eine Anwendung im fortgeschrittenen Ermittlungsstadium (s. Abschnitt 3.1). Der *GAT* als Modifikation des *TWT* zielt aber gerade auf eine Beseitigung dieses speziellen Anwendungsdefizits ab, indem er die Möglichkeit bieten soll, zwischen Schuldigen und Unschuldigen mit Tatwissen zu differenzieren (vgl. Abschnitt 2.2.3.1). Insofern stellt der *GAT* z.Z. die meistversprechende Methode

der psychophysiologischen Glaubhaftigkeitsbeurteilung dar und soll konsequenterweise in der vorliegenden Untersuchung zur Erprobung kommen.

Das in Abschnitt 3.2.3 beschriebene experimentelle Grunddesign bedarf im Hinblick auf die Aussagekraft bzw. Interpretierbarkeit der damit erzielten Ergebnisse noch einiger **Erweiterungen**. Beide unabhängigen Variablen (UVn) müssen gegenüber dem oben dargestellten Grunddesign jeweils um eine Ausprägung ergänzt werden (s. Tabelle 11). Die Gründe hierfür seien im folgenden erläutert.

Tabelle 11. Versuchsplan der vorliegenden Studie

		Status der aussagenden Person		
		Täter	*Zeugen*	*falsche Zeugen*
Methode der Glaubhaftigkeits-beurteilung	inhaltsorientiert *GAT*			
	naiv			

Wie Tabelle 11 zu entnehmen ist, wird die UV **„Status der aussagenden Person“** erweitert, indem neben den *Tätern* und *Zeugen* noch eine dritte Gruppe in den Versuchsplan aufgenommen wird, die ***falschen Zeugen***. Bei den *falschen Zeugen* handelt es sich um Pbn, die weder als Täter noch als (echte) Zeugen in den inszenierten Diebstahl verwickelt werden. Sie können auch **keine Kenntnis der kritischen Tatdetails** erlangen, die im *GAT* als relevante Items verwendet werden. Stattdessen werden sie aufgefordert, eine **„Zeugenaussage“** zu **erfinden**, in welcher sie eine fiktive Person des Diebstahls bezichtigen. Neben dem Abgeben der erfundenen „Zeugenaussage“, welche inhaltsanalytisch auf ihre Glaubhaftigkeit hin untersucht wird, wird auch mit den *falschen Zeugen* ein *GAT* vorgenommen, um die Glaubhaftigkeit der Abstreitung der Täterschaft zu überprüfen. Die *falschen Zeugen* werden instruiert, daß durch den *GAT* überprüft werde, ob sie durch ihre Zeugenaussagen und die damit verbundene Bezichtigung einer anderen Person möglicherweise versuchten, die eigene Täterschaft zu vertuschen. Den *falschen Zeugen* wird, ebenso wie den *Tätern* und den (echten) *Zeugen*, als motivationaler Anreiz eine finanzielle Belohnung in Aussicht gestellt, falls es ihnen gelinge, hinsichtlich ihrer Bekundungen als glaubhaft eingestuft zu werden.

Die Gruppe *falsche Zeugen* dient als Kontrollbedingung in zweierlei Hinsicht. Zum einen dient die **Bedingung *falsche Zeugen – inhaltsorientierte Glaubhaftigkeitsbeurteilung* als Kontrollbedingung für die Bedingung *Täter – inhaltsorientierte Glaubhaftigkeitsbeurteilung*** (vgl. Tabelle 11). Dies sei im folgenden erläutert. Wie in Abschnitt 3.2.3 dargestellt wurde, sollen die *Täter* eine „Zeugenaussage“ erfinden, in der sie den wahren *Zeugen* der Täterschaft beschuldigen. Es wäre allerdings **denkbar, daß**

die *Täter* beim Erfinden der Zeugenaussagen auf die eigenen realen Erfahrungen bei der Begehung des Diebstahls zurückgreifen und diese in ihren Schilderungen auf die von ihnen beschuldigte Person (den wahren *Zeugen*) übertragen. Dies hätte zur Folge, daß die vermeintlich erfundenen Aussagen der *Täter* doch zu einem gewissen Ausmaß auf einer Erlebnisgrundlage beruhten, was wiederum die inhaltliche Qualität der Aussagen steigern und somit zu einer Überschätzung der Glaubhaftigkeit der Aussagen führen könnte. Um nun kontrollieren zu können, ob die vermeintlich vollständig erfundenen Zeugenaussagen der *Täter* nicht doch zu einem gewissen Anteil auf einer Erlebnisgrundlage basieren, die sich in einer erhöhten inhaltlichen Aussagequalität niederschlägt, muß eine Kontrollbedingung eingeführt werden, in welcher Pbn, die eindeutig nicht auf eine Erlebnisgrundlage zurückgreifen können, „Zeugenaussagen" zu dem gleichen Delikt machen. Diese Voraussetzungen erfüllt die Gruppe der *falschen Zeugen*. Sollte sich bei der Inhaltsanalyse zeigen, daß die Aussagen der *Täter* eine höhere inhaltliche Qualität aufweisen als die Aussagen der *falschen Zeugen*, so würde dies bedeuten, daß die Voraussetzungen für die Anwendung der inhaltsorientierten Glaubhaftigkeitsbeurteilung bei den *Tätern* nicht (in vollem Umfang) erfüllt wären bzw. daß die diesbezüglich ermittelte Trefferquote nur eingeschränkt interpretierbar wäre.

Zum anderen dient die **Bedingung *falsche Zeugen – GAT* als Kontrolle für die Bedingungen *(echte) Zeugen – GAT* und *Täter – GAT*** (vgl. Tabelle 11). Zur Erläuterung: Der Grundannahme des *GAT* zufolge ist dieses Verfahren geeignet, zwischen drei Personengruppen zu differenzieren: (1) Schuldige, (2) Unschuldige mit Kenntnis von Tatdetails, (3) Unschuldige ohne Kenntnis von Tatdetails (vgl. Abschnitt 2.2.3.1). Die ersten beiden Personengruppen sind im vorliegenden Versuchsplan in Form der Gruppen *Täter* bzw. *Zeugen* repräsentiert. Durch Einführung der Kontrollgruppe *falsche Zeugen* findet auch die Gruppe der **Unschuldigen ohne Kenntnis von Tatdetails** Berücksichtigung[15]. Somit kann auch überprüft werden, inwiefern sich unschuldige Pbn ohne Tatwissen mit Hilfe *des GAT* von den beiden anderen Personengruppen differenzieren lassen, d.h. erst durch Aufnahme der Kontrollgruppe ist die Gundannahme *des GAT* einer vollständigen Überprüfung zugänglich.

Aus Tabelle 11 geht weiterhin hervor, daß die UV **„Methode der Glaubhaftigkeitsbeurteilung“** gegenüber dem in Abschnitt 3.2.3 beschriebenen experimentellen Grunddesign um die Ausprägung ***naive Glaubhaftigkeitsbeurteilung*** erweitert wird. Die *naive Glaubhaftigkeitsbeurteilung* wird als Kontrollbedingung in den Versuchsplan aufgenommen, um überprüfen zu können, inwiefern die inhaltsanalytische und die psychophysiologische Methode der Glaubhaftigkeitsbeurteilung einer rein intuitiven Beurtei-

[15] Der Begriff „unschuldig“ bezieht sich in diesem Zusammenhang ausschließlich auf die Nichttäterschaft der *falschen Zeugen* und ist losgelöst von der Tatsache zu sehen, daß diese Pbn andererseits absichtlich eine falsche Zeugenaussage machen.

lung überlegen sind bzw. inwiefern sie einen zusätzlichen Beitrag zu den Informationen erbringen, die einem völlig unbedarften Beurteiler ohnehin zur Verfügung stehen.

4 Methode

4.1 Äußere Bedingungen

Die vorliegende Untersuchung wurde in der Zeit vom 26.10.1999 bis zum 25.02.2000 im Psychologischen Institut der Universität Mainz durchgeführt. Die Rekrutierung der ausschließlich weiblichen Versuchsteilnehmer (s. Abschnitt 4.2) erfolgte über Plakate, die in der Stadt Mainz, insbesondere auf dem Universitätsgelände plaziert wurden, sowie über direkte Kontaktaufnahme im Rahmen von Lehr- und Informationsveranstaltungen des Psychologischen Instituts. Die Dauer eines einzelnen Experiments mit anschließender Aufklärung der jeweiligen Probandin betrug zwischen zweieinhalb und drei Stunden. Keine der Teilnehmerinnen begann ihr Experiment vor 8:00 Uhr bzw. nach 19:00 Uhr.

4.2 Stichprobe

An dem Experiment nahmen insgesamt 108 Probandinnen (Pbn) teil. Es wurden ausschließlich **weibliche Teilnehmer** rekrutiert, um etwaige Geschlechtseffekte konstantzuhalten. Fünf Datensätze gingen nicht in die Auswertung ein, weil es sich bei den Pbn um elektrodermale „Nonresponder“ handelte, d.h. um Personen, die keine reizbezogenen Hautleitfähigkeitsveränderungen zeigten und demzufolge nicht sinnvoll mit dem *GAT* untersucht werden konnten. Unter den „Nonrespondern“ hatte eine Probandin (Pb) zudem erhebliche Sprachschwierigkeiten. Eine weitere Pb wurde nicht für die Auswertung berücksichtigt, weil der Versuchsablauf durch eine erhebliche Sehschwäche beeinträchtigt wurde.

Somit verblieb eine Stichprobe von **N = 102** Pbn, deren Alter zwischen 17 und 67 Jahren lag und durchschnittlich ca. 26 Jahre betrug (Standardabweichung: 9 Jahre; Median: 23 Jahre). Unter diesen Pbn befanden sich 79 Studentinnen. Davon waren 27 Psychologie-Studentinnen (ausschließlich aus dem Vordiplom-Abschnitt). Die übrigen studentischen Pbn entstammten den verschiedensten Fachrichtungen, wobei am zahlreichsten die Fächer Medizin (8 Pbn), Publizistik (6), Politik (5), Fremdsprachen (4), Theaterwissenschaften (3), Filmwissenschaften (3) und Pädagogik (3) vertreten waren. Bei den 23 nicht-studentischen Pbn handelte es sich um Oberstufen-Schülerinnen (9) sowie um Angehörige verschiedenster Berufsgruppen. Eine Pb war Rentnerin. Bei 57 Pbn lag nach eigenen Angaben eine diagnostizierte Sehschwäche vor. Davon trugen 39 zum Zeitpunkt der Untersuchung eine Brille oder Kontaktlinsen, während 18 Pbn angaben, keine Sehhilfen zu nutzen und sich dadurch auch nicht beeinträchtigt zu fühlen. Keiner

Pb bereitete das Lesen der schriftlichen Instruktionen oder das Erkennen der dargebotenen Reize Schwierigkeiten. Die Muttersprache war bei 93 Pbn Deutsch. Auch die neun Pbn mit anderer Muttersprache (Kroatisch, Türkisch [2], Polnisch, Portugiesisch, Griechisch, Russisch, Jugoslawisch und Spanisch) sprachen fließend deutsch, und das Lesen und Verstehen der Instruktionen sowie der *GAT*-Fragen und -Items bereitete ihnen keine Probleme. Fünfundvierzig Pbn berichteten, früher schon einmal an psychologischen Experimenten teilgenommen zu haben. Zwei davon hatten nach eigenen Angaben auch schon einmal an einer Untersuchung zur psychophysiologischen Glaubhaftigkeitsbeurteilung partizipiert. Dreiundzwanzig Pbn gaben an, aus den Medien vage Vorkenntnisse zur forensischen, insbesondere zur psychophysiologischen Glaubhaftigkeitsbeurteilung zu besitzen. Zehn Pbn teilten mit, sie seien vor der Untersuchung schon einmal Zeugin eines Diebstahls gewesen.

4.3 Versuchsplan

Die Grundzüge des vorliegenden Versuchsplans wurden bereits in den Abschnitten 3.2.3 und 3.2.4 skizziert. Im folgenden wird genauer auf die Details des Versuchsplans eingegangen.

4.3.1 Unabhängige und abhängige Variablen

Der vorliegenden Untersuchung liegt ein zweifaktorielles Versuchsdesign mit dem Gruppenfaktor **Status der aussagenden Person** und dem Meßwiederholungsfaktor **Methode der Glaubhaftigkeitsbeurteilung** zugrunde (vgl. Abschnitt 3.2.4, Tabelle 11).

Der **Status der aussagenden Person** bzw. die Glaubhaftigkeit der Aussagen wurde variiert, indem eine Gruppe von Pbn im Rahmen einer Verbrechenssimulation einen Gelddiebstahl beging (***Täterinnen***), während eine zweite Gruppe von Pbn (***Zeuginnen***) das Delikt beobachtete. Eine dritte Gruppe (***falsche Zeuginnen***) wurde gar nicht in das inszenierte Delikt verwickelt und konnte auch keine Kenntnis der kritischen Tatdetails erlangen, die im *GAT* als relevante Items verwendet wurden. Alle drei Gruppen sollten im Rahmen einer „Vernehmung“ eine Zeugenaussage zu dem Diebstahl machen und einen *GAT* zur Überprüfung der Täterschaft absolvieren. Dementsprechend machten die *Täterinnen* nur unglaubhafte Angaben, d.h. sie stritten die Täterschaft wahrheitswidrig ab und stellten sich in einer erfundenen „Zeugenaussage“ als Zeugin dar bzw. belasteten die eigentliche *Zeugin* als Täterin. Die *Zeuginnen* machten nur glaubhafte Angaben, d.h.

sie legten eine erlebnisbasierende Aussage über den Tathergang ab und stritten die Täterschaft wahrheitsgemäß ab. Die *falschen Zeuginnen* erfanden einerseits eine Zeugenaussage, in welcher sie eine fiktive Person des Diebstahls bezichtigten; sie stritten jedoch andererseits die Täterschaft wahrheitsgemäß ab. Die Glaubhaftigkeit der vermeintlichen Zeugenaussagen analysierte man jeweils inhaltsanalytisch. Die Glaubhaftigkeit der Täterschaftsabstreitung wurde jeweils psychophysiologisch, mit dem *GAT*, untersucht. Die Zuteilung der Pbn zu den drei experimentellen Gruppen erfolgte per Randomisierung.

Das **Scheinverbrechenszenario** war so angelegt, daß sich jeweils eine *Täterin* und eine *Zeugin* mehrere Minuten gemeinsam am Tatort aufhielten, daß umfangreiche Handlungsketten auftraten und daß es zu komplexen (verbalen) Interaktionen zwischen den *Täterinnen* und *Zeuginnen* kommen konnte. Dadurch wurde gewährleistet, daß das Geschehen genügend Stoff für eine inhaltsanalytisch auswertbare Zeugenaussage bot. Die *Zeuginnen* beobachteten nicht nur den Diebstahl, sondern führten vor, während und nach der Beobachtung des Delikts auch noch eine neutrale Aufgabe (Aufräumarbeiten am Tatort) aus, die sie später als Alibi schildern konnten. Andererseits konnten jedoch die *Täterinnen* die Aufräumarbeiten der *Zeuginnen* nur bruchstückhaft wahrnehmen, weil sie instruktionsgemäß den Tatort später betraten und auch früher wieder verließen als die *Zeuginnen* und sich zudem auf die Suche der Diebesbeute konzentrieren mußten (s. Abschnitt 4.3.3). Dadurch war gewährleistet, daß die *Täterinnen* sich in der späteren „Vernehmung“ nicht die neutrale Aufgabe der *Zeuginnen* als Alibi zueigen machen konnten. Darüber hinaus war das Scheinverbrechenszenario so konzipiert, daß die *Täterinnen* und *Zeuginnen* gleichermaßen mit denjenigen kritischen Tatdetails in Berührung kamen, die später im *GAT* als relevante Items verwendet wurden.

In Abschnitt 4.3.3 erfolgt eine detaillierte Beschreibung des Versuchsablaufs bzw. Scheinverbrechenszenarios. An dieser Stelle sei allerdings eine **Modifikation** gegenüber dem in Abschnitt 3.2.3 beschriebenen experimentellen Grunddesign vorweggenommen. Anstelle zweier naiver Pbn nahm jeweils nur eine naive Pb an der Verbrechenssimulation teil. Dieser wurde – ohne ihr Wissen – eine in den Versuchsplan eingeweihte experimentelle **„Strohfrau“** zur Seite gestellt, die je nach Gruppenzugehörigkeit der naiven Pbn (*Täterinnen* oder *Zeuginnen*) die komplementäre Rolle übernahm. Der Einsatz einer eingeweihten „Strohfrau“ anstelle der zweiten naiven Pb hatte in erster Linie organisatorische bzw. durchführungstechnische Gründe. So hätte eine experimentelle Simulation mit jeweils zwei naiven Pbn es erforderlich gemacht, immer zwei Personen gleichzeitig zum Experiment einzuladen. Dies hätte zum einen die Terminabsprachen erschwert. Zum anderen hätte es vermutlich die Ausfallquote erhöht. Wäre nämlich von den beiden Pbn (*Täterin* oder *Zeugin*) eine nicht zum Experiment erschie-

nen oder hätte die Teilnahme vorzeitig abgebrochen, so hätte automatisch auch die verbleibende zweite Pb die Untersuchung nicht absolvieren können. Darüber hinaus wären bei der gleichzeitigen Teilnahme zweier naiver Pbn auch zwei äquivalente Labors und zwei parallel agierende Versuchsleiter erforderlich gewesen, um beide Pbn nach der Verbrechenssimulation parallel befragen bzw. psychophysiologisch untersuchen zu können. Schließlich brachte der Einsatz der „Strohfrau“ auch noch den Vorteil mit sich, daß der Ablauf der Verbrechenssimulation besser kontrollierbar war. Besonderheiten des Versuchsablaufs konnten jeweils im Anschluß an die Verbrechenssimulation von der „Strohfrau“ protokolliert werden. Diese wichtigen Informationen wären im Falle einer ausschließlichen Teilnahme naiver Pbn nicht zugänglich gewesen.

Die **Methode der Glaubhaftigkeitsbeurteilung** wurde intraindividuell variiert, indem bei allen Pbn eine **inhaltsorientierte** Glaubhaftigkeitsbeurteilung, eine psychophysiologische Glaubhaftigkeitsbeurteilung mit dem ***GAT*** sowie eine **naive** Glaubhaftigkeitsbeurteilung vorgenommen wurde. Eine genauere Beschreibung der entsprechenden Auswertungsprozeduren erfolgt in Abschnitt 4.4.

Die Reihenfolge der Datenerhebungen wurde konstantgehalten, d.h. mit allen Pbn wurde zuerst das halbstandardisierte Interview zur Erhebung der Zeugenaussage und anschließend der *GAT* durchgeführt. Dies war erforderlich, weil bei den *falschen Zeuginnen* eine umgekehrte Reihenfolge nicht mit der Logik der in den Instruktionen vermittelten experimentellen Hintergrundgeschichte (s. Abschnitt 4.3.3) vereinbar gewesen wäre.

Die **abhängige Variable** der vorliegenden Untersuchung war die **Treffsicherheit der Glaubhaftigkeitsbeurteilung**.

4.3.2 Zusätzlich kontrollierte Variablen

Neben den Zeugenaussagen und den physiologischen Messungen im *GAT* wurden auch noch weitere Variablen erhoben, die als potentielle konfundierende Variablen kontrolliert werden sollten.

Es gibt Forschungsergebnisse, die darauf hindeuten, daß die Treffsicherheit der psychophysiologischen Glaubhaftigkeitsbeurteilung mit der **elektrodermalen Labilität** der Probanden zusammenhängt. Die elektrodermale Labilität ist als ein relativ stabiles psychophysiologisches Personenmerkmal definiert (z.B. Vossel, 1990; Vossel & Zimmer, 1990). Die Operationalisierung erfolgt i.d.R. über die Häufigkeit reizunabhängiger, pha-

sischer Hautleitfähigkeitsveränderungen, sog. Spontanfluktuationen (NSRs; „non-specific responses“, Venables & Christie, 1980), während einer reizfreien Ruhephase (vgl. auch Boucsein, 1988). Als NSRs gelten phasische Veränderungen der Hautleitfähigkeit mit einer Amplitude von mindestens 0.02 µS (z.B. Siddle, O'Gorman & Wood, 1979; Vossel & Rossmann, 1984; Waid, Wilson & Orne, 1981; Zimmer, Vossel & Fröhlich, 1990). Personen, die eine relativ hohe Anzahl von NSRs pro Zeiteinheit aufweisen, bezeichnet man als „elektrodermal labil“, Personen mit einer geringen Häufigkeit von NSRs hingegen als „elektrodermal stabil“ (vgl. Vossel, 1990, S. 29f.). Die Ergebnisse einiger Analogstudien zum *KFT* und *TWT* deuten darauf hin, daß elektrodermal labile schuldige Pbn stärkere quantitative Reaktionsdifferenzen zwischen relevanten und Kontrollfragen bzw. zwischen relevanten und irrelevanten Items zeigen als elektrodermal stabile schuldige Pbn (z.B. Waid & Orne, 1980; Horneman & O'Gorman, 1987). Im Hinblick auf die Treffsicherheit der psychophysiologischen Glaubhaftigkeitsbeurteilung ergibt sich hieraus für elektrodermal stabile Schuldige im Vergleich zu elektrodermal labilen Schuldigen ein erhöhtes Risiko für irrtümlich negative Klassifikationen („unschuldig“). Dagegen wurden elektrodermal labile Unschuldige im Vergleich zu elektrodermal stabilen Unschuldigen häufiger als „schuldig“ fehlklassifiziert (falsch positive Befunde).

Darüber hinaus ist die Kontrolle der elektrodermalen Labilität grundsätzlich als sinnvoll anzusehen, sofern SCRs als abhängige Variable erfaßt werden. Elektrodermal labile und stabile Personen unterscheiden sich hinsichtlich der Stärke ihrer SCRs auf gleiche Reize und in bezug auf die Habituationsverläufe bei wiederholter Reizdarbietung. In zahlreichen Untersuchungen wurden Zusammenhänge zwischen der Habituationsgeschwindigkeit der SCRs und der Häufigkeit von NSRs dokumentiert (z.B. Coles, Gale & Kline, 1971; Crider & Lunn, 1971; Dickinson & Smith, 1973; Katkin & McCubbin, 1969; Schell, Dawson & Filion, 1988). Insbesondere bei bedeutungsvollen Stimuli habituieren elektrodermal Labile langsamer als Stabile (Vossel, 1990). Nach Fahrenberg (1969) kann die Nichtbeachtung derartiger interindividueller Differenzen im Rahmen psychophysiologischer Untersuchungen eine Verfälschung der experimentellen Befunde nach sich ziehen.

Da das Ergebnis *des GAT* davon abhängt, daß sich tatbeteiligte Pbn auch tatsächlich an die relevanten Items erinnern können, wurden mit allen Pbn am Ende der Untersuchung zwei Gedächtnistests durchgeführt. Zunächst wurde das **freie Erinnern der kritischen Tatdetails** überprüft. Dann wurde ein Multiple-Choice-Test appliziert, um die **Wiedererkennensleistung** zu testen.

Die Ergebnisse einiger Untersuchungen deuten darauf hin, daß die Treffsicherheit der psychophysiologischen Glaubhaftigkeitsbeurteilung positiv mit der **Motivation der Probanden glaubwürdig zu erscheinen** sowie deren **Überzeugung von der Zuverlässigkeit der Methode** korreliert ist (vgl. zusammenfassend Ben-Shakhar & Furedy, 1990; Steller, 1987). Um die Einflüsse dieser beiden Variablen zu kontrollieren, sollten die Pbn am Ende der Untersuchung sowohl retrospektiv einschätzen, wie sehr sie beim *GAT* motiviert gewesen waren, einen unschuldigen Eindruck zu hinterlassen, als auch angeben, für wie wahrscheinlich sie es hielten, daß ihnen dies gelungen sei. Es liegen zwar keine empirischen Erkenntnisse zur Rolle der beiden genannten Variablen im Rahmen der inhaltsorientierten Glaubhaftigkeitsbeurteilung vor. Gleichwohl ist es denkbar, daß auch die inhaltliche Qualität von Aussagen mit der Höhe der Motivation der aussagenden Person, glaubwürdig zu erscheinen, sowie deren Annahmen über die Treffsicherheit der inhaltsorientierten Glaubhaftigkeitsbeurteilung zusammenhängt. Daher sollten die Pbn am Ende der Untersuchung auch retrospektiv die Höhe ihrer Motivation einschätzen, bei der Abgabe der Zeugenaussage glaubhaft zu erscheinen, und angeben, für wie wahrscheinlich sie es hielten, daß ihnen dies gelungen sei.

Sowohl psychophysiologische als auch inhaltsorientierte Glaubhaftigkeitsbeurteilungen sind anfällig für **Manipulationstechniken** seitens der Probanden (vgl. Abschnitt 2.2.4 bzw. 2.1.5). Um etwaige gezielte Manipulationsversuche zu kontrollieren, wurden die Pbn am Ende des Experiments bezüglich der Anwendung entsprechender Strategien, Taktiken oder Techniken befragt. Außerdem befragte man die Pbn im Verlauf der Untersuchung nach etwaigen **Vorkenntnissen auf dem Gebiet der forensischen Glaubhaftigkeitsbeurteilung** sowie nach früheren **Teilnahmen an entsprechenden Untersuchungen** (s. auch Abschnitt 4.2).

Die inhaltliche Qualität einer Aussage hängt neben dem Erlebnisbezug auch von der **intellektuellen Begabung** der aussagenden Person ab. Um den Einfluß dieser Variablen zu kontrollieren, bearbeiteten alle Pbn am Ende der Untersuchung einen Test zur Erfassung der allgemeinen Intelligenz. Da die Pbn angesichts der langen Dauer des Experiments nicht noch zusätzlich stark belastet werden sollten, handelte es sich bei dem Test allerdings nur um ein kurzes Screening-Verfahren, das normalerweise nicht für individualdiagnostische Zwecke geeignet ist. Hiermit war es jedoch immerhin möglich, etwaige Pbn zu identifizieren, deren Intelligenz stark vom Durchschnitt abwich.

Die inhaltliche Qualität einer Zeugenaussage zu einem simulierten Diebstahl hängt möglicherweise auch damit zusammen, ob die aussagende Person früher schon einmal **reale Erfahrungen mit einem Diebstahl** (als Zeuge oder auch als Täter) gemacht hat. Insbesondere ist es denkbar, daß eine erfundene Zeugenaussage im Rahmen einer expe-

rimentellen Simulation eine höhere inhaltliche Qualität aufweist, wenn die Pb auf entsprechende autobiographische Erfahrungen rekurrieren kann. Daher wurden die Pbn auch nach persönlichen früheren Erfahrungen mit Diebstählen befragt.

4.3.3 Versuchsablauf

Tabelle 12 bietet einen Überblick über den komplexen Versuchsablauf. Das Experiment wurde in Form von **Einzeluntersuchungen** durchgeführt, d.h. es nahm jeweils nur eine Pb am Experiment teil. An der Durchführung der Untersuchung waren neben der jeweiligen Pb zwei Versuchsleiter (Vl) und eine „Strohfrau“ beteiligt. Versuchsleiter 1 (männlich) war zuständig für die Begrüßung der Pbn, die Instruierung hinsichtlich des Ablaufs der experimentellen Simulation, die Nachbefragung der Pbn im Anschluß an die Simulationsphase sowie die abschließende Aufklärung, Entlohnung und Verabschiedung der Pbn. Versuchsleiter 2 (weiblich) führte jeweils im Anschluß an die Verbrechenssimulation mit den Pbn das Interview zur Erhebung der Zeugenaussage und die psychophysiologischen Untersuchungen durch. Die „Strohfrau“ spielte in der Verbrechenssimulation je nach der experimentellen Gruppenzugehörigkeit der jeweiligen Pb die komplementäre Rolle der *Täterin* bzw. *Zeugin* (und kam nicht zum Einsatz, wenn die Pb der Bedingung *falsche Zeuginnen* angehörte).

Tabelle 12. Ablauf des Experiments

	Phasen	**Dauer**	**Raum**	**Vl**
I.	Instruktionen zum Ablauf der Verbrechenssimulation	15 min	02-131	Vl 1
II.	Verbrechenssimulation: „Tatbegehung“ bzw. „Tatbeobachtung“	15 min	02-527	„Strohfrau“
III.	Allgemeine Instruktionen zum Ablauf der „Vernehmung“	10 min	02-131	Vl 1
IV.	Vorbereitung auf „Vernehmung“	15 min	02-527	
V.	„Vernehmung durch gerichtspsychologische Expertin“		02-524	Vl 2
	• Halbstandardisiertes Interview (Zeugenaussage)	15 min		
	• Psychophysiologische Untersuchungen	70 min		
	- Ruhemessung			
	- „Stimulationstest“			
	- *GAT*			
VI.	Nachbefragung	15 min	02-131	Vl 1

Anmerkung: Die Gruppen *Täterinnen* und *Zeuginnen* durchliefen alle Phasen. Für die Gruppe *falsche Zeuginnen* stellte Phase III den Beginn des Versuchsablaufs dar.

Phase I: Instruktionen zum Ablauf der Verbrechenssimulation (vgl. Tabelle 12)
Bei der Ankunft im Psychologischen Institut wurden die Pbn von Vl 1 außerhalb des Laborbereichs begrüßt. Dieser führte die Pbn in sein Büro (Raum 02-131) und gab ih-

nen dort die Instruktionen für den Ablauf der Verbrechenssimulation. Sämtliche Anweisungen erfolgten schriftlich. (Die schriftlichen Instruktionen sind im Anhang A nachzulesen.) Falls die Pbn nach dem Lesen einer Instruktion noch Fragen hatten, wurden diese von Vl 1 beantwortet.

Zunächst wurden die Pbn über die allgemeinen Bedingungen für die Teilnahme an der Untersuchung informiert (Datenschutz, Höhe der Teilnahmevergütung, Verschwiegenheit bis zum Abschluß der Untersuchungsreihe etc.; vgl. Anhang A.1). Hatten die Pbn nichts gegen die Konditionen einzuwenden, so bekundeten sie per Unterschrift ihr Einverständnis mit der Teilnahme am Experiment. Nachdem die Pbn in die Teilnahme eingewilligt hatten, wurden sie über ihre Aufgabe im Rahmen der experimentellen Simulation informiert. Dementsprechend unterschieden sich von nun an die Instruktionen je nachdem, welcher der drei experimentellen Gruppen die Pbn zugeteilt worden waren.

Die Pbn der Gruppe *Täterinnen* bekamen den Auftrag, einen Gelddiebstahl zu verüben. Sie wurden vorab darüber in Kenntnis gesetzt, daß sie dabei von einer anderen „Pb“ (bei der es sich in Wirklichkeit um eine „Strohfrau“ handelte, s. Abschnitt 4.3.1) beobachtet würden (vgl. Anhang A.2 und A.3). Die Pbn der Gruppe *Zeuginnen* wurden instruiert, am Tatort Aufräum- bzw. Reinigungsarbeiten durchzuführen und dabei zu beobachten, wie eine andere „Pb“ (in Wirklichkeit eine „Strohfrau“) einen Diebstahl begehe (vgl. Anhang A.6 und A.7).

Sowohl die *Täterinnen* als auch die *Zeuginnen* wurden bereits vor der Begehung bzw. Beobachtung des Diebstahls darüber informiert, daß anschließend beide beteiligte Pbn (sowohl die *Täterin* als auch die *Zeugin*) der Täterschaft verdächtigt würden und daß beide sich einer Untersuchung durch eine „gerichtspsychologische Expertin“ zu unterziehen hätten, durch welche die wahre Diebin ermittelt werden solle. Die Untersuchung durch die „gerichtspsychologische Expertin“ bestehe aus einer „mündlichen Vernehmung“ und einem „Lügendetektortest“. Sowohl die *Täterinnen* als auch die *Zeuginnen* müßten der „gerichtspsychologischen Expertin“ gegenüber versuchen, sich selbst zu entlasten und die andere beteiligte Person als Täterin zu belasten. Den *Täterinnen* wie auch den *Zeuginnen* wurde schon vor der Begehung bzw. Beobachtung des Diebstahls eine Belohnung von 15,- DM in Aussicht gestellt, falls es ihnen später gelingen würde, die „gerichtspsychologische Expertin“ von der eigenen Unschuld bzw. von der Täterschaft der anderen beteiligten „Pb“ zu überzeugen (vgl. Anhang A.2 und A.6).

Die Simulation des Diebstahls bzw. seiner Beobachtung wurde in folgende Hintergrundgeschichten eingebettet: Den Pbn der Gruppe *Täterinnen* wurde mitgeteilt, das zu stehlende Geld befinde sich in einer Geldkassette, welche irgendwo im Büro von „Prof.

Kunze“[16] (Raum 02-527) versteckt sei. Das Schloß der Geldkassette sei nur mit Hilfe einer zehnstelligen Zahlenkombination zu öffnen. Da „Prof. Kunze“ sich Zahlen nicht gut merken könne, habe er die einzelnen Ziffern der Zahlenkombination an unterschiedlichen Stellen seines Büros schriftlich niedergelegt. Die Pb (*Täterin*) habe nun durch Zufall erfahren, wo sich der Notizzettel mit der ersten Ziffer der Zahlenkombination befinde (nämlich in der linken oberen Schublade des großen Schreibtischs, auf welcher ein Deutschlandaufkleber angebracht sei). Dieser Notizzettel enthalte außerdem auch Angaben über den Ort, an dem der Notizzettel mit der zweiten Ziffer der Zahlenkombination versteckt sei. Der Notizzettel mit der zweiten Ziffer der Zahlenkombination wiederum enthalte außerdem auch Angaben über den Ort, an dem sich der Notizzettel mit der dritten Ziffer der Zahlenkombination befinde u.s.w.. Auf dem Zettel mit der letzten Ziffer der zehnstelligen Zahlenkombination sei schließlich auch noch vermerkt, wo die Geldkassette verstaut sei.

Entsprechend wurden die Pbn der Gruppe *Täterinnen* aufgefordert, in das Büro von „Prof. Kunze“ zu gehen, dort die zehn Ziffern der Zahlenkombination und das Versteck der Geldkassette ausfindig zu machen, die Geldkassette zu öffnen, das ganze Geld herauszunehmen und anschließend damit in das Büro von Vl 1 zurückzukehren.[17] Ferner wurden die *Täterinnen* angewiesen, sich jeweils der anderen im Büro von „Prof. Kunze“ anwesenden „Pb“ gegenüber möglichst unauffällig zu benehmen. Falls sie von dieser „Pb“ auf ihr Tun hin angesprochen würden, sollten sie sich möglichst plausible Ausreden einfallen lassen (vgl. Anhang A.3).

Die Pbn der Gruppe *Zeuginnen* sollten sich in die Situation hineinversetzen, sie arbeiteten als Reinigungskraft im Psychologischen Institut und hätten die Aufgabe, das Büro von „Prof. Kunze“ aufzuräumen bzw. zu reinigen. Dementsprechend wurden sie aufgefordert, sich in das Büro von „Prof. Kunze“ zu begeben. Dort würden sie auf einem Schreibtisch einen Zettel vorfinden, auf welchem „Prof. Kunze“ zehn kleinere Aufräum- bzw. Reinigungsarbeiten für sie aufgelistet habe. Die Pbn sollten mit der Verrichtung dieser Arbeiten beginnen. Nach einer gewissen Zeit werde eine weitere weibliche Person das Büro betreten, um etwas ganz Bestimmtes zu suchen und zu entwenden. Die *Zeuginnen* sollten das Verhalten der anderen Person ganz genau beobachten und versuchen, diese in ein Gespräch zu verwickeln, um herauszufinden, was sie stehlen wolle. Wenn die Diebin das Büro von „Prof. Kunze“ wieder verlassen habe, sollten die *Zeuginnen* die Aufräum- bzw. Reinigungsarbeiten abschließen und danach in das Büro von Vl 1 zurückkehren (vgl. Anhang A.7).

---

[16] „Prof. Kunze“ ist eine fiktive Person. Dies wurde den Pbn jedoch erst nach dem Experiment mitgeteilt.

[17] Zur räumlichen Orientierung wurde den *Täterinnen* wie auch den *Zeuginnen* eine Wegskizze ausgehändigt (s. Anhang A.12).

Phase II: Verbrechenssimulation: „Tatbegehung“ bzw. „Tatbeobachtung“ (vgl. Tabelle 12)

Die eigentliche Simulationsphase hatte konkret folgenden Ablauf: Zuerst betrat die *Zeugin* das Büro von „Prof. Kunze“ (Raum 02-527). Dann las sie den von „Prof. Kunze“ hinterlegten Aufgabenzettel (vgl. Anhang A.13) und begann, der Reihe nach die zehn Aufräum- bzw. Reinigungsarbeiten zu verrichten: (1) Sie zog den Griff der Schreibtischschublade, auf der sich der *Deutschland-Aufkleber* befand, fest. (2) Sie goß den *Kaktus*. (3) Sie entstaubte das *Bild mit Kühen*. (4) Sie wischte den *Porzellanhund* ab. (5) Sie nahm die Pfeile von der *Dartscheibe* ab. (6) Sie hängte die schwarze *Jacke* im Schrank auf den Bügel. (7) Sie saugte den *roten Teppich* mit einem Staubsauger. (8) Sie stellte das *gelbe Fahrrad* auf den Fahrradständer. (9) Sie räumte den *Getränkekasten mit Wasserflaschen* voll. (10) Sie legte die *Äpfel* zurück in die Obstschale.

Sobald die *Zeugin* mit dem Teppichsaugen (Aufgabe 7, s.o.) begonnen hatte, betrat die *Täterin* das Büro von „Prof. Kunze“[18]. Sie begann dort mit der Suche nach den zehn Ziffern der Zahlenkombination für das Schloß der Geldkassette. Den Notizzettel[19] mit der ersten Ziffer fand sie, wie in der Instruktion beschrieben, in der Schreibtischschublade mit dem *Deutschland-Aufkleber*. Den Notizzettel mit der zweiten Ziffer der Zahlenkombination fand sie unter dem *Kaktus*. Den Notizzettel mit der dritten Ziffer fand sie unter dem *Porzellanhund*, den mit der vierten Ziffer fand sie auf der Satteltasche des *gelben Fahrrads*, den mit der fünften Ziffer hinter dem *Bild mit Kühen*, den mit der sechsten Ziffer unter dem *Getränkekasten mit Wasserflaschen*, den mit der siebten Ziffer in der rechten Tasche der *Jacke* im Holzschrank, den mit der achten Ziffer unter der Obstschale mit *Äpfeln*, den mit der neunten Ziffer hinter der *Dartscheibe* an der Wand, den Notizzettel mit der zehnten Ziffer der Zahlenkombination fand sie schließlich unter dem *roten Teppich*. Die *Täterin* notierte sich jeweils die entsprechende Ziffer der Zahlenkombination und den Ort, an dem der Notizzettel mit der nächsten Ziffer der Zahlenkombination zu finden sei.[20] Der Notizzettel mit der letzten Ziffer der Zahlenkombination verriet der *Täterin* schließlich auch den Ort der Geldkassette. Diese war in einem roten Karton im untersten Regalfach deponiert. Die *Täterin* ging also zum Versteck der

[18] Die zeitliche Abstimmung für den Auftritt der *Täterin* erfolgte so: Wie bereits erwähnt, wurde die komplementäre Rolle zur Rolle der naiven Pb jeweils von einer „Strohfrau“ gespielt. Handelte es sich bei der naiven Pb um eine *Zeugin*, so wartete die „Strohfrau“ im Nebenraum des Büros von „Prof. Kunze“, bis das gut hörbare Geräusch des Staubsaugers ertönte. Das Starten des Staubsaugers diente der „Strohfrau“ als Signal, jetzt in das Büro von „Prof. Kunze“ zu gehen und die Rolle der *Täterin* zu spielen. Handelte es sich bei der naiven Pb um eine *Täterin*, so saugte die „Strohfrau“ im Büro von „Prof. Kunze“ so lange den Teppich, bis die naive Pb den Raum betrat, und fuhr anschließend fort, die Rolle der *Zeugin* zu spielen.

[19] Die zehn Notizzettel mit den Ziffern der Zahlenkombination sind zur Veranschaulichung im Anhang abgedruckt (s. Anhang A.15).

[20] Zum Aufschreiben der Ziffern für die Zahlenkombination sowie der Orte, an denen die Notizzettel von „Prof. Kunze“ und die Geldkassette versteckt waren, wurde den *Täterinnen* von Vl 1 ein speziell vorbereitetes Formular (s. Anhang A.16) und ein Kugelschreiber mitgegeben.

Geldkassette, öffnete diese[21], nahm das Geld (100 DM-Schein) heraus, verließ anschließend den Tatort und kehrte in das Büro von Vl 1 zurück.

Die *Zeugin* beobachtete genau das Verhalten der *Täterin*. Sie versuchte außerdem, die *Täterin* in ein Gespräch zu verwickeln, um herauszufinden, was sie stehlen wolle. Die *Täterin* versuchte, sich auf die Fragen der *Zeugin* möglichst plausible Ausreden einfallen zu lassen. Nachdem die *Täterin* den Raum verlassen hatte, führte die *Zeugin* noch die restlichen Aufräum- bzw. Reinigungsarbeiten aus. Dann verließ sie den Tatort und ging zurück in das Büro von Vl 1. Die *Täterinnen* und *Zeuginnen* kamen also mit denselben zehn Details des Tatorts (oben jeweils durch Hervorhebung gekennzeichnet) in Kontakt. Diese zehn Tatortdetails wurden später im *GAT* als relevante Antwortalternativen verwendet. Die Ansicht des Tatorts und die zehn kritischen Tatortdetails sind im Anhang B.1 auf den Abbildungen B.1.1 bis B.1.11 dargestellt.

Phase III: Allgemeine Instruktionen zum Ablauf der „Vernehmung" (s. Tabelle 12)

Nach der Begehung bzw. Beobachtung des Diebstahls erhielten die *Täterinnen* und *Zeuginnen* von Vl 1 weitere Instruktionen. Ihnen wurde mitgeteilt, daß beide zur Tatzeit am Tatort anwesenden Personen, also sowohl die wahre *Täterin* als auch die *Zeugin*, der Täterschaft verdächtigt würden und daß eine „gerichtspsychologische Expertin" damit beauftragt sei herauszufinden, welche von beiden Verdächtigen das Geld gestohlen habe. Hierzu werde die „Expertin" mit beiden Tatverdächtigen eine „mündliche Vernehmung" und einen „Lügendetektortest" durchführen. Sowohl den *Täterinnen* als auch den *Zeuginnen* wurde eine Belohnung von 15,- DM versprochen, falls es ihnen gelinge, die „Expertin" von der eigenen Unschuld zu überzeugen bzw. davon, daß die andere verdächtige Person das Geld gestohlen habe. Dies sei nur möglich, wenn die Pbn sowohl in der „mündlichen Vernehmung" einen glaubwürdigen Eindruck hinterlassen als auch den „Lügendetektortest" bestehen würden (vgl. Anhang A.4 und A.8).

In einer weiteren Instruktion wurden die Pbn auf die „mündliche Vernehmung" durch die „gerichtspsychologische Expertin" vorbereitet (vgl. Anhang A.5 und A.9). Sowohl die *Täterinnen* als auch die *Zeuginnen* wurden darauf hingewiesen, daß es nicht ausreiche, einfach nur zu behaupten, die andere Person habe das Geld gestohlen. Vielmehr

[21] Es sei erwähnt, daß die Geldkassette gar nicht mit einem Zahlenschloß verriegelt war. Dadurch wurde sichergestellt, daß alle *Täterinnen* die Kassette öffnen und das Geld entnehmen konnten. Das Öffnen von Zahlenschlössern stellt nämlich mitunter hohe Anforderungen an die Feinmotorik bzw. setzt Übung mit der Schloßmechanik voraus. Dies hätte möglicherweise dazu geführt, daß manche Pbn nicht in der Lage gewesen wären die Kassette zu öffnen und deshalb das Experiment hätten abbrechen müssen. Da sich das Fehlen des Zahlenschlosses für die *Täterinnen* jedoch erst am Ende der Verbrechenssimulation herausstellte bzw. nachdem sie mit allen kritischen Tatortdetails in Berührung gekommen waren, war es der Logik des Versuchsplans bzw. -ablaufs nicht abträglich.

komme es darauf an, eine umfassende Beschreibung des gesamten Tathergangs abzugeben (*Zeuginnen*) bzw. sich eine möglichst umfangreiche und zugleich plausibel erscheinende Geschichte über den Diebstahl auszudenken und bei deren Schilderung möglichst glaubwürdig zu erscheinen (*Täterinnen*). Den *Täterinnen* und *Zeuginnen* wurde ferner mitgeteilt, daß die „Gerichtspsychologin" zu Beginn der „Vernehmung" von ihnen einen zusammenhängenden Bericht über den gesamten Tathergang verlangen werde, welchem besondere Bedeutung bei der Beurteilung ihrer Glaubwürdigkeit zukomme.

Auch die Pbn in der Bedingung *falsche Zeuginnen* wurden instruiert, eine „Zeugenaussage" zu dem Gelddiebstahl zu machen. Allerdings hatten sie vorher den Diebstahl weder selbst begangen noch als *Zeuginnen* beobachtet. Den *falschen Zeuginnen* wurden lediglich die wesentlichen Eckdaten zu dem Gelddiebstahl mitgeteilt (vgl. Anhang A.10 und A.11): Im Psychologischen Institut seien am gleichen Tag 100,- DM aus dem Büro von „Prof. Kunze" gestohlen worden. Es sei bekannt, daß sich während des Diebstahls neben dem Täter bzw. der Täterin noch eine weitere Person am Tatort aufgehalten habe, die möglicherweise das Delikt beobachtet habe und möglicherweise auch mit dem Dieb bzw. der Diebin in Kontakt getreten sei. Für Hinweise, die zur Aufklärung des Diebstahls führten, sei eine Belohnung von 15,- DM ausgesetzt. Die Pbn der Bedingung *falsche Zeuginnen* sollten nun versuchen, die ausgesetzte Belohnung zu erhalten, indem sie sich als *Zeuginnen* ausgäben und eine entsprechende Zeugenaussage erfänden, in welcher sie eine fiktive Person des Diebstahls beschuldigten. Ob sie die Belohnung schließlich erhalten würden oder nicht, hänge davon ab, ob ihre Aussagen von einer „gerichtspsychologischen Expertin" als glaubwürdig beurteilt würden. Die „Gerichtspsychologin" sei nicht vorab darüber informiert, ob die Pbn den Diebstahl wirklich beobachtet hätten oder nicht. Sie gründe ihr Urteil bezüglich der Glaubwürdigkeit der vermeintlichen Zeugenaussagen auf eine „mündliche Vernehmung" und auf einen „Lügendetektortest", denen sich die Pbn unterziehen müßten. Auch die *falschen Zeuginnen* wurden darauf hingewiesen, daß sie zu Beginn der „mündlichen Vernehmung" einen zusammenhängenden Bericht über den gesamten Tathergang abgeben müßten, der für die Beurteilung ihrer Glaubwürdigkeit von besonderem Gewicht sei.

Phase IV: Vorbereitung auf „Vernehmung" (vgl. Tabelle 12)

Alle Pbn hatten 15 Minuten Zeit, um sich auf ihre Aussagen vorzubereiten. Um die Vorbereitung zu erleichtern, durften sich die *Täterinnen*, *Zeuginnen* und *falschen Zeuginnen* in der Vorbereitungsphase am Tatort (Büro von „Prof. Kunze") aufhalten. Immer wenn sich *falsche Zeuginnen* zur Vorbereitung ihrer Ausssage am Tatort aufhielten, waren zuvor die zehn kritischen Tatortdetails (s.o., Hervorhebung im Text), die im *GAT* als relevante Items dienten, gegen – im Hinblick auf den *GAT* – neutrale Details ausgetauscht worden. Anstelle des *Deutschland-Aufklebers*, des *Kaktus'*, des *Bilds mit Kühen*,

des *Porzellanhunds*, der *Dartscheibe*, der *Jacke*, des *roten Teppichs*, des *gelben Fahrrads*, des *Getränkekastens mit Wasserflaschen* und der *Äpfel* befanden sich an den entsprechenden Stellen des Tatorts nun ein *Schweiz-Aufkleber*, eine *Rose*, ein *Bild mit Fröschen*, ein *Porzellanelefant*, ein *Gymnastikreifen*, ein *Schal*, ein *grauer Teppich*, ein *silbernes Fahrrad*, ein *Kasten Limonade* und eine Obstschale mit *Nüssen*. Durch diesen Kunstgriff war gewährleistet, daß die *falschen Zeuginnen* die relevanten Items des *GAT* nicht kannten bzw. daß es sich diesbezüglich bei ihnen um unschuldige Personen ohne Tatwissen handelte. Die Ansicht des Tatorts und die zehn kritischen Tatortdetails in der Bedingung *falsche Zeuginnen* sind im Anhang B.2 auf den Abbildungen B.2.1 bis B.2.11 dargestellt.

Phase V: „Vernehmung durch gerichtspsychologische Expertin“ (vgl. Tabelle 12)
Der weitere Versuchsablauf war für alle drei Gruppen (*Täterinnen*, *Zeuginnen*, *falsche Zeuginnen*) gleich. Nach Ablauf der Vorbereitungsphase führte Vl 1 die Pbn in den Laborraum, in dem die Befragungen und die psychophysiologischen Messungen durchgeführt wurden (Raum 02-524). Hier stellte Vl 1 den Pbn zunächst die „gerichtspsychologische Expertin“ (Vl 2) vor.[22] Vl 2 führte mit den Pbn zuerst eine „**Vernehmung“ zum Tathergang in Form eines halbstandardisierten Interviews** durch. Zu Beginn des Interviews wurden die Pbn gebeten, eine ausführliche und detaillierte Schilderung des gesamten Tathergangs abzugeben. Nachfragen wurden in diesem Teil nur gestellt, wenn sich Unklarheiten ergaben oder die Pbn Schwierigkeiten hatten, der Aufforderung zum freien Erzählen nachzukommen. Im Anschluss an diesen Teil des Interviews folgten offene Fragen. Diese zielten zum einen auf eine genaue Beschreibung der Täterin und zum anderen auf eine präzise Beschreibung des Tatorts ab. Nach der Beantwortung dieser Fragen wurde den Pbn die Möglichkeit gegeben ihre Aussagen zu ergänzen. Sie wurden explizit darauf hingewiesen, daß sie Zeit hätten, die Vollständigkeit ihrer Einlassungen nochmals zu überdenken. Hatten die Pbn nichts mehr hinzuzufügen, war das Interview beendet. Die Befragungen wurden auf Videokassetten und Tonträgern aufgezeichnet.

Nach der mündlichen „Vernehmung“ führte Vl 2 mit den Pbn die **psychophysiologischen Untersuchungen** durch, die aus einer Ruhemessung, einem „Stimulationstest“ sowie dem eigentlichen *GAT* bestanden (vgl. Tabelle 12). Die entsprechenden Instruktionen erfolgten schriftlich und sind im Anhang A nachzulesen. Zunächst wurden die Pbn grob über die Zielsetzung, die „Logik“ und den Ablauf der psychophysiologischen

[22] Bei Vl 2 handelte es sich in Wirklichkeit nicht um eine professionelle Gerichtspsychologin, sondern um eine Diplomandin, die die Rolle einer gerichtspsychologischen Expertin spielte. Hierüber wurden die Pbn jedoch erst nach dem Experiment informiert. Aus organisatorischen Gründen gab es zwei Personen, die als Vl 2 fungierten. Die Zuteilung dieser beiden Vl zu den Pbn erfolgte nach dem Zufallsprinzip.

Glaubhaftigkeitsbeurteilung informiert. Ihnen wurde mitgeteilt, mit Hilfe des „Lügendetektortests“ solle festgestellt werden, ob sie den Diebstahl begangen hätten oder nicht. Die Funktionsweise des Tests beruhe darauf, daß man beim Lügen aufgeregter sei als beim Wahrheitsagen, daß sich diese Aufregung in bestimmten körperlichen Veränderungen manifestiere und daß man dementsprechend anhand der bei einer Befragung gemessenen körperlichen Reaktionen den Wahrheitsgehalt der Antworten bestimmen könne. Um die Annahmen über die Zuverlässigkeit der „Lügendetektion“ zu manipulieren, wurde den Pbn pauschal mitgeteilt, die Treffsicherheit der Methode sei wissenschaftlich erwiesen. Um überdies die Motivation der Pbn zu steigern, erhielten sie die fingierte Information, daß hochintelligente und zugleich emotional kontrollierte Personen dennoch den Lügendetektor überlisten könnten (s. Anhang A.17).

Nach diesen einleitenden Informationen legte Vl 2 den Pbn die Meßfühler an, führte sie anschließend in die Meßkabine und schloß sie dort an die Meßapparaturen an (vgl. Abschnitt 4.3.4). Anschließend wurden einige soziodemographische Daten der Pbn erfragt (Alter, Beruf etc.). Die Pbn sollten außerdem angeben, inwiefern sie Vorerfahrungen mit psychologischen Experimenten im allgemeinen und mit Untersuchungen zur forensischen Glaubhaftigkeitsbeurteilung im speziellen hätten. Außerdem wurden sie nach Vorkenntnissen auf dem Gebiet der forensischen Glaubhaftigkeitsbeurteilung befragt (vgl. Anhang C.1). Es folgte eine sechsminütige **Ruhemessung**, während der die Pbn mit offenen Augen bequem und entspannt in der Meßkabine sitzen sowie sich möglichst wenig bewegen und nicht sprechen sollten (vgl. Anhang A.18). Die Ruhemessung diente der Erfassung der elektrodermalen Spontanaktivität (vgl. Abschnitt 4.3.2). Hierzu wurden die in den letzten fünf Minuten der Ruhemessung auftretenden NSRs gezählt.

Im Anschluß daran wurde ein fingierter Zahlentest (**„Stimulationstest“**, vgl. Abschnitt 2.2.1.2 und 2.2.2.1) durchgeführt, welcher angeblich der Erfassung der „typischen körperlichen Reaktionen beim Lügen und Wahrheitsagen“ diente. In Wahrheit sollten durch den Zahlentest jedoch die Annahmen der Pbn über die Treffsicherheit der „Lügendetektion“ manipuliert werden. Hierzu wurden die Pbn aufgefordert, eine Zahl zwischen drei und sieben auszuwählen und auf einem vor ihnen befindlichen Blatt zu notieren. Anschließend wurden den Pbn in standardisierter Form sieben Fragen gestellt, die auf die ausgewählte Zahl abzielten. In der ersten Frage wurde gefragt, ob sie die Zahl zwei aufgeschrieben hätten. In der zweiten Frage wurde gefragt, ob sie die Zahl drei notiert hätten. In den übrigen fünf Fragen wurde entsprechend gefragt, ob es sich bei der aufgeschriebenen Zahl um die vier, fünf, sechs, sieben bzw. acht handle (vgl. Anhang A.19). Die Fragen wurden zugleich optisch (Monitor) und akustisch (Lautsprecher) dargeboten (vgl. Abschnitt 4.3.4). Die Pbn wurden aufgefordert, sämtliche Fragen mit nein zu beantworten, sobald die jeweilige Frage auf dem Bildschirm ausgeblendet werde.

Während der Befragungsprozedur wurden die körperlichen Reaktionen gemessen. Nach der Befragung gab Vl 2 den Pbn – unabhängig von deren tatsächlichen körperlichen Reaktionen – mündlich die fingierte Rückmeldung, sie hätten stark reagiert, als sie die Frage, die die aufgeschriebene Zahl beinhaltete, wahrheitswidrig verneinten. Dagegen hätten sie bei den übrigen sechs Fragen, welche wahrheitsgemäß verneint wurden, kaum körperlich reagiert. Dies zeige, daß die körperlichen Reaktionen beim Lügen und Wahrheitsagen leicht unterscheidbar seien und daß die Pbn sich somit für den „Lügendetektortest“ eigneten. In letzterem würden die körperlichen Reaktionsunterschiede zwischen wahrheitsgemäßen Antworten und Lügen gegebenenfalls sogar noch deutlicher ausfallen, da Lügen dort eine größere persönliche Relevanz besäßen.

Nun wurde der eigentliche „Lügendetektortest“ (***GAT***) durchgeführt. Dieser bestand aus zehn Mehrfachwahlfragen mit jeweils sechs Antwortalternativen (s. Tabelle 13). Die Fragen zielten auf dieselben zehn Tatortdetails ab, mit denen sowohl die *Täterinnen* als auch die *Zeuginnen* bei der Begehung des Diebstahls bzw. bei der Erledigung der Aufräumarbeiten in Berührung gekommen waren (s.o.). Wie Tabelle 13 zu entnehmen ist, variierte die Position der relevanten Items zwischen den einzelnen Fragen in unsystematischer Weise. Den eigentlichen Fragen waren jeweils einleitende Sätze vorgeschaltet, welche vorweg den Gegenstandsbereich der Fragen eingrenzten. Hierdurch sollte eine Voraktivierung der entsprechenden relevanten Gedächtnisinhalte bei den Pbn mit Tatwissen erzielt werden. Die Fragen bzw. Items wurden jeweils in simultaner Weise optisch (Schriftzug auf Monitor) und akustisch (Frauenstimme auf Lautsprecher) dargeboten (s. Abschnitt 4.3.4). Simultan bedeutet, daß die entsprechenden optischen und akustischen Reize jeweils das gleiche zeitliche Onset hatten, d.h. im selben Moment, in dem der Schriftzug auf dem Monitor erschien, setzte auch die Frauenstimme ein. Die Dauer der akustischen Reizdarbietungen variierte in Abhängigkeit von der Satz-, Fragen- bzw. Itemlänge (Aussprechdauer) und war nie länger als die optische Präsentationsdauer. Die optische Darbietung der einleitenden Sätze dauerte jeweils 9.9 Sekunden. Gleiches gilt für die optische Darbietung der Fragen. Die Dauer der schriftlichen Einblendung der Items auf dem Bildschirm variierte nach dem Zufallsprinzip zwischen acht und zehn Sekunden. Die reizfreien Intervalle zwischen den optischen Darbietungen der einleitenden Sätze und der Fragen betrugen jeweils 3.6 Sekunden. Gleiches gilt für die reizfreien Intervalle zwischen den Fragen und den ersten Items (Pufferitems). Die Dauer der reizfreien Intervalle zwischen den Items variierte zufällig zwischen 20 und 22 Sekunden. Gleiches gilt für die reizfreien Intervalle zwischen den jeweils letzten Items und den darauf folgenden einleitenden Sätzen. Die optische Ausblendung der Items auf dem Bildschirm signalisierte auch den Zeitpunkt der Antwortgabe, d.h. die Pbn sollten immer mit nein antworten, sobald eine Antwortalternative auf dem Bildschirm ausgeblendet wurde.

Tabelle 13. Befragungsschema des *Guilty Actions Tests*

1. In dem Raum, in dem die 100,- DM gestohlen wurden, befand sich ein Fahrrad mit einer ganz bestimmten Farbe.

   Welche Farbe hatte das Fahrrad in dem Raum, in dem Sie die 100,- DM gestohlen haben? War es ...

   a) rot? d) gelb?*
   b) weiß? e) schwarz?
   c) blau? f) grün?

2. In dem Raum, in dem die 100,- DM gestohlen wurden, befand sich ein Getränkekasten mit einem ganz bestimmten Getränk.

   Welches Getränk war in dem Getränkekasten in dem Raum, in dem Sie die 100,- DM gestohlen haben? War es ...

   a) Milch? d) Wein?
   b) Bier? e) Fruchtsaft?
   c) Cola? f) Wasser?*

3. In dem Raum, in dem die 100,- DM gestohlen wurden, befand sich eine Porzellanfigur, die ein ganz bestimmtes Tier darstellte.

   Welches Tier stellte die Porzellanfigur in dem Raum dar, in dem Sie die 100,- DM gestohlen haben? War es ...

   a) eine Giraffe? d) ein Hase?
   b) ein Hund?* e) eine Katze?
   c) ein Schwan? f) ein Pferd?

4. In dem Raum, in dem die 100,- DM gestohlen wurden, befand sich eine ganz bestimmte Pflanze.

   Welche Pflanze befand sich in dem Raum, in dem Sie die 100,- DM gestohlen haben? War es ...

   a) eine Primel? d) ein Efeu?
   b) eine Palme? e) ein Farn?
   c) ein Kaktus?* f) eine Geranie?

5. In dem Raum, in dem die 100,- DM gestohlen wurden, befand sich eine Obstschale mit einer ganz bestimmten Obstsorte darin.

   Welche Obstsorte befand sich in der Obstschale in dem Raum, in dem Sie die 100,- DM gestohlen haben? Waren es ...

   a) Birnen?
   b) Trauben?
   c) Bananen?
   d) Orangen?
   e) Pflaumen?
   f) Äpfel?*

6. Auf einer der Schreibtischschubladen in dem Raum, in dem die 100,- DM gestohlen wurden, befand sich ein Aufkleber von einem ganz bestimmten Land.

   Welcher Landesaufkleber befand sich auf einer der Schubladen in dem Raum, in dem Sie die 100,- DM gestohlen haben? War es ...

   a) ein Großbritannien-Aufkleber?
   b) ein Portugal-Aufkleber?
   c) ein Deutschland-Aufkleber?*
   d) ein Frankreich-Aufkleber?
   e) ein Italien-Aufkleber?
   f) ein Spanien-Aufkleber?

7. In dem Raum, in dem die 100,- DM gestohlen wurden, hing ein ganz bestimmtes Sportgerät an der Wand.

   Welches Sportgerät hing an der Wand des Raumes, in dem Sie die 100,- DM gestohlen haben? War es ...

   a) ein Tennisschläger? d) ein Golfschläger?
   b) ein Schlittschuh? e) eine Dartscheibe?*
   c) eine Taucherbrille? f) ein Springseil?

8. In dem Raum, in dem die 100,- DM gestohlen wurden, hing ein Bild an der Wand, auf dem ganz bestimmte Tiere abgebildet waren.

   Welche Tiere waren auf dem Bild in dem Raum, in dem Sie die 100,- DM gestohlen haben? Waren es ...

   a) Delphine? d) Kühe?*
   b) Schafe? e) Ziegen?
   c) Hirsche? f) Hühner?

9. In dem Raum, in dem die 100,- DM gestohlen wurden, war ein Holzschrank, in dem sich ein ganz bestimmtes Kleidungsstück befand.

   Welches Kleidungsstück befand sich in dem Holzschrank des Raumes, in dem Sie die 100,- DM gestohlen haben? War es ...

   a) ein Hemd? d) ein Pullover?
   b) eine Jacke?* e) eine Hose?
   c) ein Kleid? f) eine Krawatte?

10. In dem Raum, in dem die 100,- DM gestohlen wurden, befand sich ein Teppich mit einer ganz bestimmten Farbe.

    Welche Farbe hatte der Teppich in dem Raum, in dem Sie die 100,- DM gestohlen haben? War er ...

    a) blau? d) grün?
    b) weiß? e) rot?*
    c) gelb? f) schwarz?

Anmerkung: * relevantes Item.

Vor der Befragungsprozedur wurde den Pbn per Instruktion folgende „Logik“ des „Lügendetektortests“ vermittelt (vgl. Anhang A.20): Unschuldige Testpersonen hätten nichts zu befürchten, da sie bei allen Antwortalternativen die Wahrheit sagten, sofern sie instruktionsgemäß mit nein antworteten. Handle es sich bei einer Testperson jedoch um die Täterin, so reagiere diese unweigerlich bei allen zehn Fragen auf jeweils eine Antwortalternative mit einer Lüge, indem sie instruktionsgemäß mit nein antworte. Diese Lügen seien anhand der starken damit einhergehenden körperlichen Reaktionen der Testperson erkennbar. Vor dem eigentlichen *GAT* wurde noch ein Probedurchgang durchgeführt. Die Befragungsprozedur beim *GAT* war ebenso wie diejenige bei dem vorgeschalteten „Stimulationstest“ (s.o.) vollständig automatisiert. Die Aufgabe von Vl 2 bestand lediglich darin, jeweils die entsprechenden Computerprogramme für die Reizdarbietung und Aufzeichnung der physiologischen Messungen (s. Abschnitt 4.3.4) zu starten, den Versuchsablauf zu überwachen und die Antworten der Pbn zu protokollieren (vgl. Anhang C.1).

Nach Abschluß des *GAT* führte Vl 2 die Pbn zurück in das Büro von Vl 1. Dabei erklärte Vl 2, sie benötige noch ca. 15 Minuten, um sich anhand der Ergebnisse der mündlichen Vernehmung und des „Lügendetektortests“ ein Urteil zu bilden. In der Zwischenzeit sollten die Pbn im Büro von Vl 1 warten und diesem noch ein paar Fragen zum Versuchsablauf beantworten.

Phase VI: Nachbefragung (vgl. Tabelle 12)

Vl 1 ließ die Pbn noch einen Nachbefragungsbogen zu dem simulierten Diebstahl und zu der Untersuchung durch die „gerichtspsychologische Expertin“ ausfüllen (vgl. Anhang A.21). Darin wurde zuerst das freie Erinnern der zehn kritischen Tatortdetails überprüft, die als relevante Items im *GAT* dienten. Weiterhin sollten die Pbn jeweils auf sechsstufigen Skalen beurteilen, wie sehr sie bei der „mündlichen Vernehmung“ und bei dem „Lügendetektortest“ motiviert gewesen waren, glaubwürdig bzw. unschuldig zu erscheinen. Sie sollten ferner auf jeweils elfstufigen Skalen die Wahrscheinlichkeit einschätzen, mit der es ihnen gelungen sei, bei der „mündlichen Vernehmung“ bzw. bei dem „Lügendetektortest“ einen glaubwürdigen bzw. unschuldigen Eindruck zu hinterlassen. Als nächstes wurde ein Wiedererkennungstest bezüglich der zehn kritischen Tatortdetails durchgeführt. Dabei wurden zehn Mehrfachwahlfragen mit den gleichen Antwortalternativen wie im *GAT* verwendet. In der Bedingung *falsche Zeuginnen* war der Widererkennungstest so konzipiert, daß pro Frage jeweils eine Antwortalternative aus dem *GAT* durch dasjenige Tatortdetail ersetzt worden war, das sich während des Aufenthalts der *falschen Zeuginnen* am Tatort (Phase IV: Vorbereitung auf „Vernehmung“) dort befand. Somit konnte auch überprüft werden, wie gut die *falschen Zeuginnen* die für sie zutreffenden Tatortdetails wiedererkannten. Außerdem sollten alle Pbn

sowohl für die „mündliche Vernehmung“ als auch für den „Lügendetektortest“ angeben, inwiefern sie spezielle Maßnahmen (Tricks) ergriffen hätten, um glaubwürdig bzw. unschuldig zu erscheinen. Schließlich bearbeiteten die Pbn noch den *Mehrfachwahl-Wortschatz-Intelligenztest* (*MWT-B*, Lehrl, 1977), ein Screening-Verfahren zur Erfassung der allgemeinen Intelligenz.

Nach dem Ausfüllen des Nachbefragungsbogens und des *MWT-B* wurden die Pbn von Vl 1 und Vl 2 vollständig über den Sinn und Zweck der Untersuchung aufgeklärt, für die Teilnahme entlohnt und verabschiedet. Die durchschnittliche Dauer der Untersuchung betrug für die *Täterinnen* und *Zeuginnen* gut zweieinhalb Stunden, für die *falschen Zeuginnen* ca. zwei Stunden (s. Tabelle 12).

4.3.4 Technischer Versuchsaufbau und physiologische Messungen

Wie im letzten Abschnitt deutlich wurde, hielten sich die Pbn im Laufe des Experiments in drei verschiedenen Räumen des Psychologischen Instituts auf (vgl. Tabelle 12, Abschnitt 4.3.3). Raum 02-131, in welchem VL 1 die Instruktionen für den Ablauf der Verbrechenssimulation und der „Vernehmung“ erteilte und die abschließende Nachbefragung durchführte, war ein gewöhnliches Mitarbeiterbüro. Raum 02-527 war ein Experimentierraum, der speziell für die vorliegende Untersuchung als „Professorenzimmer“ hergerichtet worden war. Das Interieur des Büros von „Prof. Kunze“ ist im Anhang B abgebildet. Die Erhebung der Zeugenaussagen und die psychophysiologischen Untersuchungen (Ruhemessung, Zahlentest, *GAT*) erfolgten in Raum 02-524, einem weiteren Laborraum des Psychologischen Instituts.

Während des halbstandardisierten Interviews zur Erhebung der Zeugenaussage saßen sich die Interviewerin (Vl 2) und die Pb an einem Tisch frontal gegenüber. Rechts neben der Interviewerin war auf einem Stativ eine Videokamera (*Panasonic S-VHS Movie Camera, AG-455 E*) positioniert, die die Zeugenaussagen in Bild und Ton aufzeichnete. Außerdem wurden die Aussagen mit einem Mini-Disc-Gerät (*Sony, MZ-R 35*) aufgenommen.

Während der psychophysiologischen Untersuchungen befanden sich die Pbn in einer elektrisch und akustisch abgeschirmten Meßkabine (*Industrial Accustics Company, Typ 403-A*). Die Temperatur in der Meßabine schwankte zwischen 21 und 23 Grad Celsius, die Luftfeuchtigkeit lag zwischen 25 und 35%. In der Kabine herrschte Dämmerlicht. Das Innere der Kabine ist im Anhang D auf den Abbildungen D.1 und D.2 dargestellt. In der Kabine befand sich ein gepolsterter Stuhl. Dieser war so plaziert, daß eine be-

quem darauf sitzende Pb bei gerade nach vorne gerichtetem Blick durch ein Glasfenster in der Kabinenwand auf einen Computer-Monitor sah. Über diesen Monitor (*Iiyama Vision Master 17, Model No. MF-8617T*) erfolgte die optische Reizdarbietung. Der Abstand zwischen dem Monitor und den Augen der Pb betrug ca. 130 cm. Die Fragen bzw. Items des Zahlentests und des *GAT* wurden auf dem Bildschirm zentriert in der Schriftart *Helvetica Sans Serif* mit einer Buchstabenhöhe von 1 cm (Großbuchstaben) bzw. 0.7 cm (Kleinbuchstaben) dargeboten. Die Breite der Schriftzeichen schwankte je nach Buchstabe zwischen 0.2 und 1 cm. Umfaßte eine Frage mehrere Zeilen, so betrug der Zeilenabstand 1.5 cm. Die akustische Darbietung der Fragen bzw. Items (gesprochen von einer Frauenstimme in normaler Konversationslautstärke) erfolgte über zwei Lautsprecherboxen (*Jamo HI-FI 7084*), die auf dem Boden rechts und links hinter dem Stuhl plaziert waren. Rechts vor dem Stuhl war in Augenhöhe der Pb eine Videokamera (*Panasonic VHS-C Movie Camera, NV-RX17EG/E*) angebracht, deren Objektiv auf die Pb gerichtet war. Die Kamera war an einen Videorecorder (*Toshiba V-227G*) außerhalb der Kabine angeschlossen, der das Verhalten der Pb während der psychophysiologischen Untersuchungen aufzeichnete. Zudem war an die Videokamera außerhalb der Kabine ein Fernsehmonitor angeschlossen, über den Vl 2 das Verhalten der Pb beobachten bzw. deren Äußerungen mithören konnte.

Die Reizdarbietung und Datenerfassung im Rahmen der psychophysiologischen Untersuchungen waren vollständig automatisiert. Vl 2 hatte lediglich die Aufgabe, die entsprechenden Computerprogramme zu starten und deren ordnungsgemäßen Ablauf zu überwachen. Der zeitliche Ablauf der Stimuluspräsentation wurde mit Hilfe des im Haus entwickelten Programms *Edamess* (Münch, 1999) gesteuert, welches auf einem PC (486 DX, 8 MB RAM) installiert war. Das Programm *Edamess* steuerte ein auf einem weiteren PC (Penthium-S 100, 32 MB RAM) installiertes Graphik- und Sprachpräsentationsprogramm (*Visual C*) an, in welches die optischen und akustischen Reize in Form von Bitmap- bzw. Wave-Dateien implementiert waren. Immer wenn das Graphik- und Sprachpräsentationsprogramm vom Programm *Edamess* die entsprechenden Signale erhielt, wurden die Fragen und Items über den Bildschirm bzw. die Lautsprecherboxen (s.o.) dargeboten, wobei die akustischen Stimuli zuvor mittels einer Stereoanlage (*Technics Stereo Integrated Amplifier SU-Z400* und *Technics Stereo Graphic Equalizer SH-8065*) verstärkt bzw. abgemischt wurden.

Das Programm *Edamess* steuerte auch die Aufzeichnung der physiologischen Messungen. Bezüglich der Hautleitfähigkeitsmessung signalisierte es jeweils Beginn und Ende der Aufzeichnungsintervalle und fügte zudem zeitliche Marker für On- und Offset der Reize in die Aufzeichnungen ein. Die Hautleitfähigkeit wurde im Konstant-Spannungsverfahren mit einer Spannung von 0.5 V gemessen (Lykken & Venables, 1971). Die

Ableitung erfolgte bipolar von der thenaren und hypothenaren Erhebung (Walschburger, 1975) der linken Hand. Es wurden Ag/AgCl-Elektroden (Elektrodenfläche: 1 cm²) der Firma *Hellige* und eine 0.05-molare NaCl-Elektrolytpaste (auf „Unibase"-Grundlage; Marke *Aponorm*) verwendet. Vor dem Anbringen der Elektroden wurde die Haut mit Äthylalkohol (70%) gereinigt. Die aus der Anwendung des Konstant-Spannungsverfahrens resultierende Stromstärke wurde unmittelbar nach der Registrierung mittels eines im Haus entwickelten Transformators (*CEDA-12*) in Spannung umgewandelt und frequenzmoduliert. Anschließend wurde das Signal auf den Zählerbaustein eines PCs (486 DX, 8 MB RAM; s.o.) übertragen. Der Zähler wurde mit einer Abtastrate von 10 Hz ausgelesen. Die so erhaltenen digitalisierten Werte wurden in Leitfähigkeitswerte (Mikro-Siemens, µS) umgerechnet und auf Festplatte gespeichert.

Die übrigen physiologischen Messungen (Elektrokardiogramm [EKG], periphere Durchblutung, Bauchatmung, Hauttemperatur[23]) wurden mit dem *Kölner Varioport-System* vorgenommen. Das EKG wurde mit Hilfe von EKG-Elektroden (Ag/AgCl) und Elektrodengel der Firma *Hellige* abgeleitet. Vor dem Anbringen der Elektroden wurde die Haut mit Äthylalkohol (70%) gereinigt. Die Messung erfolgte als Brustwandableitung zwischen dem Manubrium sterni und dem linken untersten Rippenbogen, bei Erdung auf dem rechten untersten Rippenbogen. Das EKG-Signal wurde vom *Varioport-System* mit einer Abtastrate von 512 Hz digitalisiert. Die periphere Durchblutung wurde mit Hilfe eines im Haus entwickelten Photoplethysmographen gemessen, der am äußersten Glied des rechten Zeigefingers angebracht wurde. Das plethysmographische Signal wurde über einen im Haus gebauten Verstärker an das *Varioport-System* weitergeleitet. Hier wurde es mit einer Abtastrate von 512 Hz digitalisiert. Die Atmung wurde mittels eines zum *Varioport-System* gehörenden Atemgurts mit integriertem Dehnungsrezeptor gemessen. Zur Erfassung der Bauchatmung wurde der Gurt leicht oberhalb des Bauchnabels angelegt (vgl. Schandry, 1989; Walschburger, 1976). Das Atemsignal wurde vom *Varioport-System* mit einer Abtastfrequenz von 32 Hz digitalisiert.

Auf drei weiteren Kanälen des *Varioport-Systems* wurden die Hauttemperatur (Abtastrate: 32 Hz), die Antwortzeitpunkte der Pb (Abtastrate: 32 Hz) sowie die Ereignismarker zur Kennzeichnung von Reizbeginn und Reizende (Abtastrate: 512 Hz) registriert. Die Hauttemperatur wurde mit Hilfe eines Temperatursensors gemessen, der am äußersten Glied des rechten Mittelfingers angeklebt wurde. Die Antwortzeitpunkte wurden mit Hilfe eines Mikrophons erfaßt, das links neben dem Meßkabinenstuhl plaziert war und bei jeder stimmlichen Äußerung der Pb ein Signal an das *Varioport-*

[23] In Anlehnung an die Mehrzahl der publizierten Studien zum *TWT* und *GAT* sind ausschließlich die gemessenen Hautleitfähigkeitsveränderungen Gegenstand der vorliegenden Arbeit (s. Abschnitt 5.3). Die übrigen physiologischen Maße wurden rein explorativ miterhoben; eine Auswertung wurde bisher nicht vorgenommen.

System schickte. Die Ereignismarker wurden gesetzt, indem ein vom Programm *Edamess* (s.o.) angesteuerter Verstärker (*Mark-1*, Entwicklung des Hauses) bei jeder Einblendung eines optischen Reizes ein Rechtecksignal auslöste, welches erst bei der Reizausblendung wieder zur Grundlinie zurückkehrte, so daß die Reizdarbietungsintervalle auf dem entsprechenden Kanal des *Varioport-Systems* in Form von Rechtecksignalen registriert wurden.

Sämtliche Messungen mit dem *Varioport-System* wurden zunächst auf einer Speicherkarte gespeichert und anschließend auf die Festplatte eines angeschlossenen PCs (Penthium MMX 233, 64 MB RAM) ausgelesen. Über diesen PC bzw. einen daran angeschlossenen Monitor erfolgte auch die Bedienung des *Varioport-Systems* und die Online-Kontrolle der entsprechenden Messungen.

4.4 Auswertung

4.4.1 Inhaltsanalytische Auswertung der experimentellen Zeugenaussagen

Für jede der 102 experimentell gewonnenen Aussagen wurden die **Ausprägungsgrade der ersten 18 Glaubhaftigkeitskriterien der *Kriterienorientierten Inhaltsanalyse*** (vgl. Abschnitt 2.1.2) auf jeweils **vierstufigen Ratingskalen** mit den Stufen 0 (= *nicht vorhanden*), 1 (= *schwach vorhanden*), 2 (= *mittel vorhanden*) und 3 (= *stark vorhanden*) kodiert. Der verwendete Ratingbogen ist im Anhang C.2 abgedruckt. Kriterium 19 fand in der Auswertung keine Berücksichtigung, da es bislang ausschließlich für den Bereich der Sexualdelikte definiert ist (s. Abschnitt 2.1.2).

Da einerseits die Bestimmung des Ausprägungsgrades ein komplexer Einschätzungsprozess ist (s. Abschnitt 2.1.3), der prinzipiell subjektiven Interpretationsspielraum zuläßt, und andererseits in früheren Studien zur *Kriterienorientierten Inhaltsanalyse* mitunter eine mangelhafte Operationalisierung der Glaubhaftigkeitskriterien beklagt wurde (z.B. Landry & Brigham, 1992; Krahé & Kundrotas, 1992, Steller, 1989), sollte in der vorliegenden Untersuchung auch die Auswertungsobjektivität überprüft werden. Daher wurden alle 102 Aussagen jeweils von **drei unabhängigen Auswertern** analysiert. Die drei Auswerter (Psychologie-Studierende im Hauptstudium) kannten weder den Ablauf der experimentellen Diebstahlsimulation noch den Wahrheitsstatus der Aussagen. Für die Inhaltsanalysen standen ihnen lediglich Aussagetranskripte zur Verfügung. Vor der Auswertung der Aussagen absolvierten sie ein intensives **Ratertraining**, dessen Aufbau sich an einem an der Universität Kiel entwickelten und evaluierten Schulungsprogramm zur *Kriterienorientierten Inhaltsanalyse*, dem sog. *Kieler Trainingsprogramm zur Be-*

urteilung der Glaubwürdigkeit von Zeugenaussagen (*KTBG*; Krause, 1997; Petersen, 1997; vgl. auch Höfer, Krause, Petersen, Sievers & Köhnken, 1999) orientierte.[24] Der Ablauf des hier durchgeführten Ratertrainings ist im Anhang E genauer beschrieben. Die hier verwendeten Operationalisierungsvorschriften zu den einzelnen Kriterien wurden dem *KTBG* entnommen.

Neben der Quantifizierung der 18 Glaubhaftigkeitskriterien sollten die drei Auswerter jeweils noch auf **zehnstufigen bipolaren Skalen** einschätzen, für **wie glaubhaft bzw. unglaubhaft** sie die betreffende Aussage hielten und wie sicher bzw. unsicher sie sich bezüglich dieses Urteils waren (vgl. Anhang C.2). Die Auswerter sollten sich in ihrem Glaubhaftigkeitsurteil zwar an den vorgenommenen Kriterienratings orientieren; allerdings wurden ihnen diesbezüglich keine diagnostischen Entscheidungsrichtlinien vorgegeben (klinisch-intuitive Urteilsbildung, vgl. Abschnitt 2.1.5 und 2.1.6.2). Dabei ist allerdings zu betonen, daß den Auswertern nicht sämtliche Informationsquellen zur Verfügung standen, die grundsätzlich für die diagnostische Urteilsbildung relevant sind (Ergebnisse gezielter Persönlichkeits- und Motivanalysen etc.; vgl. Abschnitt 2.1.3).

Um etwaige **Sequenzeffekte** bei der Auswertung zu **kontrollieren**, wurde die Reihenfolge der auszuwertenden Aussagen zwischen den drei Auswertern folgendermaßen variiert. Auswerter A bearbeitete die 102 Aussagen in der Reihenfolge 1, 2, 3, ... 102. Auswerter B bearbeitete die Aussagen in umgekehrter Reihenfolge (102, 101, 100, ... 1). Auswerter C schließlich bearbeitete die Aussagen in der Reihenfolge 51, 52, 50, 53, 49, 54, ... 1, 102.

4.4.2 Analyse der psychophysiologischen Daten aus dem *Guilty Actions Test*

Als abhängiges physiologisches Maß im *GAT* wurden die **Amplituden der phasischen Hautleitfähigkeitsreaktionen (SCRs) auf die Testitems** quantifiziert. Die Auswertung der SCR-Amplituden erfolgte per EDV mittels eines interaktiven Auswertungsprogramms (*EDAVIEW*, Münch, Psychologisches Institut der Universität Mainz). Die Amplitude einer SCR ergab sich rechnerisch als Änderung der Hautleitfähigkeit (in µS) zwischen Reaktionsminimum und dem darauffolgenden Maximum. In jenen Zweifelsfällen, wenn eine Spontanfluktuation ohne eine merkliche Absenkung von einer nachfolgenden Reaktion überlagert wurde, dienten auch deutlich erkennbare Übergänge der Anstiegsflanken (z.B. „Einkerbung“ des Kurvenverlaufs) als Minima (vgl. Boucsein,

[24] Herrn Prof. Dr. [redacted] sei an dieser Stelle für die freundliche Bereitstellung der Schulungsmaterialien aus dem *KTBG* gedankt.

1988; Edelberg, 1967). Um als SCR gewertet zu werden, mußte die Amplitude eines Hautleitfähigkeitsanstiegs mindestens 0.02 µS betragen.

Es wurden diejenigen SCRs quantifiziert, die sich unmittelbar nach dem Reizbeginn (optische Einblendung eines *GAT*-Items auf dem Bildschirm bzw. Erklingen der Frauenstimme über Lautsprecher) manifestierten. Dabei kamen zwei Quantifizierungsmethoden zur Anwendung, die sich hinsichtlich des gewählten Latenzzeitkriteriums unterschieden. Zum einen wurde in Anlehnung an übliche Empfehlungen im Rahmen experimenteller Studien (z.B. Venables & Christie, 1980) ein Latenzzeitkriterium von einer bis drei Sekunden angelegt, d.h. Hautleitfähigkeitsanstiege wurden nur dann als reizbezogene SCRs gewertet, wenn der Anstieg frühestens eine und spätestens drei Sekunden nach Reizbeginn einsetzte (**SCR-Quantifizierungsmethode A**). Zum anderen kam auch noch ein breiteres Latenzzeitkriterium zur Anwendung, indem alle SCRs quantifiziert wurden, deren Anstiegspunkt in einem Zeitfenster von einer bis zehn Sekunden nach Reizbeginn lag, wobei neben dem Anstiegspunkt auch der Gipfelpunkt der Reaktion innerhalb dieses neunsekündigen Intervalls liegen mußte. Traten innerhalb des definierten Zeitfensters mehrere Reaktionen (Hautleitfähigkeitsanstiege mit anschließender Steigungsumkehr) auf, so wurde diejenige mit der größten Amplitude als reizbezogene SCR gewertet (**SCR-Quantifizierungsmethode B**). Diese Quantifizierungsmethode entspricht eher den in der forensischen Praxis geltenden Standards und korrespondiert zudem in etwa mit der Vorgehensweise in den bisherigen Untersuchungen zum *GAT* durch die Bradley-Gruppe (vgl. Abschnitt 2.2.3.2).

Im Hinblick auf die statistischen Analysen wurden die nach den beiden Quantifizierungsmethoden (s.o.) bestimmten SCR-Amplitudenwerte zusätzlich einer **logarithmischen Transformation** nach der Formel log (x + 1) nach Venables und Christie (1980, S. 17) unterzogen, um so eine Annäherung der Verteilungscharakteristika der Amplitudenwerte an eine Normalverteilung zu begünstigen. Die i.d.R. signifikante Linksschiefe und Steilheit der SCR-Verteilung wird durch diese Logarithmierung weitgehend normalisiert, und mittels der Addition von 1 sind auch Null-Reaktionen mathematisch definiert (vgl. auch Boucsein, 1988).

Zudem wurden die nach den beiden Quantifizierungsmethoden ermittelten SCR-Amplituden jeweils der in Abschnitt 2.2.2.1 beschriebenen **numerischen Auswertungsmethode** nach Lykken (1959) unterzogen. Die Quantifizierung der SCR-Amplituden erfolgte durch Auswerter, die nicht über die experimentelle Gruppenzugehörigkeit der physiologischen Aufzeichnungen informiert waren.

4.4.3 Naive Auswertung der experimentellen Zeugenaussagen

Um eine naive Glaubhaftigkeitsbeurteilung durchzuführen, wurden aus dem Gesamtpool von 102 Aussagen zwei (sich nicht überlappende) Zufallsstichproben von jeweils 15 Aussagen (5 *Täterinnen*, 5 *Zeuginnen*, 5 *falsche Zeuginnen*) gezogen. Diese beiden Aussagenstichproben wurden jeweils 16 naiven Beurteilern (insgesamt also **32 Beurteiler**) in Bild und Ton (Videoaufzeichnungen) präsentiert. Die Präsentation erfolgte im Rahmen von Gruppensitzungen, an denen jeweils 4 Beurteiler teilnahmen. Um Sequenzeffekte zu kontrollieren, bearbeitete bei beiden Aussagenstichproben jeweils eine Hälfte der naiven Beurteiler die 15 Aussagen in der Reihenfolge 1, 2, ... 15, während der anderen Hälfte der Beurteiler die Aussagen in der Sequenz 15, 14, ... 1 dargeboten wurden.

Bei den naiven Beurteilern handelte es sich um Personen, die **keine theoretischen oder praktischen Vorkenntnisse auf dem Gebiet der Glaubhaftigkeitsbeurteilung** besaßen und dementsprechend in ihrer Beurteilung rein intuitiv bzw. unter Rückgriff auf alltagspsychologische Heurismen vorgehen mußten. Sie waren weder über den tatsächlichen Ablauf des Scheinverbrechens noch über die experimentelle Gruppenzugehörigkeit der Aussagen informiert.

Die Beurteilung wurde auf einer **achtstufigen bipolaren Ratingskala** mit den Extremen *äußerst unglaubwürdig* (Skalenstufe 1) und *äußerst glaubwürdig* (Skalenstufe 8) vorgenommen. Außerdem sollten die naiven Auswerter nach der Beurteilung einer Aussage jeweils kurz schriftlich erläutern, worauf sich ihre Einschätzung stützte. Der Auswertungsbogen ist im Anhang C.3 abgedruckt.

5 Ergebnisse

Die statistischen Analysen wurden, sofern es in der folgenden Darstellung nicht anders vermerkt ist, mit dem Programmpaket *SPSS für Windows*, Version *10.0.5*) gerechnet.

5.1 Ausprägungen der Kontrollvariablen

Bevor eine Darstellung der Ergebnisse der verschiedenen Methoden der Glaubhaftigkeitsbeurteilung erfolgt, soll zunächst darauf eingegangen werden, wie die berücksichtigten Kontrollvariablen (vgl. Abschnitt 4.3.2) in der vorliegenden Untersuchung ausgeprägt waren und ob sich diesbezüglich bedeutsame Gruppenunterschiede ergaben.

Elektrodermale Labilität

Die während der fünfminütigen Ruhemessung erfaßte Anzahl elektrodermaler Spontanfluktuationen (NSRs) betrug in der Gruppe *Täterinnen* durchschnittlich 24.06 (SD = 19.45), in der Gruppe *Zeuginnen* 17.09 (SD = 19.94) und in der Gruppe *falsche Zeuginnen* 19.59 (SD = 16.25). Die Unterschiede zwischen den drei Gruppen erwiesen sich in einer einfaktoriellen Varianzanalyse (ANOVA) mit dem Gruppenfaktor Status der aussagenden Person nicht als statistisch signifikant (s. Anhang F, Tabelle F.1).

Freies Erinnern bzw. Wiedererkennen der kritischen Tatortdedails

In dem freien Erinnerungstest bezüglich der zehn kritischen Tatortdetails, die im *GAT* als relevante Items dienten, reproduzierten die *Täterinnen* durchschnittlich 9.79 korrekte Details (SD = 0.48). Die *Zeuginnen* gaben im Mittel 9.91 korrekte Antworten (SD = 0.29). Dagegen konnten die *falschen Zeuginnen* durchschnittlich nur 0.29 Details wiedergeben (SD = 0.63). Eine einfaktorielle ANOVA ergab einen signifikanten Gruppeneffekt, $F(2,99) = 4390.455$, $p < .01$ (vgl. Anhang F, Tabelle F.2). Scheffé-Tests für paarweise Gruppenvergleiche zeigten, daß sich die *falschen Zeuginnen* sowohl von den *Täterinnen* als auch von den *Zeuginnen* signifikant unterschieden, jeweils $p < .01$. Der Unterschied zwischen den *Täterinnen* und *Zeuginnen* war nicht signifikant (vgl. Anhang F, Tabelle F.3).

Die Ergebnisse im Wiedererkennungstest korrespondieren mit denjenigen im Test für das freie Erinnern. Auch hier zeigte sich in einer einfaktoriellen ANOVA ein signifikanter Effekt des Gruppenfaktors, $F(2,99) = 1910.819$, $p < .01$ (vgl. Anhang F, Tabelle F.4). Dieser beruhte den Ergebnissen von Scheffé-Tests zufolge darauf, daß die *falschen Zeuginnen* ($M = 0.44$, $SD = 1.16$) durchschnittlich weniger kritische Tatortdetails wie-

dererkannten als die *Täterinnen* (M = 9.85, SD = 0.36) und als die *Zeuginnen* (M = 9.88, SD = 0.33), jeweils p < .01 (vgl. Anhang F, Tabelle F.5).

Motivation, im *GAT* einen unschuldigen Eindruck zu hinterlassen

Die auf der sechsstufigen Skala retrospektiv eingeschätzte Motivation, im *GAT* einen unschuldigen Eindruck zu hinterlassen, betrug in der Gruppe *Täterinnen* durchschnittlich 4.50 (SD = 1.13), d.h. die Pbn dieser Gruppe waren „ziemlich motiviert" bis „stark motiviert". Die *Zeuginnen* waren im Durchschnitt „ziemlich motiviert" (M = 4.03, SD = 1.17). Die *falschen Zeuginnen* waren „ziemlich" bis „stark motiviert" (M = 4.44, SD = 0.99). Die Gruppenunterschiede erwiesen sich in einer einfaktoriellen ANOVA als nicht signifikant (vgl. Anhang F, Tabelle F.6).

Motivation, beim Ablegen der Zeugenaussage glaubhaft zu erscheinen

Die auf der sechsstufigen Skala retrospektiv eingeschätzte Motivation, beim Ablegen der Zeugenaussage einen glaubhaften Eindruck zu hinterlassen, war sowohl bei den *Täterinnen* (M = 4.79, SD = 0.84) als auch bei den *Zeuginnen* (M = 4.68, SD = 0.88) und den *falschen Zeuginnen* (M = 4.79, SD = 0.81) „ziemlich" bis „stark" ausgeprägt. Eine einfaktorielle ANOVA ergab keinen signifikanten Gruppeneffekt (vgl. Anhang F, Tabelle F.7).

Subjektiv eingeschätzte persönliche Erfolgswahrscheinlichkeit im *GAT*

Die auf der elfstufigen Skala retrospektiv eingeschätzte Wahrscheinlichkeit, im *GAT* einen unschuldigen Eindruck hinterlassen zu haben, unterlag einem signifikanten Gruppeneffekt, F(2,99) = 22.020, p < .01 (vgl. Anhang F, Tabelle F.8). Scheffé-Tests zeigten, daß die *Täterinnen* (M = 33.24%, SD = 23.19%) ihre Erfolgschancen im Mittel signifikant geringer einschätzten als die *Zeuginnen* (M = 67.06%, SD = 23.55%) und die *falschen Zeuginnen* (M = 64.41%, SD = 23.38%), jeweils p < .01. Zwischen den beiden letztgenannten Gruppen bestand kein signifikanter Unterschied (vgl. Anhang F, Tabelle F.9).

Subjektiv eingeschätzte persönliche Erfolgswahrscheinlichkeit beim Ablegen der Zeugenaussage

Die auf der elfstufigen Skala retrospektiv eingeschätzte Wahrscheinlichkeit, beim Ablegen der Zeugenaussage einen glaubhaften Eindruck hinterlassen zu haben, unterschied sich zwischen den Gruppen. Der signifikante Gruppeneffekt der einfaktoriellen ANOVA, F(2,99) = 15.590, p < .01 (vgl. Anhang F, Tabelle F.10), beruhte darauf, daß die *Zeuginnen* (M = 76.47%, SD = 15.74%) ihre Erfolgschancen im Durchschnitt signifikant höher einschätzten als die *Täterinnen* (M = 49.41%, SD = 23.09%) und die *fal-*

schen Zeuginnen (M = 57.94%, SD = 21.71%), jeweils $p < .01$. Letztere zwei Gruppen unterschieden sich nicht signifikant (vgl. Anhang F, Tabelle F.11).

Manipulationsmaßnahmen im *GAT*

Auf die Frage, ob sie im *GAT* irgendeine Strategie, Taktik oder Technik angewendet hätten, um einen unschuldigen Eindruck zu hinterlassen, antworteten 67.6% der *Täterinnen* mit ja. Dagegen berichteten nur 26.5% der *Zeuginnen* und 35.3% der *falschen Zeuginnen*, entsprechende Maßnahmen ergriffen zu haben. Ein 3×2-Chi-Quadrat-Test (nach Pearson) mit den Faktoren Status der aussagenden Person (*Täterinnen* vs. *Zeuginnen* vs. *falsche Zeuginnen*) und Manipulationsmaßnahmen im *GAT* („nein“ vs. „ja“) erbrachte ein signifikantes Ergebnis, $\chi^2(2) = 13.030$, $p < .01$. Um zu überprüfen, welche Gruppen sich signifikant voneinander unterschieden, wurde eine einfaktorielle ANOVA mit dem Gruppenfaktor als UV und der Variable Manipulationsmaßnahmen im *GAT* (0 [= „nein“], 1 [= „ja“]) als AV gerechnet. Bortz (1999, S. 493) weist darauf hin, daß eine einfaktorielle ANOVA mit einer dichotomen AV zu den gleichen statistischen Entscheidungen führt wie ein entsprechender k×2-Chi-Quadrat-Test, sofern die Stichprobenumfänge genügend groß sind. Im vorliegenden Fall erbrachte die einfaktorielle ANOVA, ebenso wie der 3×2-Chi-Quadrat-Test, ein signifikantes Resultat, $F(2,99) = 7.249$, $p < .01$ (vgl. Anhang F, Tabelle F.12). Anhand paarweiser Scheffé-Tests wurde überprüft, auf welche Gruppenunterschiede die Overall-Signifikanz zurückzuführen war. Es zeigte sich, daß die *Täterinnen* (M = 0.68, SD = 0.47) sich signifikant von den *Zeuginnen* (M = 0.27, SD = 0.45) und den *falschen Zeuginnen* (M = 0.35, SD = 0.49) unterschieden, $p < .01$ bzw. $p < .05$. Zwischen den beiden letztgenannten Gruppen ergab sich kein signifikanter Unterschied (vgl. Anhang F, Tabelle F.13).

Manipulationsmaßnahmen beim Ablegen der Zeugenaussage

Die *Täterinnen*, *Zeuginnen* und *falschen Zeuginnen* gaben mit 55.9%, 41.2% bzw. 50.0% in etwa gleich häufig an, beim Ablegen der Zeugenaussage irgendeine Strategie, Taktik oder Technik zur Steigerung der Glaubhaftigkeit angewendet zu haben. Weder ein 3×2-Chi-Quadrat-Test mit den Faktoren Status der aussagenden Person und Manipulationsmaßnahmen beim Ablegen der Zeugenaussage („nein“ vs. „ja“), $\chi^2(2) = 1.491$, $p = .475$, noch eine einfaktorielle ANOVA mit dem Gruppenfaktor als UV und Manipulationsmaßnahmen beim Ablegen der Zeugenaussage (0 [= „nein“], 1 [= „ja“]) als AV (vgl. Anhang F, Tabelle F.14) erbrachten signifikante Ergebnisse.

Intellektuelle Begabung

Die im MWT-B erzielten Scores unterschieden sich nicht signifikant zwischen den drei Gruppen (vgl. Anhang F, Tabelle F.15). Sowohl die *Täterinnen* (M = 30.00, SD = 3.77) als auch die *Zeuginnen* (M = 29.44, SD = 4.72) und die *falschen Zeuginnen* (M = 30.15,

SD = 2.83) erzielten im Durchschnitt Testergebnisse, die auf eine gut durchschnittliche Intelligenz verweisen.

Vorerfahrung mit Diebstahl

Die Frage, ob sie vor der aktuellen Untersuchung schon einmal Zeugin eines Diebstahls gewesen seien, wurde von 20.6% der *Zeuginnen* mit ja beantwortet. Dagegen machten die *Täterinnen* und die *falschen Zeuginnen* nur zu 5.9% bzw. 2.9% entsprechende Angaben. Ein 3×2-Chi-Quadrat-Test, $\chi^2(2) = 6.874$[25], $p < .05$, und eine einfaktorielle ANOVA, $F(2,99) = 3.577$, $p < .05$ (vgl. Anhang F, Tabelle F.16) verwiesen auf signifikante Gruppenunterschiede. In Scheffé-Tests erwies sich der Unterschied zwischen den *Zeuginnen* und den *falschen Zeuginnen* als statistisch bedeutsam, $p < .05$, während die beiden anderen paarweisen Gruppenvergleiche keine signifikanten Ergebnisse erbrachten (vgl. Anhang F, Tabelle F.17).

Vorkenntnisse zur Glaubhaftigkeitsbeurteilung

Gefragt, ob sie irgendwelche Vorkenntnisse zur „Lügendetektion bzw. Glaubwürdigkeitsbegutachtung“ hätten, antworteten 14.7% der *Täterinnen*, 29.4% der *Zeuginnen* und 23.5% der *falschen Zeuginnen* mit „ja“. Die Gruppenunterschiede erwiesen sich weder in einem 3×2-Chi-Quadrat-Test, $\chi^2(2) = 2.133$, $p = .344$, noch in einer einfaktoriellen ANOVA (vgl. Anhang F, Tabelle F.18) als statistisch signifikant.

Frühere Teilnahme an Untersuchungen zur Glaubhaftigkeitsbeurteilung

Keine der *Täterinnen* und *Zeuginnen* gab an, früher schon einmal an einer Untersuchung „zur Lügendetektion“ bzw. „Glaubwürdigkeitsbegutachtung“ teilgenommen zu haben. Lediglich zwei *falsche Zeuginnen* (5.9%) machten entsprechende Angaben. Diese Häufigkeit wich jedoch nicht überzufällig von denen in den beiden anderen Gruppen ab – die Ergebnisse eines 3×2-Chi-Quadrat-Tests, $\chi^2(2) = 4.080$[26], $p = .130$, und einer einfaktoriellen ANOVA (vgl. Anhang F, Tabelle F.19) waren nicht signifikant.

---

[25] Der k×l-Chi-Quadrat-Test setzt eigentlich voraus, daß der Anteil der erwarteten Häufigkeiten, die kleiner als 5 sind, 20% nicht überschreitet (Bortz, 1999, S. 167, 170). Im vorliegenden Fall war die erwartete Häufigkeit jedoch in 3 Zellen (50%) jeweils kleiner als 5. Das Ergebnis des Chi-Quadrat-Tests findet hier dennoch Berücksichtigung, weil es sich mit dem Ergebnis der einfaktoriellen ANOVA deckt.

[26] Für die Interpretation dieses χ^2-Wertes gelten die gleichen Einschränkungen, wie sie in Fußnote 25 erläutert wurden.

5.2 Resultate der inhaltsorientierten Glaubhaftigkeitsbeurteilung

5.2.1 Auswertungsobjektivität der *Kriterienorientierten Inhaltsanalyse*

Pro Aussage lagen zu jedem der ersten 18 Glaubhaftigkeitskriterien der *Kriterienorientierten Inhaltsanalyse* (Steller & Köhnken, 1989) die Werte von drei Beurteilern vor. Daher mußte zunächst überprüft werden, inwieweit sich die drei Rater bei der Einschätzung der Ausprägungsgrade der einzelnen Kriterien auf der vierstufigen Ratingskala (0 = *nicht vorhanden*, 1 = *schwach vorhanden*, 2 = *mittel vorhanden*, 3 = *stark vorhanden*) einig waren. Hierbei handelt es sich um die Feststellung der Auswertungsobjektivität (z.B. Lienert & Raatz, 1998) bzw. Interrater-Reliabilität (z.B. Crocker & Algina, 1986, S. 143). Deren quantitative Bestimmung erfolgt im allgemeinen über die Berechnung von Produkt-Moment-Korrelationen, Kappa-Koeffizienten sowie auf varianzanalytischem Wege (vgl. Amelang & Zielinski, 1997, S. 144; Bortz, Lienert & Boehnke, 1990, S. 482). Diese drei Bestimmungsmethoden kamen auch hier zur Anwendung. Zudem wurde als rein deskriptives Maß der Urteilskonkordanz auch noch die prozentuale Übereinstimmung zwischen den Ratern berechnet (vgl. Frick & Semmel, 1978, S. 164; Bortz & Döring, 1995, S. 253f.). Mit Ausnahme der Produkt-Moment-Korrelationen mußten alle Statistiken von Hand bzw. unter Zuhilfenahme des Tabellenkalkulationsprogramms M*icrosoft Excel 97 für Windows* berechnet werden, da dies mit den verfügbaren Statistik-Programmpaketen nicht in befriedigender Weise möglich war.

Abweichend von der üblichen Vorgehensweise sollen im folgenden die Ergebnisse der Objektivitätsanalysen nicht nur beschrieben, sondern auch schon interpretiert bzw. diskutiert werden. Die vorgezogene Diskussion der Ergebnisse ist erforderlich, da die weitere Auswertungsprozedur (Analyse der Validität bzw. Differenzierungsfähigkeit der *Kriterienorientierten Inhaltsanalyse*) davon abhängt, wie die Auswertungsobjektivität der einzelnen Kriterien bewertet wird. So ist die angestrebte Zusammenfassung der Scores aller drei Rater zu jeweils einem Score (arithmetische Mittelung) nur zulässig, wenn für das entsprechende Kriterium eine ausreichende Auswertungsobjektivität nachgewiesen werden kann.

5.2.1.1 Einfache prozentuale Übereinstimmung

Pro Kriterium wurde für die drei möglichen Zweierkombinationen von Ratern (Rater A – Rater B, Rater A – Rater C, Rater B – Rater C) berechnet, bei wie vielen der insgesamt 102 auszuwertenden Aussagetranskripte beide Rater die gleiche Skalenstufe gewählt hatten. Als Maß der Gesamtübereinstimmung zwischen allen drei Ratern wurde

zum einen pro Kriterium der Mittelwert der drei paarweisen prozentualen Raterübereinstimmungen berechnet (vgl. Asendorpf und Wallbott, 1979, S. 250). Als konservativeres Maß der Gesamtübereinstimmung wurde außerdem auch noch pro Kriterium die direkte prozentuale Übereinstimmung aller drei Rater bestimmt. Dabei galt es als Übereinstimmung, wenn alle drei Rater (A, B und C) einer Aussage die gleiche Kriteriumsausprägung zuordneten. Die Ergebnisse sind im Anhang F, Tabelle F.20, dargestellt.

Legt man die Untergrenze für eine als hoch zu bewertende prozentuale Übereinstimmung bei 70% fest (vgl. Wellershaus & Wolf, 1989, zitiert nach Krause & Petersen, keine Jahresangabe, S. 22), so fällt auf, daß zwischen den Ratern A und B ebenso wie zwischen den Ratern A und C und zwischen den Ratern B und C jeweils nur bei neun Kriterien eine als hoch einzustufende Übereinstimmung vorlag (vgl. Tabelle F.20). Auch von den mittleren prozentualen Übereinstimmungen wurde der Wert von 70% lediglich bei 9 Kriterien überschritten (Kriterien 1, 5, 7, 10, 11, 15, 16, 17 und 18; vgl. Tabelle F.20). Hinsichtlich der direkten prozentualen Übereinstimmung aller drei Rater (A – B – C) ergaben sich nur bei sechs Kriterien hohe Prozentwerte (Kriterien 1, 7, 10, 11, 16 und 17).

5.2.1.2 Erweiterte prozentuale Übereinstimmung

Bei einer vierstufigen Ratingskala ist es jedoch durchaus noch akzeptabel, wenn die Urteile der Rater nur um einen Skalenpunkt voneinander abweichen, da beide Urteile in diesem Fall immer noch gleiche Tendenzen in der Einschätzung der Kriteriumsausprägung darstellen. Daher wurde zusätzlich zur einfachen auch noch die erweiterte prozentuale Übereinstimmung zwischen den Ratern bestimmt, bei der es als Übereinstimmung galt, wenn die Rater einer Aussage die gleiche oder zwei direkt benachbarte Skalenstufen zuordneten. Auch hier wurden pro Kriterium die drei paarweisen Raterübereinstimmungen (A – B, A – C, B – C) sowie deren Mittelwert als Maß der Gesamtübereinstimmung berechnet. Als konservativeres Maß der Gesamtübereinstimmung wurde auch hier zudem die direkte Übereinstimmung aller drei Rater (A – B – C) bestimmt, wobei es als Übereinstimmung galt, wenn von allen drei Ratern die gleiche Skalenstufe (z.B.: Alle drei Rater wählten die Skalenstufe 1.) oder zwei direkt benachbarte Skalenstufen (z.B.: Rater A wählte Skalenstufe 1; Rater B und Rater C wählten Skalenstufe 2.) gewählt wurden. Die Ergebnisse finden sich im Anhang F, Tabelle F.21. Wie Tabelle F.21 zu entnehmen ist, wies Rater B sowohl mit Rater A als auch mit Rater C bei allen 18 Kriterien eine hohe bis sehr hohe erweiterte Urteilsübereinstimmung auf – die Prozentwerte lagen meist sehr deutlich über 70%. Zwischen den Ratern A und C betrug die

erweiterte Übereinstimmung nur bei einem Kriterium (9) – wenn auch unwesentlich – weniger als 70% und lag ansonsten meist deutlich darüber. Der Durchschnitt der drei paarweisen erweiterten prozentualen Übereinstimmungen war bei allen 18 Kriterien hoch. Sogar die direkte erweiterte prozentuale Übereinstimmung aller drei Rater war bei 15 Kriterien hoch bis sehr hoch. Die Kriterien 3 und 14 verfehlten die Marke von 70% nur knapp. Lediglich bei Kriterium 9 lag die direkte erweiterte Übereinstimmung der drei Rater deutlich unter 70% (vgl. Tabelle F.21).

5.2.1.3 Produkt-Moment-Korrelationen

Pro Kriterium wurden die Korrelationen zwischen den Beurteilungen von je zwei Ratern berechnet. Bei drei möglichen Zweierkombinationen von Ratern (A – B, A – C, B – C) ergaben sich somit 3 Korrelationskoeffizienten pro Kriterium. Als Maß der Gesamtübereinstimmung zwischen allen drei Ratern wurde pro Kriterium der Durchschnittswert der drei paarweisen Korrelationen berechnet (Amelang & Zielinski, 1997, S.144). Bei der Mittelungsprozedur kam Fishers Z-Transformation zur Anwendung (vgl. Bortz, 1999, S. 209). Tabelle F.22 im Anhang F faßt die Ergebnisse der Korrelationsberechnungen zusammen.

Es fällt auf, daß die Urteile der Rater A und B nur bei fünf Kriterien (3, 4, 6, 12, 13) in akzeptabler Höhe ($r \geq .40$) miteinander korrelierten. Die Korrelationen zwischen den Ratern A und C erreichten nur bei acht Kriterien (2, 3, 4, 6, 12, 13, 14, 18), die zwischen den Ratern B und C lediglich bei sieben Kriterien (3, 5, 6, 8, 12, 13, 15) eine akzeptable Höhe. Dementsprechend waren auch die gemittelten Korrelationen nur bei fünf Kriterien substantiell (3, 4, 6, 12, 13; vgl. Tabelle F.22), wobei allerdings keine Signifikanzprüfung vorgenommen werden konnte.

Es ist jedoch zu betonen, daß die Aussagekraft vieler Korrelationskoeffizienten im vorliegenden Fall beträchtlich herabgesetzt ist. Der Grund hierfür liegt darin, daß für viele Kriterien die Variationsbreite der entsprechenden Ratings in der vorliegenden Stichprobe sehr stark eingeschränkt war. Dies führte dazu, daß die einschlägigen Stichprobenkorrelationskoeffizienten die entsprechenden Populationskorrelationen, d.h. den wirklichen Zusammenhang der Beurteilungen der Rater, erheblich unterschätzen (vgl. Bortz, 1999, S. 205f.). Um diesen Sachverhalt zu verdeutlichen, sind im Anhang F, Abbildung F.1, die nach Kriterien und Ratern getrennten Häufigkeitsverteilungen der Ratings mit den korrespondierenden Varianzwerten abgetragen. Wie aus Abbildung F.1 hervorgeht, waren bei sechs Kriterien (1, 7, 10, 11, 16, 17) die Häufigkeitsverteilungen der Ratings aller drei Rater jeweils sehr rechts- bzw. linksschief, d.h. die Rater ordneten

den Aussagen fast ausschließlich die Skalenstufen 0 bzw. 3 zu. Die Einseitigkeit dieser Ratings findet auch in entsprechend niedrigen Varianzwerten ihren Ausdruck (vgl. Abbildung F.1). Bei eben diesen sechs Kriterien mit den gestutzten Häufigkeitsverteilungen der entsprechenden Ratings fielen auch die Produkt-Moment-Korrelationskoeffizienten am niedrigsten aus bzw. konnten aufgrund des völligen Fehlens von Varianz in den Ratings von mindestens einem Rater erst gar nicht berechnet werden (vgl. Tabelle F.22). Auch weitere drei Kriterien, für die sich keine substantiellen Korrelationskoeffizienten ergaben (8, 15, 18), waren dadurch gekennzeichnet, daß die entsprechenden Ratings meist eine vergleichsweise geringe Variationsbreite aufwiesen (vgl. Abbildung F.1). Insgesamt spricht die dargelegte Problematik dafür, daß die Produkt-Moment-Korrelation im vorliegenden Fall nicht als Kennwert für die Interrater-Reliabilität geeignet ist.

5.2.1.4 Gewichtete Kappa-Koeffizienten

Pro Kriterium wurde für jede mögliche Zweierkombination von Ratern (A – B, A – C, B – C) der gewichtete Kappa-Koeffizient berechnet. Kappa gibt den zufallskorrigierten Anteil der Raterübereinstimmungen wieder. Bei der Berechnung des gewichteten Kappa-Koeffizienten werden zudem Nichtübereinstimmungen je nach der Distanz zwischen den gewählten Skalenstufen gewichtet, so daß hier das Ausmaß der Unterschiedlichkeit der abgegebenen Urteile Berücksichtigung findet (vgl. Cohen, 1968). Da die Urteilskategorien (Skalenstufen) ein eindimensionales Kontinuum mit konstanten Abständen zwischen den Kategorien repräsentierten und somit faktisch von einer Intervallskala ausgegangen wurde, wurde eine quadratische Gewichtungsstruktur zugrunde gelegt (Cohen, 1968, S. 218f.; vgl. auch Asendorpf & Wallbott, 1979, S. 249; Bortz, 1999, S. 203f.). Als Maß der Gesamtübereinstimmung zwischen allen drei Ratern wurde pro Kriterium der Durchschnittswert der drei paarweisen Kappa-Koeffizienten berechnet (Michel, 1965). Die Ergebnisse sind im Anhang F, Tabelle F.23 zusammengefaßt.

Kappa-Koeffizienten in einem Bereich von .40 bis .74 gelten als Hinweis auf eine akzeptable bis gute Urteilerübereinstimmung, Kappa-Werte ≥ .75 sollen eine sehr hohe Übereinstimmung indizieren (Fleiss, 1981, S. 223; Landis & Koch, 1977). Wie Tabelle F.23 zu entnehmen ist, erreichten die Kappa-Koeffizienten bezüglich der Rater A und B nur bei vier Kriterien (3, 6, 12, 13) eine zumindest akzeptable Höhe. Bei dem Raterpaar A – C waren die Koeffizienten lediglich bei fünf Kriterien (4, 6, 12, 13, 18) befriedigend, ebenso bei dem Raterpaar B – C (Kriterien 3, 6, 8, 13, 15). Was die Gesamtübereinstimmung zwischen allen drei Ratern angeht, so deuteten die über die drei Raterpaare

gemittelten gewichteten Kappa-Koeffizienten (s. Tabelle F.23) nur bei fünf Kriterien (3, 4, 6, 12, 13) auf eine zumindest akzeptable Urteilerübereinstimmung hin.

Die Interpretierbarkeit der Kappa-Koeffizienten ist im vorliegenden Fall jedoch ebenso eingeschränkt wie die der Korrelationskoeffizienten (s.o.). Der Grund hierfür liegt erneut in den Verteilungseigenschaften der Ratings. Das Kappa-Maß reagiert empfindlich auf die Verteilung der Raterurteile über die Kategorien. Verteilen sich – wie es hier bei vielen Kriterien der Fall war (s. Abbildung F.1) – die Ratings zweier Auswerter jeweils nur schwach über die verfügbaren Skalenstufen, so kann dies dazu führen, daß viele Zellen der Übereinstimmungsmatrix, die die Grundlage für die Berechnung von Kappa bildet, leer sind. Sind jedoch nur wenige Zellen der Matrix besetzt, so unterschätzt der resultierende Kappa-Koeffizient häufig die tatsächliche Raterübereinstimmung (Asendorpf & Wallbott, 1979, S. 250). Zur Veranschaulichung dieses Sachverhalts sind in Tabelle F.24 im Anhang F die Übereinstimmungsmatrizen getrennt nach Kriterien und Raterpaaren abgetragen. Wie aus Tabelle F.24 hervorgeht, waren die sechs Kriterien mit den niedrigsten Kappa-Koeffizienten (Kriterien 1, 7, 10, 11, 16, 17; mittleres Kappa jeweils < .1; vgl. Tabelle F.23) dadurch gekennzeichnet, daß in den entsprechenden Übereinstimmungsmatrizen jeweils eine erhebliche Anzahl von Zellen unbesetzt war. Bei vier weiteren Kriterien mit unbefriedigenden Kappa-Koeffizienten (Kriterien 5, 8, 15, 18; mittleres Kappa jeweils < .4) nahm die Anzahl leerer Zellen in den entsprechenden Übereinstimmungsmatrizen ebenfalls jeweils ein beträchtliches Ausmaß an. Lediglich bei den Kriterien 2, 9 und 14 konnten die unbefriedigenden Kappa-Koeffizienten nicht auf eine zu hohe Anzahl unbesetzter Zellen in den entsprechenden Übereinstimmungsmatrizen zurückgeführt werden, wobei allerdings Kriterium 2 die Marke von Kappa = .4 nur knapp verfehlte. Alles in allem spricht die erörterte Problematik dafür, daß auch das gewichtete Kappa-Maß im vorliegenden Fall keinen angemessenen Kennwert der Auswertungsobjektivität darstellt.[27]

5.2.1.5 Varianzanalytische Bestimmung der Auswertungsobjektivität

Für jedes Kriterium wurde eine Varianzanalyse (ANOVA) mit zwei Faktoren (Aussagen und Auswerter) und einer Beobachtung pro Zelle berechnet (ANOVA-Modell der Zufallseffekte, vgl. Iseler, 1967, S. 136). Der Objektivitätsindex, der sich mit dieser Methode ermitteln läßt, ist der Anteil der Variabilität zwischen den Aussagen an der Gesamtvarianz, welcher als Maß der Konsistenz oder Werteübereinstimmung innerhalb

[27] Daher wurde darauf verzichtet, als weiteres Maß der Gesamtübereinstimmung zwischen allen drei Ratern pro Kriterium auch noch den gewichteten Kappa-Koeffizienten für mehrere (> 2) Rater zu berechnen, welcher eigentlich ein angemessenerer Index der Gesamtübereinstimmung ist als der Mittelwert der paarweisen Kappa-Koeffizienten (Cohen, 1972).

von Fällen (**Intraklassenkorrelation** [r_{AA}]; vgl. z.B. Iseler, 1967) aufzufassen ist. Für jedes Kriterium wurden sowohl die Intraklassenkorrelation für die Ratings einzelner Auswerter ($r_{AA[single\ measure]}$) als auch die Intraklassenkorrelation für die über alle drei Auswerter gemittelten Ratings ($r_{AA[average\ measure]}$) berechnet. Letztgenannter Koeffizient ist als Objektivitätsmaß vorzuziehen, sofern, wie im vorliegenden Fall, eine Mittelung der Scores mehrerer Rater anvisiert ist (Crocker & Algina, 1986, S. 167).

Die Ergebnisse der 18 ANOVAs und die entsprechenden Intraklassenkorrelationskoeffizienten sind im Anhang F, Tabelle F.25, zusammengefaßt. Tabelle F.25 ist zu entnehmen, daß die Intraklassenkorrelationen für die über alle drei Auswerter gemittelten Ratings lediglich bei elf Kriterien (2, 3, 4, 5, 6, 8, 12, 13, 14, 15, 18) akzeptabel bis gut ausfielen ($r_{AA[average\ measure]} \geq .40$). Für die übrigen sieben Kriterien war die Höhe der Koeffizienten unbefriedigend.

Die Höhe einer Intraklassenkorrelation hängt jedoch nicht nur von der Höhe der Auswerterübereinstimmung ab, sondern auch von der Varianz zwischen den Aussagen, d.h. davon, wie stark die über die Auswerter gemittelten Ratings streuen. Ist die Variabilität zwischen den Aussagen sehr gering, so kann die Intraklassenkorrelation die Höhe der Auswerterübereinstimmung nicht hinreichend wiedergeben (Asendorpf & Wallbott, 1979, S. 245; Frick & Semmel, 1978, S. 164). Betrachtet man nun im vorliegenden Fall die Varianz zwischen den Aussagen (s. Anhang F, Tabelle F.25), so fällt auf, daß diese bei den meisten Kriterien recht niedrig war, daß sie jedoch insbesondere bei sechs Kriterien (1, 7, 10, 11, 16, 17) drastisch reduziert war. Für diese sechs Kriterien fielen auch die Intraklassenkorrelationskoeffizienten mit Abstand am niedrigsten aus.

Im Falle zu geringer Varianz zwischen den Aussagen stellt der Anteil der Varianz zwischen den Auswertern an der Gesamtvarianz (z.B. Michel, 1965, S.167ff.) einen ersten Anhaltspunkt bezüglich der tatsächlichen Urteilskonkordanz zwischen den Ratern dar. Dieser Quotient gibt an, wie sehr die Intraklassenkorrelation durch mangelnde Raterübereinstimmung gesenkt wird. Je niedriger das Verhältnis ausfällt, desto höher ist die Übereinstimmung zwischen den Ratern einzustufen. Wie Tabelle F.25 im Anhang F zu entnehmen ist, lag das Verhältnis der Auswertervarianz zur Gesamtvarianz bei zwölf Kriterien (1, 2, 4, 5, 6, 7, 10, 11, 15, 16, 17, 18) in einem vernachlässigbar niedrigen Bereich und war bei den übrigen sechs Kriterien immerhin noch akzeptabel. So besagt der insgesamt schlechteste Quotient von .259 bei Kriterium 8, daß die Auswertungsobjektivität hier um 25.9% durch Uneinigkeiten zwischen den Auswertern gesenkt wurde.

Als alternativer Kennwert zur Intraklassenkorrelation im Falle zu geringer Varianz zwischen den Aussagen ist der **Finn-Koeffizient** zu berechnen, welcher den Anteil der

nicht fehlerbedingten Varianz an der Gesamtvarianz reflektiert (Finn, 1970; vgl. auch Asendorpf und Wallbott, 1979, S. 245). Variabilität zwischen den Auswertern gilt hierbei als Fehlervarianz. Tabelle F.26 im Anhang F faßt die Resultate der Finn-Statistik zusammen. Für 13 Kriterien (1, 2, 5, 6, 7, 8, 10, 11, 13, 15, 16, 17, 18), darunter auch sechs der insgesamt sieben Kriterien mit unbefriedigenden Intraklassenkorrelationen (s.o.), ergaben sich hohe bis sehr hohe Finn-Koeffizienten (≥ .75). Bei weiteren vier Kriterien (3, 4, 12, 14) waren die Finn-Koeffizienten akzeptabel bis gut (zwischen .40 und .75). Lediglich für Kriterium 9 war der Finn-Koeffizient (ebenso wie die Intraklassenkorrelation) unbefriedigend.

5.2.1.6 Zusammenfassende Würdigung der Auswertungsobjektivität

Tabelle 14 bietet einen Gesamtüberblick über die Ergebnisse der durchgeführten Objektivitätsanalysen. Pro Kriterium und pro statistischer Bestimmungsmethode wird angegeben, ob der entsprechende Objektivitätsindex mindestens eine akzeptable Höhe erreichte (+) oder unbefriedigend ausfiel (–). Die Produkt-Moment-Korrelationen, gewichteten Kappa-Koeffizienten und Intraklassenkorrelationen sind im vorliegenden Fall nur eingeschränkt interpretierbar (s.o.). Daher orientiert sich die abschließende Bewertung der Auswertungsobjektivität bzw. Interrater-Reliabilität an den erweiterten prozentualen Übereinstimmungen und an den Finn-Koeffizienten. Bezüglich der Interpretation der prozentualen Übereinstimmungen ist jedoch einschränkend zu berücksichtigen, daß bei diesem Maß keine zufallskritische Absicherung möglich ist (z.B. Asendorpf & Wallbott, 1979, S. 248; Frick & Semmel, 1978, S. 168).

Wie aus Tabelle 14 hervorgeht, ergaben sich bei den erweiterten prozentualen Übereinstimmungen und bei den Finn-Koeffizienten äquivalente Schlußfolgerungen bezüglich der Auswertungsobjektivität der einzelnen Kriterien. Beide Maße sprechen bei 17 Kriterien für eine mindestens akzeptable Urteilskonkordanz der drei Rater. Lediglich für Kriterium 9 (*Nebensächliches*) waren beide Kennwerte unbefriedigend. Zusammenfassend wird geschlußfolgert, daß die Auswertungsobjektivität bei 17 Kriterien akzeptabel war (vgl. Tabelle 14, rechte Spalte). Bei diesen 17 Kriterien konnten daher die Scores der drei Rater für die weiteren statistischen Analysen ohne Bedenken zusammengefaßt werden (arithmetische Mittelung). Nur bei Kriterium 9 war die Auswertungsobjektivität als unzureichend einzustufen, so daß eine Zusammenfassung der Scores der drei Rater nur unter Vorbehalt erfolgen konnte.

Tabelle 14. Überblick über die Ergebnisse der Objektivitätsanalysen, getrennt nach Kriterien und statistischen Bestimmungsmethoden

	Kriterium	Erweit. proz. Übereinst.	Prod.-Mom.-Korrelation	Gewichtetes Kappa	Intraklassen-korrelation	Finn-Koeffizient	Gesamturteil
1.	*Konsistenz*	+	–	–	–	+	+
2.	*Unordnung*	+	–	–	+	+	+
3.	*Details*	+	+	+	+	+	+
4.	*Verknüpfungen*	+	+	+	+	+	+
5.	*Interaktionen*	+	–	–	+	+	+
6.	*Gespräche*	+	+	+	+	+	+
7.	*Komplikationen*	+	–	–	–	+	+
8.	*Ausgefallenes*	+	–	–	+	+	+
9.	*Nebensächliches*	–	–	–	–	–	–
10.	*Unverstandenes*	+	–	–	–	+	+
11.	*Indirektes*	+	–	–	–	+	+
12.	*Eigenseelisches*	+	+	+	+	+	+
13.	*Fremdseelisches*	+	+	+	+	+	+
14.	*Verbesserungen*	+	–	–	+	+	+
15.	*Erinnerungslücken*	+	–	–	+	+	+
16.	*Selbsteinwände*	+	–	–	–	+	+
17.	*Eigenbelastung*	+	–	–	–	+	+
18.	*Fremdentlastung*	+	–	–	+	+	+

Anmerkung: „+“ = Ergebnis spricht für eine akzeptable Auswertungsobjektivität; „–“ = Ergebnis spricht für eine unzureichende Auswertungsobjektivität; grau unterlegte Felder bilden die Grundlage für die Gesamtbeurteilung.

5.2.2 Differenzierung der experimentellen Gruppen

5.2.2.1 Gruppenunterschiede in den Ausprägungen der 18 Glaubhaftigkeitskriterien

Angesichts der insgesamt akzeptablen Auswertungsobjektivität bei der Quantifizierung der einzelnen Glaubhaftigkeitskriterien (s. Abschnitt 5.2.1) wurden die Ratings der drei Beurteiler für die weiteren statistischen Analysen jeweils durch arithmetische Mittelung zu einem Wert zusammengefaßt. Lediglich für Kriterium 9 (*Nebensächliches*), das als einziges keine ausreichende Auswertungsobjektivität aufwies, wurden zusätzliche statistische Analysen vorgenommen, in denen die drei Rater als Einflußvariable berücksichtigt wurden, um so etwaige systematische Unterschiede zwischen den Ratern zu identifizieren (s.u.).

Anhand der gemittelten Werte wurde zunächst überprüft, inwieweit die 18 Glaubhaftigkeitskriterien in der vorliegenden Untersuchung als voneinander unabhängige Variablen betrachtet werden können. Hierfür wurde jedes einzelne Kriterium mit jedem anderen Kriterium paarweise korreliert, wobei sowohl die Produkt-Moment-Korrelationen nach Pearson als auch die nichtparametrischen Rangkorrelationen nach Spearman berechnet wurden. Die Ergebnisse der Korrelationsanalysen sind in den Tabellen F.27 und F.28 im Anhang F zusammengefaßt. Auf der Grundlage der hier analysierten 102 experimentellen Aussagen erwiesen sich 37 der 153 möglichen Interkorrelationen (entspricht 24.18%) als signifikant positiv ($\alpha = .05$, zweiseitiger Test). Ferner zeigten sich drei (entspricht 1.96%) signifikant negative Korrelationen. Dies galt sowohl für die Produkt-Moment- als auch für die Rangkorrelationen (vgl. Anhang F, Tabellen F.27 und F.28). Somit konnten die 18 Glaubhaftigkeitskriterien im vorliegenden Fall nicht als voneinander unabhängig angesehen werden.

Aus diesem Grund wurde zur Überprüfung von Gruppenunterschieden bezüglich der Ausprägungsgrade der 18 Glaubhaftigkeitskriterien zunächst eine einfaktorielle multivariate Varianzanalyse (MANOVA) mit dem Gruppenfaktor Status der aussagenden Person als UV und allen 18 Kriterien als AVn gerechnet. Es zeigte sich ein statistisch bedeutsamer simultaner Effekt des Gruppenfaktors auf die Ausprägungsgrade der 18 Kriterien, PS = 0.872, $F(36,166) = 3.565$, $p < .01$. Neben *Pillais Spurkriterium PS* wurden als weitere multivariate Teststatistiken auch noch *Wilks Likelihood-Quotient* Λ, *Hotellings Spurkriterium T* sowie *Roys größter Eigenwert* berechnet, aus denen sich ebenfalls approximativ F-verteilte Prüfgrößen ableiten lassen. Im vorliegenden Fall ist PS die konservativste Statistik. Die anderen Statistiken erbrachten allerdings äquivalente Resultate. Die Ergebnisse der multivariaten Teststatistiken sind in Tabelle F.29 im Anhang F zusammengefaßt.

Um zu überprüfen, auf welche Gruppenunterschiede der multivariate Effekt in erster Linie zurückging, wurden als Anschlußtests drei MANOVAs mit jeweils nur zwei Gruppen (*Täterinnen* vs. *Zeuginnen*, *Täterinnen* vs. *falsche Zeuginnen*, *Zeuginnen* vs. *falsche Zeuginnen*) gerechnet, wobei jeweils eine Bonferroni-Korrektur des Alpha-Fehlers vorgenommen wurde ($\alpha' = .05/3 = .017$). Die Ergebnisse der drei Anschluß-MANOVAs sind in den Tabellen F.30, F.31 und F.32 im Anhang F zusammengefaßt. Es zeigte sich, daß die *Zeuginnen* sich signifikant sowohl von den *Täterinnen*, PS = 0.545, $F(18,49) = 3.263$, $p < .017$, als auch von den *falschen Zeuginnen*, PS = 0.761, $F(18,49) = 8.690$, $p < .017$, unterschieden (vgl. Anhang F, Tabellen F.30 bzw. F.32). Dagegen verfehlte der Unterschied zwischen den Gruppen *Täterinnen* und *falsche Zeuginnen* knapp die korrigierte Signifikanzgrenze, PS = 0.440, $F(18,49) = 2.137$, $p = .018$ (vgl. Anhang F, Tabelle F.31).

Wenn die in eine MANOVA eingehenden AVn untereinander korreliert sind, hat die Aufklärung eines resultierenden signifikanten multivariaten Effekts, d.h. die Analyse der Bedeutung der einzelnen AVn für die Unterscheidung der untersuchten Stichproben, grundsätzlich mittels einer Diskriminanzanalyse zu erfolgen (vgl. z.B. Bortz, 1999, S. 426, 576). Daher wurde eine Diskriminanzanalyse mit den 18 Glaubhaftigkeitskriterien als Prädiktoren und der Gruppenzugehörigkeit der Pbn (Status der aussagenden Person) als vorherzusagendem Kriterium gerechnet. Die Ergebnisse der Diskriminanzanalyse sind in Tabelle 15 zusammengefaßt.

Tabelle 15. Ergebnisse der Diskriminanzanalyse mit den 18 Glaubhaftigkeitskriterien als Prädiktoren und der Gruppenzugehörigkeit der Pbn als vorherzusagendem Kriterium

Faktor	Eigenwert λ	% der Varianz	Kumulierte %	Kanonische Korrelation
1	1.528	80.7	80.7	.777
2	0.366	19.3	100.0	.517

Signifikanzprüfung

Faktor(en)	Wilks Likelihood-Quotient Λ	χ^2	df	p
1 und 2	.290	112.119	36	.000
2	.732	28.196	17	.043

		Diskriminanzkoeffizienten		Faktorladungen	
	Prädiktor	Faktor 1	Faktor 2	Faktor 1	Faktor 2
1.	*Konsistenz*	0.156	0.045	-.018	-.028
2.	*Unordnung*	-0.237	0.398	.105	.267
3.	*Details*	0.387	-1.161	.389	-.407
4.	*Verknüpfungen*	-0.607	0.263	-.300	-.072
5.	*Interaktionen*	-0.008	-0.133	-.049	-.309
6.	*Gespräche*	0.396	-0.021	.377	-.251
7.	*Komplikationen*	0.041	0.080	-.094	-.002
8.	*Ausgefallenes*	0.341	0.528	.458	.355
9.	*Nebensächliches*	0.503	0.306	.416	.051
10.	*Unverstandenes*	0.096	0.109	.143	.038
11.	*Indirektes*	0.100	0.107	.115	.030
12.	*Eigenseelisches*	-0.462	0.047	-.104	-.034
13.	*Fremdseelisches*	-0.186	-0.030	-.035	-.205
14.	*Verbesserungen*	0.193	0.205	.244	-.059
15.	*Erinnerungslücken*	0.231	0.256	.214	.070
16.	*Selbsteinwände*	-0.084	-0.157	.005	-.085
17.	*Eigenbelastung*	-0.129	-0.053	-.075	.064
18.	*Fremdentlastung*	0.177	0.434	-.006	.380

Wie Tabelle 15 zu entnehmen ist, resultierten zwei Diskriminanzfaktoren, die beide zusammen die experimentellen Gruppen signifikant trennten. Der Anteil des ersten

Faktors am gesamten Diskriminanzpotential betrug 80.7%, durch den zweiten Diskriminanzfaktor wurden 19.3% der Varianz aufgeklärt. Der zweite Diskriminanzfaktor trug auch für sich allein genommen, d.h. bei Nichtberücksichtigung des ersten Diskriminanzfaktors, noch signifikant zur Trennung der Gruppen bei (s. Tabelle 15).

Abbildung 2 veranschaulicht, in welcher Weise die beiden Diskriminanzfaktoren zur Trennung der experimentellen Gruppen beitrugen. Wie Abbildung 2 zu entnehmen ist, resultierten auf dem ersten Diskriminanzfaktor für die *Zeuginnen* fast ausschließlich positive Diskriminanzwerte; der Gruppen-Zentroid (durchschnittlicher Diskriminanzwert innerhalb der Gruppe) betrug hier 1.589. Dagegen ergaben sich für die *falschen Zeuginnen* nahezu nur negative Diskriminanzwerte (Zentroid = -1.370). Die Werte der *Täterinnen* auf Diskriminanzfaktor 1 verteilten sich in etwa gleichmäßig über den positiven und den negativen Bereich (Zentroid = -0.219); die Werteverteilung dieser Gruppe überschnitt sich zudem deutlich mit den Werteverteilungen der *Zeuginnen* und insbesondere der *falschen Zeuginnen* auf Faktor 1. Somit differenzierte der erste Diskriminanzfaktor klar zwischen den Gruppen der *Zeuginnen* und der *falschen Zeuginnen*, wohingegen diese beiden Gruppen, insbesondere die *falschen Zeuginnen*, nur schwach von der Gruppe der *Täterinnen* abgegrenzt werden konnten.

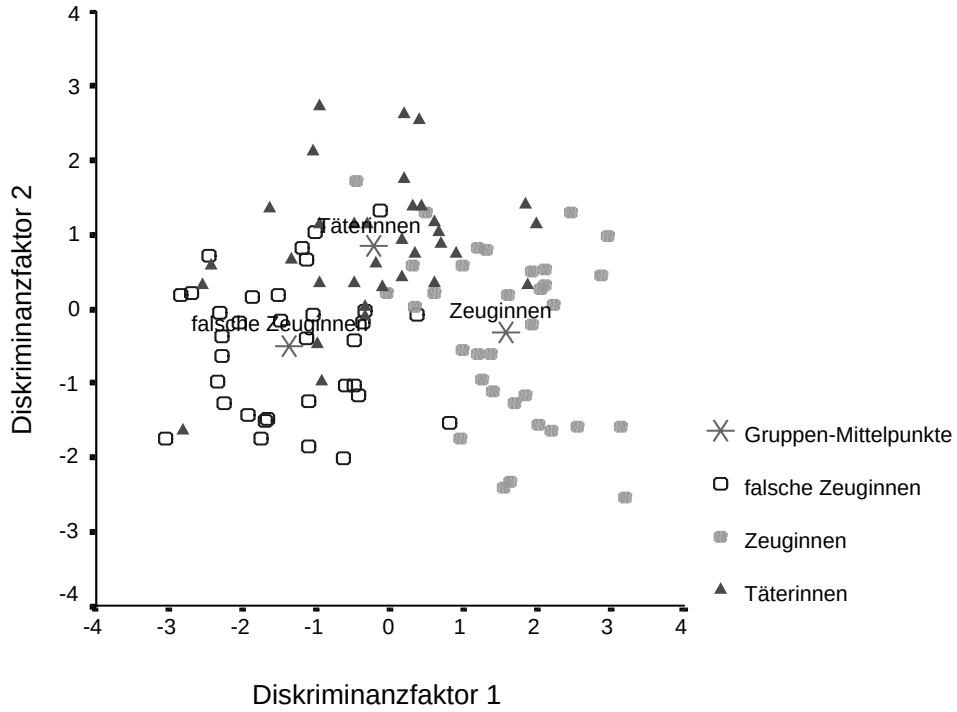


Abbildung 2. Positionen der Probandinnen auf den aus den 18 Glaubhaftigkeitskriterien extrahierten Diskriminanzfaktoren.

Auf Diskriminanzfaktor 2 ergaben sich für die *Täterinnen* weitestgehend positive Werte (Zentroid = 0.836). Dagegen lagen die Werte der *Zeuginnen* und der *falschen Zeuginnen*

auf dem zweiten Diskriminanzfaktor zwar überwiegend im negativen Bereich (Zentroide = -0.325 bzw. -0.511), gleichwohl reichten die Werteverteilungen dieser beiden Gruppen deutlich bis in den positiven Bereich hinein und überschnitten sich stark mit der Werteverteilung der Gruppe *Täterinnen*. Die Werteverteilungen der *Zeuginnen* und *falschen Zeuginnen* auf Faktor 2 waren praktisch nicht unterscheidbar (s. Abbildung 2). Somit trug der zweite Diskriminanzfaktor nicht zur Differenzierung zwischen den *Zeuginnen* und *falschen Zeuginnen* bei, trennte diese beiden Gruppen jedoch schwach von der Gruppe der *Täterinnen*.

Aus der Betrachtung der Diskriminanzkoeffizienten (s. Tabelle 15) ergibt sich, daß für den Diskriminanzfaktor 1 insbesondere die Glaubhaftigkeitskriterien 4 (*Verknüpfungen*), 9 (*Nebensächliches*), 12 (*Eigenseelisches*), 6 (*Gespräche*), 3 (*Details*) und 8 (*Ausgefallenes*) relevant waren. Diese Kriterien trugen also am meisten zur Unterscheidung der experimentellen Gruppen, insbesondere zur Differenzierung der *Zeuginnen* von den *falschen Zeuginnen*, bei. Dabei ist jedoch zu betonen, daß die Kriterien 4 (*Verknüpfungen*) und 12 (*Eigenseelisches*) jeweils mit negativen Gewichtungen in den ersten Diskriminanzfaktor eingingen, d.h. die Aussagen der *Zeuginnen* waren (entgegen der „Undeutsch-Hypothese“) durch relativ geringe Ausprägungen dieser Glaubhaftigkeitskriterien gekennzeichnet, wohingegen die Aussagen der *falschen Zeuginnen* (und in geringerem Ausmaß die der *Täterinnen*) vergleichsweise mehr *Verknüpfungen* und *Eigenseelisches* aufwiesen. Ferner zeichneten sich die Aussagen der *Zeuginnen* gegenüber den Aussagen der *falschen Zeuginnen* (und in geringerem Ausmaß gegenüber den Aussagen der *Täterinnen*) durch mehr *Nebensächliches*, *Gespräche*, *Details* und *Ausgefallenes* aus. Diese diskriminanzanalytischen Schlußfolgerungen sind auch anhand der jeweiligen Gruppenmittelwerte in den einzelnen Glaubhaftigkeitskriterien nachvollziehbar, welche weiter unten in Tabelle 16 aufgelistet sind.

Im zweiten Diskriminanzfaktor wurde Kriterium 3 (*Details*) mit Abstand am stärksten gewichtet (s. Tabelle 15). Somit leistete dieses Kriterium hier den deutlichsten Beitrag zur Differenzierung der *Täterinnen* einerseits von den *Zeuginnen* und *falschen Zeuginnen* andererseits. Die Aussagen der *Täterinnen* zeichneten sich also gegenüber den Aussagen der beiden anderen Gruppen durch relative Detailarmut aus (negative Gewichtung des Kriteriums in Diskriminanzfaktor 2, s. Tabelle 15; bezüglich der Gruppenmittelwerte s. Tabelle 16).

Die Deutung der beiden Diskriminanzfaktoren anhand der jeweiligen Diskriminanzkoeffizienten deckt sich weitgehend mit der Beschreibung der Faktoren anhand der Faktorladungen. Wie Tabelle 15 zu entnehmen ist, wiesen die betragsmäßig höchsten Diskriminanzkoeffizienten (Kriterien 4, 9, 12, 6, 3, 8 bei Faktor 1; Kriterium 3 bei

Faktor 2; s.o.) zum einen die gleichen Vorzeichen auf wie die entsprechenden Faktorladungen. Zum anderen waren fünf der sechs Kriterien mit den höchsten Diskriminanzkoeffizienten bei Faktor 1 (Kriterien 4, 9, 6, 3, 8) zugleich auch die fünf Kriterien mit den höchsten Ladungen auf diesem Faktor. Kriterium 3, das mit Abstand den betragsmäßig höchsten Diskriminanzkoeffizienten bei Faktor 2 aufwies (s.o.), hatte zugleich auch die höchste Ladung auf diesem Faktor (vgl. Tabelle 15).

Im Falle eines statistisch bedeutsamen multivariaten Effekts können neben der Diskriminanzanalyse auch separate univariate Tests Hinweise darauf liefern, welche abhängigen Variablen zur multivariaten Overall-Signifikanz beitragen. Dabei ist allerdings prinzipiell zu beachten, daß die aus den univariaten Tests gezogenen Schlußfolgerungen, sofern die AVn untereinander korrelieren, irreführend sein können, da etwaige Suppressionseffekte unberücksichtigt bleiben (Bortz, 1999, S. 732). Im vorliegenden Fall ist die Durchführung univariater Tests auch deshalb naheliegend, weil die 18 Kriterien theoretisch als „voneinander unabhängige Instrumente zur Beurteilung der Glaubhaftigkeit“ angesehen werden, demzufolge eine Überprüfung der Validität des Kriterienkataloges „auf dem Wege der Überprüfung von [...] Einzelhypothesen zu erfolgen hat“ (Steller et al., 1992, S. 164). Zudem dient die Durchführung univariater Tests der Vergleichbarkeit mit den einschlägig publizierten Studien (vgl. Abschnitt 2.1.6.1), in denen überwiegend auf den univariaten Ansatz zurückgegriffen wurde. Daher wurden 18 einfaktorielle ANOVAs gerechnet. Angesichts der Interkorrelationen zwischen den 18 AVn wurde jedoch eine Adjustierung des Alpha-Fehlers vorgenommen, womit bei einer Durchführung von 18 ANOVAs ein Effekt noch auf dem 5%-Niveau signifikant war, wenn er das Alpha-Niveau von .0028 (= .05/18) unterschritt. Tabelle 16 bietet einen Überblick über die Ergebnisse der 18 ANOVAs bzw. der jeweiligen Anschlußtests für paarweise Gruppenvergleiche. Eine ausführliche Dokumentation der varianzanalytischen Ergebnisse sowie der an signifikante ANOVA-Effekte sich anschließenden Scheffé-Tests findet sich im Anhang F, Tabellen F.33 bzw. F.34.

Wie Tabelle 16 zu entnehmen ist, manifestierten sich in den ANOVAs statistisch bedeutsame Effekte der Gruppenzugehörigkeit auf die Ausprägungen der Glaubhaftigkeitskriterien 3, 4, 6, 8 und 9. Die Scheffé-Tests für paarweise Gruppenvergleiche ergaben, daß die Kriterien 3 (*Details*) und 6 (*Gespräche*) jeweils in den Aussagen der *Zeuginnen* signifikant stärker ausgeprägt waren als in den Aussagen der *Täterinnen* und der *falschen Zeuginnen* (s. Tabelle 16). Hinsichtlich der Ausprägung von Kriterium 9 (*Nebensächliches*) hoben sich die Aussagen der *Zeuginnen* signifikant nur von den Aussagen der *falschen Zeuginnen* ab. Bei Kriterium 8 (*Ausgefallenes*) erzielten neben den *Zeuginnen* auch die *Täterinnen* eine signifikant stärkere Ausprägung als die *falschen Zeuginnen*. Die statistische Signifikanz bei Kriterium 4 (*Verknüpfungen*) beruhte darauf,

daß dieses (entgegen der „Undeutsch-Hypothese“) in den Aussagen der *falschen Zeuginnen* stärker ausgeprägt war als in den Aussagen der *Zeuginnen* (s. Tabelle 16).

Tabelle 16. Mittlere Ausprägungsgrade (M) und Standardabweichungen (SD) der inhaltlichen Glaubhaftigkeitskriterien in den drei experimentellen Gruppen, Ergebnisse der Varianzanalysen (ANOVAs) und paarweisen Gruppenvergleiche (Scheffé-Tests)

		Experimentelle Gruppe						ANOVA		Gruppenvergleich		
		Gruppe 1: *Täterinnen*		Gruppe 2: *Zeuginnen*		Gruppe 3: *Falsche Zeug.*				1 - 2	1 - 3	2 - 3
	Kriterium	M	SD	M	SD	M	SD	*F*(2,99)	p	p	p	p
1.	*Konsistenz*	2.93	0.16	2.93	0.20	2.94	0.15	0.04				
2.	*Unordnung*	0.50	0.58	0.44	0.50	0.26	0.36	2.13				
3.	*Details*	1.51	0.73	2.31	0.70	1.58	0.62	14.43	*	*		*
4.	*Verknüpfungen*	1.15	0.70	0.86	0.55	1.42	0.60	6.89	*			*
5.	*Interaktionen*	0.21	0.48	0.32	0.40	0.41	0.43	1.91				
6.	*Gespräche*	1.48	0.85	2.27	0.70	1.40	0.89	11.87	*	*		*
7.	*Komplikationen*	0.08	0.18	0.05	0.15	0.10	0.19	0.67				
8.	*Ausgefallenes*	0.45	0.35	0.57	0.24	0.17	0.25	18.14	*		*	*
9.	*Nebensächliches*	0.70	0.61	1.07	0.60	0.40	0.38	13.11	*			*
10.	*Unverstandenes*	0.02	0.08	0.04	0.14	0.00	0.00	1.57				
11.	*Indirektes*	0.01	0.06	0.02	0.08	0.00	0.00	1.02				
12.	*Eigenseelisches*	1.20	0.55	1.11	0.59	1.29	0.64	0.84				
13.	*Fremdseelisches*	0.43	0.53	0.53	0.55	0.61	0.59	0.85				
14.	*Verbesserungen*	1.19	0.64	1.46	0.45	1.08	0.51	4.55				
15.	*Erinnerungslück.*	0.19	0.27	0.27	0.35	0.09	0.22	3.56				
16.	*Selbsteinwände*	0.02	0.08	0.03	0.10	0.03	0.10	0.13				
17.	*Eigenbelastung*	0.02	0.11	0.00	0.00	0.02	0.11	0.50				
18.	*Fremdentlastung*	0.36	0.44	0.20	0.35	0.18	0.30	2.61				

Anmerkung: * $p < .0028$ ($\alpha' = .05/18$).

Die Resultate der univariaten Varianzanalysen decken sich also weitestgehend mit den Ergebnissen der Diskriminanzanalyse (s.o.). Lediglich für Kriterium 12 (*Eigenseelisches*), das sich im Rahmen der Diskriminanzanalyse als relativ guter Prädiktor der Gruppenzugehörigkeit erwies (betragsmäßig dritthöchster Diskriminanzkoeffizient in Diskriminanzfaktor 1; s. Tabelle 15), ergab sich kein signifikanter univariater Effekt (s. Tabelle 16). Andererseits steht das Ausbleiben eines signifikanten univariaten Effekts bei Kriterium 12 jedoch im Einklang damit, daß dieses das einzige der sechs Kriterien mit den betragsmäßig höchsten Diskriminanzkoeffizienten in Faktor 1 (Kriterien 4, 9, 12, 6, 3, 8) war, welches nicht zugleich auch eine der höchsten Ladungen auf diesem Faktor aufwies (s.o.).

In Abschnitt 5.2.1 wurde gezeigt, daß die Auswertungsobjektivität von Kriterium 9 (*Nebensächliches*) unbefriedigend ausfiel. Daher wurde neben der einfaktoriellen auch

noch eine zweifaktorielle ANOVA gerechnet, in welcher die drei Rater zusätzlich als Meßwiederholungsfaktor berücksichtigt wurden (keine Mittelwertbildung über die drei Rater). Hierdurch konnte überprüft werden, ob sich die Rater möglicherweise systematisch bei der Beurteilung der drei experimentellen Gruppen hinsichtlich des Ausprägungsgrades von Kriterium 9 unterschieden. Es zeigte sich eine signifikante Rater × Gruppe-Wechselwirkung, $F(4,198) = 12.191$, $\varepsilon = .872$, $p < .01$ (s. genauer Anhang F, Tabelle F.35). Diese ist in Abbildung 3 veranschaulicht. Aus Abbildung 3 geht hervor, daß Rater B und Rater C Kriterium 9 jeweils bei den *Zeuginnen* (M = 1.68 bzw. 1.41, SD = 1.15 bzw. 0.96) als stärker ausgeprägt einstuften als bei den *Täterinnen* (M = 0.53 bzw. 1.09, SD = 0.96 bzw. 0.87) und bei den *falschen Zeuginnen* (M = 0.21 bzw. 0.88, SD = 0.48 bzw. 0.88). Im Gegensatz dazu sah Rater A Kriterium 9 bei den *Täterinnen* (M = 0.47, SD = 0.79) stärker ausgeprägt als bei den zwei anderen Gruppen (jeweils M = 0.12, SD = 0.33).

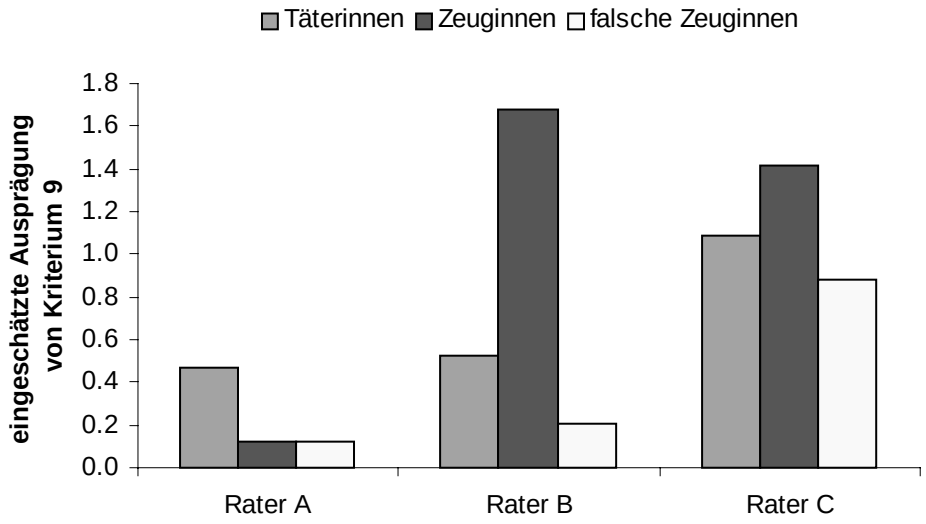


Abbildung 3. Eingeschätzter Ausprägungsgrad von Kriterium 9 (*Nebensächliches*), getrennt nach Ratern und experimentellen Gruppen.

Neben der Wechselwirkung erwiesen sich auch die beiden Haupteffekte als statistisch bedeutsam. Der Haupteffekt des Gruppenfaktors wurde oben bereits erläutert (s. Tabelle 16; vgl. außerdem Tabellen F.33 und F.34 im Anhang F). Der Rater-Haupteffekt, $F(2,198) = 39.786$, $\varepsilon = .872$, $p < .01$, wurde durch Anschlußtests (geringste signifikante Differenz, s. Tabelle F.36 im Anhang F) weiter aufgeklärt. Es zeigte sich, daß Rater C (M = 1.13, SD = 0.92) den Ausprägungsgrad von Kriterium 9 insgesamt signifikant höher einstufte als Rater B (M = 0.80, SD = 1.10) und dieser wiederum signifikant höher als Rater A (M = 0.24, SD = 0.55), jeweils $p < .01$.

5.2.2.2 Gruppenunterschiede in den Gesamtscores

Da es sich bei den einzelnen Kriterien theoretisch um gleich gepolte Indikatoren der Glaubhaftigkeit handelt, die in der vorliegenden Untersuchung zudem anhand gleich abgestufter Ratingskalen erfaßt wurden, wurden die Ausprägungsgrade der 18 Kriterien pro Pb additiv zu einem Gesamtscore zusammengefaßt. Diese Gesamtscores wurden zur Überprüfung von Gruppenunterschieden einer einfaktoriellen ANOVA unterzogen. Dabei zeigte sich ein statistisch bedeutsamer Effekt der Gruppenzugehörigkeit auf die Höhe der Gesamtscores, $F(2,99) = 5.515$, $p < .01$ (vgl. genauer Anhang F, Tabelle F.37). Der Effekt ist in Abbildung 4 graphisch illustriert. Scheffé-Tests ergaben, daß die *Zeuginnen* ($M = 14.49$, $SD = 3.07$) im Durchschnitt signifikant höhere Gesamtscores erzielten als die *Täterinnen* ($M = 12.43$, $SD = 3.85$) und als die *falschen Zeuginnen* ($M = 11.98$, $SD = 2.98$), jeweils $p < .05$, während die beiden letztgenannten Gruppen sich nicht signifikant unterschieden (vgl. Anhang F, Tabelle F.38).

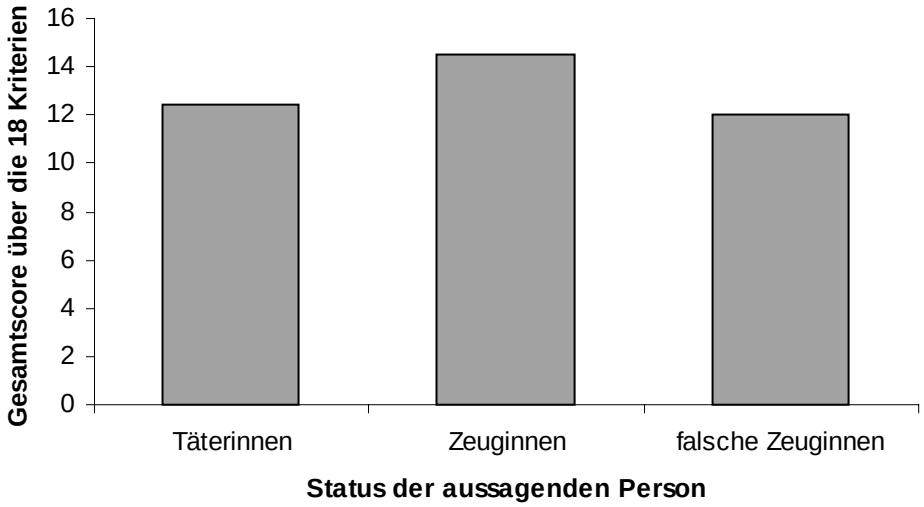


Abbildung 4. Durchschnittliche Gesamtscores über alle 18 Glaubhaftigkeitskriterien in den drei experimentellen Gruppen.

Um zu überprüfen, ob die Höhe der Gesamtscores bzw. die diesbezüglichen Gruppenunterschiede möglicherweise stark mit der differentiellen Beurteilung von Kriterium 9 (*Nebensächliches*) durch die drei Rater zusammenhingen, wurden zusätzlich drei ANOVAs gerechnet, bei denen anstelle des arithmetischen Mittels der drei Rater bei Kriterium 9 jeweils nur der von Rater A bzw. Rater B bzw. Rater C eingeschätzte Ausprägungsgrad von Kriterium 9 in den Gesamtscore einging. Die Ergebnisse der drei ANOVAs bzw. der anschließenden Scheffé-Tests finden sich in den Tabellen F.39 bis F.44 im Anhang F dokumentiert. Es machte keinen wesentlichen Unterschied, wenn anstelle des mittleren Ratings aller drei Rater bei Kriterium 9 das Rating von Rater B oder das von Rater C bei Kriterium 9 in den Gesamtscore einging. Die diesbezüglichen

ANOVAs erbrachten ebenfalls jeweils einen signifikanten Effekt des Faktors Status der aussagenden Person, $F(2,99) = 8.630$, $p < .01$ bzw. $F(2,99) = 5.222$, $p < .01$ (vgl. Anhang F, Tabellen F.41 bzw. F.43), der den Ergebnissen der jeweiligen Scheffé-Tests zufolge darauf beruhte, daß die *Zeuginnen* einen signifikant höheren Gesamtscore erzielten als die *Täterinnen* und die *falschen Zeuginnen* (s. Anhang F, Tabellen F.42 bzw. F.44). Ging jedoch anstelle des mittleren Ratings aller drei Rater bei Kriterium 9 das Rating von Rater A bei Kriterium 9 in den Gesamtscore ein, so unterschieden sich die experimentellen Gruppen im Gesamtscore nur noch marginal, $F(2,99) = 2.829$, $p = .064$ (vgl. Anhang F, Tabelle F.39). In diesem Fall ergaben die Scheffé-Tests lediglich einen tendenziell bedeutsamen Unterschied zwischen den *Zeuginnen* und den *falschen Zeuginnen*, $p = .076$ (s. Anhang F, Tabelle F.40).

5.2.2.3 Gruppenunterschiede in der klinisch-intuitiven Gesamtbeurteilung

Bei der statistischen Analyse der abschließenden klinisch-intuitiven Gesamtbeurteilungen der Aussagen durch die drei Rater wurde neben dem Faktor Status der aussagenden Person auch noch der Faktor „Rater“ als zusätzliche Einflußvariable berücksichtigt. Dies war erforderlich, weil die klinisch-intuitive Gesamtbeurteilung prinzipiell nicht nach standardisierten Richtlinien erfolgt (daher keine diesbezügliche Urteilsschulung möglich), woraus sich die Frage ergibt, ob sich die Rater möglicherweise systematisch in ihrem Urteilsverhalten unterschieden. Daher wurde eine zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Rater (Rater A, B und C) gerechnet. Es resultierte eine signifikante Wechselwirkung der beiden Faktoren, $F(4,198) = 8.243$, $\varepsilon = .874$, $p < .01$ (s. genauer Anhang F, Tabelle F.45). Diese ist in Abbildung 5 graphisch veranschaulicht.

In Abbildung 5 ist ersichtlich, daß die Rater B und C jeweils die Aussagen der *Zeuginnen* ($M = 7.09$ bzw. 8.85, $SD = 2.11$ bzw. 1.35) als glaubhafter beurteilten als die Aussagen der *Täterinnen* ($M = 4.41$ bzw. 6.62, $SD = 2.52$ bzw. 2.91) und der *falschen Zeuginnen* ($M = 5.41$ bzw. 5.50, $SD = 2.30$ bzw. 2.96). Dagegen hielt Rater A nicht nur die Aussagen der *Zeuginnen* ($M = 7.79$, $SD = 1.87$), sondern auch die der *falschen Zeuginnen* ($M = 7.82$, $SD = 1.55$) für glaubhafter als die Aussagen der *Täterinnen* ($M = 6.50$, $SD = 2.43$).

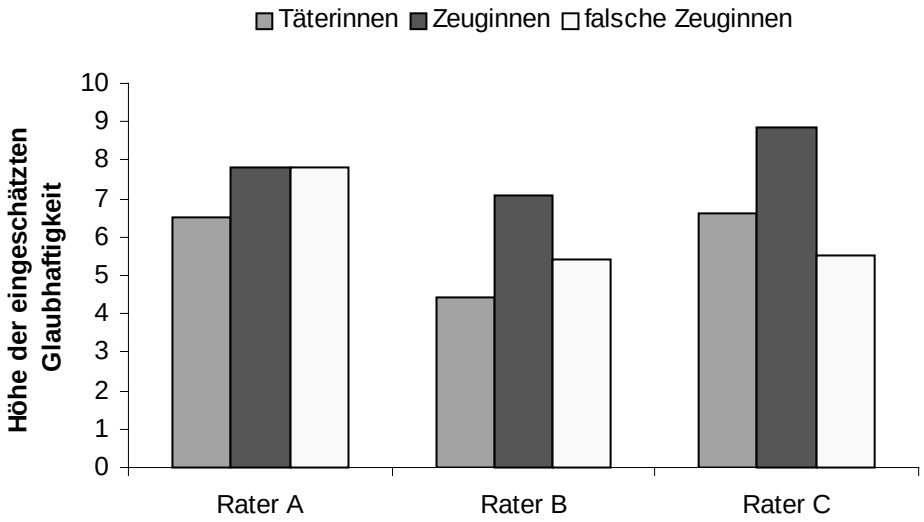


Abbildung 5. Klinisch-intuitive Gesamtbeurteilung der Glaubhaftigkeit, getrennt nach Ratern und experimentellen Gruppen.

Diese Aufschlüsselung der zweifachen Wechselwirkung wird auch durch die Ergebnisse einfaktorieller ANOVAs gestützt, die getrennt für die drei Rater gerechnet wurden. Darin zeigte sich jeweils ein signifikanter Effekt des Faktors Status der aussagenden Person (Rater A: $F(2,99) = 4.943$, $p < .01$; Rater B: $F(2,99) = 11.590$, $p < .01$; Rater C: $F(2,99) = 15.602$, $p < .01$; s. genauer Anhang F, Tabellen F.46, F.48 und F.50). Den Ergebnissen von Scheffé-Tests zufolge beruhte der ANOVA-Effekt bei Rater B darauf, daß die *Zeuginnen* sich signifikant sowohl von den *Täterinnen* als auch von den *falschen Zeuginnen* unterschieden, $p < .01$ bzw. $p < .05$, wohingegen der Unterschied zwischen den letztgenannten Gruppen nicht signifikant wurde (vgl. Anhang F, Tabelle F.49). Auch bei Rater C waren die Unterschiede zwischen *Zeuginnen* und *Täterinnen* sowie zwischen *Zeuginnen* und *falschen Zeuginnen* signifikant, jeweils $p < .01$ (vgl. Anhang F, Tabelle F.51). Dagegen lag bei Rater A kein signifikanter Unterschied zwischen den *Zeuginnen* und *falschen Zeuginnen* vor; beide Gruppen unterschieden sich jedoch von den *Täterinnen*, jeweils $p < .05$ (vgl. Anhang F, Tabelle F.47).[28]

In der zweifaktoriellen ANOVA wurden neben dem Interaktionseffekt auch die beiden Haupteffekte signifikant (Gruppe: $F(2,99) = 14.250$, $p < .01$; Rater: $F(2,198) = 24.108$, $\varepsilon = .874$, $p < .01$; s. Tabelle F.45 im Anhang F). Anschlußtests für paarweise Vergleiche der Faktorstufen (geringste signifikante Differenz) ergaben, daß Rater A ($M = 7.37$, $SD = 2.06$) und Rater C ($M = 6.99$, $SD = 2.86$) die Glaubhaftigkeit der Aussagen insgesamt signifikant höher einstuften als Rater B ($M = 5.64$, $SD = 2.55$), jeweils $p < .01$. Zwi-

[28] Es sei angemerkt, daß die Ergebnisse der drei einfaktoriellen ANOVAs auch bei Anwendung einer Bonferroni-Korrektur ($\alpha' = .05/3 = .0167$) signifikant waren. Auch die Resultate der Scheffé-Tests bezüglich Rater B und Rater C änderten sich nicht nach der Alpha-Adjustierung. Dagegen wurde bei Rater A keiner der Scheffé-Tests auf dem korrigierten Niveau signifikant (s. Anhang F, Tabellen F.46 bis F.51).

schen den Ratern A und C ergab sich kein signifikanter Unterschied (s. Tabelle F.52 im Anhang F). Der Gruppeneffekt beruhte den Ergebnissen von Scheffé-Tests zufolge darauf, daß die Aussagen der *Zeuginnen* (M = 7.91, SD = 1.31) insgesamt für glaubhafter gehalten wurden als die Aussagen der *Täterinnen* (M = 5.84, SD = 1.93) und der *falschen Zeuginnen* (M = 6.25, SD = 1.79), jeweils $p < .01$, wobei der Unterschied zwischen den beiden letztgenannten Gruppen nicht statistisch bedeutsam war (s. Tabelle F.53 im Anhang F).

5.2.2.4 Herauspartialisierung der Kontrollvariablen

Um auszuschließen, daß die oben dargestellten Ergebnisse auf einer Konfundierung der UV Status der aussagenden Person mit den Kontrollvariablen (vgl. Abschnitte 4.3.2 und 5.1) beruhten, wurden zusätzlich auch noch kovarianzanalytische Prozeduren angewendet. Zum einen wurde ergänzend zur oben dargestellten MANOVA (s. Abschnitt 5.2.2.1) eine multivariate Kovarianzanalyse (MANCOVA) gerechnet. Zum anderen wurden ergänzend zu den oben beschriebenen separaten ANOVAs der 18 Glaubhaftigkeitskriterien (s. Abschnitt 5.2.2.1) jeweils univariate Kovarianzanalysen (ANCOVAs) vorgenommen. Schließlich wurden auch noch die über alle 18 Kriterien aufsummierten Gesamtscores – in Ergänzung zur einfaktoriellen ANOVA (s. Abschnitt 5.2.2.2) – kovarianzanalytisch ausgewertet. Als Kovariaten wurden jeweils die Variablen „Motivation, beim Ablegen der Zeugenaussage glaubhaft zu erscheinen“, „subjektiv eingeschätzte persönliche Erfolgswahrscheinlichkeit beim Ablegen der Zeugenaussage“, „Manipulationsmaßnahmen beim Ablegen der Zeugenaussage“, „intellektuelle Begabung“, „Vorerfahrung mit Diebstahl“, „Vorkenntnisse zur Glaubhaftigkeitsbeurteilung“ sowie „frühere Teilnahme an Untersuchungen zur Glaubhaftigkeitsbeurteilung“ in das statistische Design mit aufgenommen (s. Abschnitt 4.3.2).

Der simultane multivariate Effekt der UV Status der aussagenden Person auf die 18 Glaubhaftigkeitskriterien blieb auch nach Herauspartialisierung der Kontrollvariablen in der MANCOVA signifikant, PS = 0.772, $F(36,152) = 2.656$, $p < .01$, wobei allerdings die in der MANCOVA resultierende Effektstärke gegenüber derjenigen in der MANOVA reduziert war ($\eta^2 = .386$ vs. $\eta^2 = .436$; vgl. Anhang F, Tabelle F.54).

Auch die Ergebnisse der separaten univariaten ANCOVAs für die 18 Glaubhaftigkeitskriterien (s. genauer Anhang F, Tabelle F.55) decken sich weitgehend mit den Resultaten der entsprechenden ANOVAs (s. Abbildung 6). Für diejenigen Glaubhaftigkeitskriterien, für die sich in den ANOVAs auf dem 5%-Niveau signifikante Gruppeneffekte ergaben (Kriterien 3, 4, 6, 8, 9, 14 und 15; vgl. Anhang F, Tabelle F.33), fielen diese

Effekte auch nach Herauspartialisierung der Kontrollvariablen in den jeweiligen ANCOVAs signifikant aus, wohingegen sich für die übrigen elf Kriterien auch nach Eliminierung der Kovariaten keine signifikanten Effekte ergaben (s. Anhang F, Tabelle F.55). Ebenso wie bei den ANOVAs führte auch bei den ANCOVAs eine Adjustierung des Alpha-Fehlers ($\alpha' = .05/18 = .0028$) dazu, daß die Gruppeneffekte bei den Kriterien 14 und 15 nicht mehr signifikant wurden. Allerdings verfehlte auch der ANCOVA-Effekt bei Kriterium 4 – im Gegensatz zum entsprechenden ANOVA-Effekt – die korrigierte Signifikanzgrenze. Zudem zeigte sich bei vier der sieben Kriterien, für die sich jeweils ein auf dem 5%-Niveau signifikanter Gruppeneffekt ergab (s.o.), eine Reduktion der Effektstärke nach Herauspartialisierung der Kontrollvariablen (Kriterien 3, 4, 6 und 14; vgl. Anhang F, Tabelle F.55).

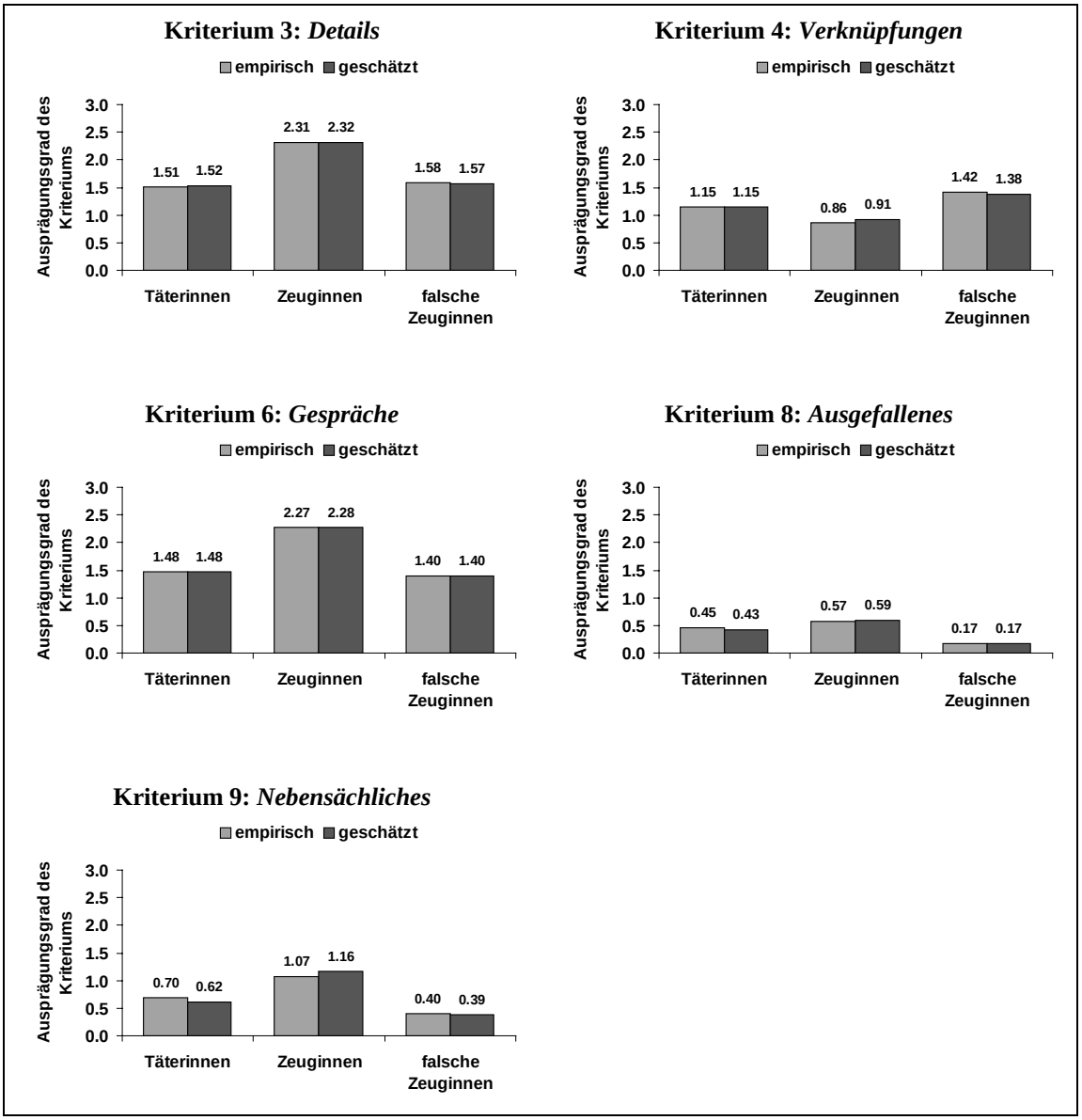


Abbildung 6. Empirische und kovarianzanalytisch geschätzte Gruppenmittelwerte der Glaubhaftigkeitskriterien 3, 4, 6, 8 und 9.

Für diejenigen Kriterien, hinsichtlich derer sich in den ANOVAs auf dem korrigierten Alpha-Niveau ($\alpha' = .0028$) signifikante Gruppeneffekte ergaben (Kriterien 3, 4, 6, 8 und 9; vgl. Tabelle 16), stellt Abbildung 6 den empirischen Gruppenmittelwerten die aufgrund der Kovarianzanalyse geschätzten Durchschnittswerte gegenüber. Wie aus der Abbildung hervorgeht, unterschieden sich die empirischen Werte und die kovarianzanalytischen Schätzwerte kaum voneinander.

Der Effekt der experimentellen Gruppenzugehörigkeit auf die Höhe der Gesamtscores erwies sich auch in der Kovarianzanalyse als statistisch bedeutsam, $F(2,92) = 6.493$, $p < .01$. Die Effektstärke wurde durch die Herauspartialisierung der Kontrollvariablen im Vergleich zur entsprechenden ANOVA sogar erhöht ($\eta^2 = .124$ vs. $\eta^2 = .100$; vgl. Anhang F, Tabelle F.56). In Abbildung 7 ist graphisch illustriert, wie sich die kovarianzanalytisch geschätzten Gruppenmittelwerte der Gesamtscores von den empirischen Durchschnittswerten unterschieden. Es wird deutlich, daß die Unterschiede nur minimal waren.

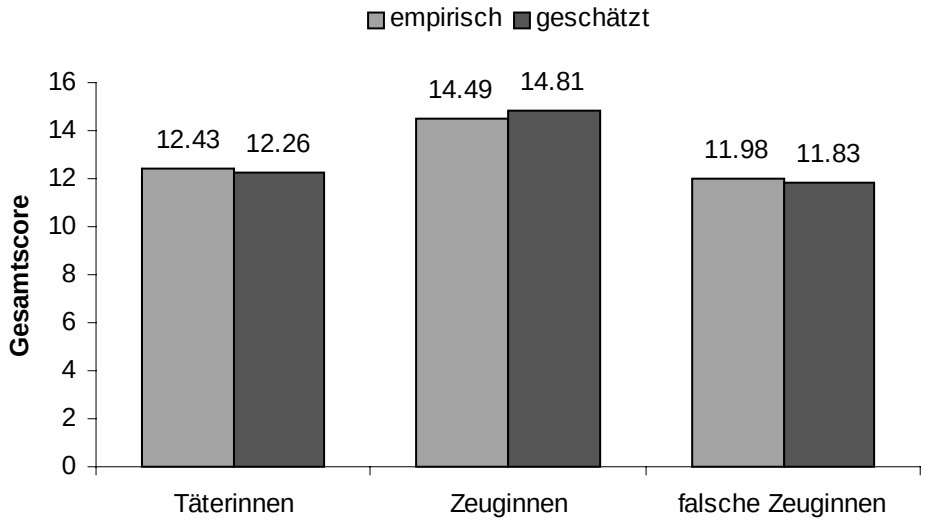


Abbildung 7. Empirische und kovarianzanalytisch geschätzte Gruppenmittelwerte in den über alle 18 Glaubhaftigkeitskriterien gebildeten Gesamtscores.

5.2.3 Treffsicherheit

5.2.3.1 Diskriminanzanalytische Klassifikationsgenauigkeit bei Zugrundelegung der 18 Glaubhaftigkeitskriterien

Um zu überprüfen, mit welcher Treffsicherheit sich auf der Grundlage der 18 Glaubhaftigkeitskriterien die Aussagen der *Zeuginnen* korrekt als glaubhaft bzw. die Schilderungen der *Täterinnen* und der *falschen Zeuginnen* korrekt als unglaubhaft einstufen

ließen, wurde eine diskriminanzanalytische Klassifikationsprozedur durchgeführt. Hierfür wurde die Gruppierungsvariable (vorherzusagendes Kriterium) dichotomisiert, indem die Aussagen der *Täterinnen* und der *falschen Zeuginnen* zur Kategorie „unglaubhaft“ zusammengefaßt wurden, während die Aussagen der *Zeuginnen* die Kategorie „glaubhaft“ repräsentierten. Alle 18 Glaubhaftigkeitskriterien gingen als eigenständige Prädiktoren in die Diskriminanzanalyse ein.

Wird eine diskriminanzanalytische Klassifikationsprozedur mit den gleichen Fällen vorgenommen, anhand derer auch der oder die Diskriminanzfaktoren bzw. die diskriminanzanalytischen Klassifizierungsregeln berechnet wurden, so überschätzen die resultierenden stichprobenbezogenen Trefferquoten in der Regel die für die Population gültigen Hitraten (Bortz, 1999, S. 604). Aus diesem Grund sollte stets eine Kreuzvalidierung der Trefferquoten erfolgen, indem die Klassifizierungsregeln auf eine weitere Stichprobe angewendet werden, die nicht in die Berechnung der Diskriminanzanalyse bzw. der Klassifizierungsregeln einging. Steht – wie im vorliegenden Fall – keine externe Validierungsstichprobe zur Verfügung, kann man alternativ auch die vorhandene Gesamtstichprobe in eine „Konstruktions-“ und eine „Klassifikationsstichprobe“ splitten („Hold-out sample“-Methode, Bortz, 1999, S. 604). Somit bleiben die zu klassifizierenden Personen bei der Berechnung der Klassifikationsregeln unberücksichtigt und eignen sich dementsprechend zur Kreuzvalidierung.

Gemäß der „Hold-out sample“-Methode wurden die Diskriminanzanalyse bzw. die Klassifikationsregeln anhand einer Zufallsstichprobe von 51 Fällen (17 glaubhafte Aussagen [*Zeuginnen*]; 34 unglaubhafte Aussagen [17 *Täterinnen*, 17 *falsche Zeuginnen*]) berechnet, so daß eine Kreuzvalidierung der Klassifikation an den verbleibenden 51 Fällen erfolgen konnte. Tabelle 17 faßt die Klassifizierungsergebnisse der Kreuzvalidierung zusammen. Eine genauere Darstellung der diskriminanzanalytischen Ergebnisse findet sich in Tabelle F.57 im Anhang F. Aus Tabelle 17 geht hervor, daß die Trefferquoten bei den glaubhaften Aussagen der *Zeuginnen* und den unglaubhaften Aussagen der *falschen Zeuginnen* deutlich über dem Zufallsniveau lagen, während die unglaubhaften Aussagen der *Täterinnen* in etwa gleich oft unzutreffend wie zutreffend klassifiziert wurden.

Tabelle 17. Ergebnisse der Kreuzvalidierung der diskriminanzanalytischen Klassifikation der Aussagen der *Täterinnen, Zeuginnen* und *falschen Zeuginnen* als glaubhaft vs. unglaubhaft (alle 18 Glaubhaftigkeitskriterien als eigenständige Prädiktoren)

		Vorhergesagte Gruppenzugehörigkeit	
Tatsächliche Gruppenzugehörigkeit		glaubhaft	unglaubhaft
glaubhaft	*Zeuginnen*	12 (70.6%)	5 (29.4%)
unglaubhaft	*Täterinnen*	8 (47.1%)	9 (52.9%)
	falsche Zeuginnen	5 (29.4%)	12 (70.6%)

Anmerkung: grau unterlegte Felder = Trefferquoten; Gesamttrefferquote = 64.7% (66.2% bei Korrektur der Basisraten von glaubhaften und unglaubhaften Aussagen [29]).

5.2.3.2 Diskriminanzanalytische Klassifikationsgenauigkeit bei Zugrundelegung der Gesamtscores

Die in Abschnitt 5.2.3.1 beschriebene Vorgehensweise täuscht gewissermaßen über die Tatsache hinweg, daß sich nicht alle 18 Glaubhaftigkeitskriterien im Sinne der „Undeutsch-Hypothese“ verhielten (s. Abschnitt 5.2.2.1), d.h. auch solche Kriterien, die sich hypothesenkonträr verhielten, konnten durch eine a posteriori vorgenommene negative Gewichtung (s. Tabelle F.57) zu einer optimalen Trennung zwischen glaubhaften und unglaubhaften Aussagen beitragen. Zum anderen besitzt auch die Gewichtungsprozedur an sich reinen Ex-post-facto-Charakter, d.h. die diskriminanzanalytisch aufgefundene Gewichtungsstruktur für die einzelnen Glaubhaftigkeitskriterien entspricht weder theoretischen Überlegungen noch praktisch-heuristischen Richtlinien.

Um Maßzahlen für die Klassifikationsgenauigkeit zu erhalten, die nicht den oben beschriebenen Verzerrungen unterliegen, wurde eine weitere Diskriminanzanalyse gerechnet, in welche der über alle 18 Glaubhaftigkeitskriterien aufsummierte Gesamtscore als einziger Prädiktor einging. Vorherzusagendes Kriterium war erneut die Glaubhaftigkeit der Aussagen (glaubhaft vs. unglaubhaft). Wie schon in Abschnitt 5.2.3.1 wurde auch hier das „Hold-out sample“-Verfahren angewandt, d.h. die Berechnung des Diskriminanzfaktors bzw. der Klassifikationsregeln erfolgte anhand einer Zufallsstichprobe von 51 Aussagen (17 glaubhafte Aussagen [*Zeuginnen*], 34 unglaubhafte Aussagen [17 *Täterinnen*; 17 *falsche Zeuginnen*]); und die Kreuzvalidierung der Klassifikation wurde an den verbleibenden 51 Fällen vorgenommen. Die „Konstruktions-“ und die „Klassifikationsstichprobe“ waren mit denjenigen in Abschnitt 5.2.3.1 identisch. Die Klassifizie-

[29] Die korrigierte Gesamttrefferquote bezieht sich auf den Fall, daß die Basisraten von glaubhaften und unglaubhaften Aussagen gleich hoch (50%) sind. Die Berechnung erfolgt, indem man die prozentualen Trefferquoten bei den glaubhaften Aussagen (*Zeuginnen*: 70.6%) und unglaubhaften Aussagen (*Täterinnen* und *falsche Zeuginnen* zusammen: 70.6%) arithmetisch mittelt.

rungsergebnisse der Kreuzvalidierung sind in Tabelle 18 dargestellt. Tabelle F.58 im Anhang F bietet eine detaillierte Beschreibung der diskriminanzanalytischen Ergebnisse. Tabelle 18 ist zu entnehmen, daß bei dieser Vorgehensweise die Trefferquote in bezug auf die glaubhaften Aussagen der *Zeuginnen* noch deutlicher über der Zufallswahrscheinlichkeit lag, wohingegen sowohl die unglaubhaften Aussagen der *Täterinnen* als auch die der *falschen Zeuginnen* in etwa genauso oft falsch wie richtig klassifiziert wurden.

Tabelle 18. Ergebnisse der Kreuzvalidierung der diskriminanzanalytischen Klassifikation der Aussagen der *Täterinnen*, *Zeuginnen* und *falschen Zeuginnen* als glaubhaft vs. unglaubhaft (Gesamtscore über alle 18 Glaubhaftigkeitskriterien als einziger Prädiktor)

		Vorhergesagte Gruppenzugehörigkeit	
Tatsächliche Gruppenzugehörigkeit		glaubhaft	unglaubhaft
glaubhaft	*Zeuginnen*	14 (82.4%)	3 (17.6%)
unglaubhaft	*Täterinnen*	9 (52.9%)	8 (47.1%)
	falsche Zeuginnen	8 (47.1%)	9 (52.9%)

Anmerkung: grau unterlegte Felder = Trefferquoten; Gesamttrefferquote = 60.8% (66.2% bei Korrektur der Basisraten von glaubhaften und unglaubhaften Aussagen).

5.2.3.3 Klassifikationsgenauigkeit der klinisch-intuitiven Gesamtbeurteilung

Um zu überprüfen, wie treffsicher die klinisch-intuitive Gesamtbeurteilung der Aussagen durch die drei Rater war, wurden die auf der zehnstufigen Skala (vgl. Anhang C) vorgenommenen Ratings dichotomisiert. Die Skalenstufen 1 bis 5 galten als „Unglaubhaft“-Urteil. Die Skalenstufen 6 bis 10 repräsentierten das Urteil „glaubhaft“. Tabelle 19 gibt die Trefferquoten sowohl getrennt für die drei Rater als auch insgesamt wieder. Es fällt auf, daß die Aussagen der *Zeuginnen* mit einer deutlich überzufälligen Häufigkeit korrekt als glaubhaft diagnostiziert wurden. Dies galt für alle drei Rater. Dagegen lag im Hinblick auf die *falschen Zeuginnen* die diagnostische Treffsicherheit aller drei Rater unter dem Zufallsniveau; Rater A diagnostizierte sogar nahezu alle *falschen Zeuginnen* fälschlich als glaubhaft. Auch die Aussagen der *Täterinnen* wurden alles in allem häufiger falsch als richtig diagnostiziert. Dabei bildete Rater B allerdings eine Ausnahme; seine Treffsicherheit bei dieser Gruppe überstieg deutlich das Zufallsniveau. Aus der rechten äußeren Spalte von Tabelle 19 geht hervor, daß die über alle drei Rater und alle drei Aussagegruppen gebildete (basisratenkorrigierte) Gesamttrefferquote über dem Zufallsniveau lag. Insbesondere die Rater B und C lagen in ihrer jeweiligen (basisratenkorrigierten) Gesamttrefferquote deutlich über dem Zufallsniveau, Rater A hingegen kaum.

Tabelle 19. Treffsicherheit der klinisch-intuitiven Gesamtbeurteilung

	Status der aussagenden Person													
	Täterinnen (unglaubhaft)				*Zeuginnen* (glaubhaft)				*falsche Zeuginnen* (unglaubhaft)				Gesamt-trefferquote	
	Diagnose				Diagnose				Diagnose					
	glaubhaft		unglaubh.		Glaubhaft		unglaubh.		glaubhaft		unglaubh.			
Rater	n	%	n	%	n	%	n	%	n	%	n	%	%	%*
A	21	61.8	13	38.2	30	88.2	4	11.8	31	91.2	3	8.8	45.1	55.9
B	10	29.4	24	70.6	27	79.4	7	20.6	20	58.8	14	41.2	63.7	67.7
C	24	70.6	10	29.4	33	97.1	1	2.9	19	55.9	15	44.1	56.9	66.9
gesamt		53.9		46.1		88.2		11.8		68.6		31.4	55.2	63.5

Anmerkung: grau unterlegte Felder = Trefferquoten; * Gesamttrefferquote bei Korrektur der Basisraten von glaubhaften und unglaubhaften Aussagen.

5.3 Resultate der psychophysiologischen Glaubhaftigkeitsbeurteilung mit dem *Guilty Actions Test*

In Abschnitt 4.4.2 wurde dargestellt, daß bei der Auswertung der Hautleitfähigkeitsreaktionen (SCRs) auf die *GAT*-Items zwei unterschiedliche Quantifizierungsmethoden zur Anwendung kamen. Diese seien an dieser Stelle nochmals in Erinnerung gerufen:
SCR-Quantifizierungsmethode A: SCRs nach Item-Onset mit Latenz 1 – 3 s,
SCR-Quantifizierungsmethode B: SCRs nach Item-Onset mit Latenz 1 – 10 s.
Im folgenden werden die Ergebnisse getrennt nach den zugrundeliegenden SCR-Quantifizierungsmethoden dargestellt.

5.3.1 Resultate bei Anwendung von SCR-Quantifizierungsmethode A

5.3.1.1 Differenzierung der experimentellen Gruppen

5.3.1.1.1 Gruppenunterschiede in den numerischen Scores

Um zu überprüfen, ob die drei experimentellen Gruppen sich in der Höhe der numerischen Scores (vgl. Abschnitt 4.4.2) unterschieden, wurden die auf SCR-Quantifizierungsmethode A basierenden numerischen Scores einer einfaktoriellen ANOVA mit dem Faktor Status der aussagenden Person unterzogen. Dabei manifestierte sich ein statistisch bedeutsamer Effekt des Gruppenfaktors, $F(2,99) = 17.362$, $p < .01$ (s. genauer Anhang F, Tabelle F.59). Der Effekt ist in Abbildung 8 graphisch dargestellt. Scheffé-Tests für paarweise Gruppenvergleiche ergaben, daß die *falschen Zeuginnen* (= *Un-*

schuldige ohne Tatwissen; M = 3.65, SD = 2.65) signifikant niedrigere numerische Scores erzielten als die *Täterinnen* (= *Schuldige*; M = 10.18, SD = 5.72) und die *Zeuginnen* (= *Unschuldige mit Tatwissen*; M = 9.38, SD = 5.90), jeweils $p < .01$. Der Unterschied zwischen den *Täterinnen* und den *Zeuginnen* war nicht signifikant (s. genauer Anhang F, Tabelle F.60).

Abbildung 8. Durchschnittliche numerische *GAT*-Scores in den experimentellen Gruppen (SCR-Quantifizierungsmethode A).

5.3.1.1.2 Gruppenunterschiede hinsichtlich der Stärke der Hautleitfähigkeitsreaktionen auf die relevanten und irrelevanten Items

Um zu überprüfen, ob die drei experimentellen Gruppen sich hinsichtlich ihrer Reaktionsstärken auf die relevanten bzw. irrelevanten *GAT*-Items unterschieden, wurden die logarithmisch (log [x + 1]; s. Abschnitt 4.4.2) transformierten, mit Quantifizierungsmethode A bestimmten SCR-Amplitudenwerte einer zweifaktoriellen ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Itemtyp unterzogen. Der Faktor Itemtyp besaß neben den Ausprägungen relevant und irrelevant auch noch die Ausprägung Puffer. Dieser Faktorstufe gehörten die zehn Items an, die bei den zehn *GAT*-Fragen jeweils am Beginn der Itemsequenz standen und bei der numerischen Auswertung üblicherweise nicht berücksichtigt werden. Für die ANOVA wurden pro Pb jeweils sämtliche SCR-Amplitudenwerte eines Itemtyps durch arithmetische Mittelung zusammengefaßt, d.h. die drei SCR-Magnituden einer Pb resultierten aus der Mittelung der zehn SCRs bei den Pufferitems bzw. der zehn SCRs bei den relevanten Items bzw. der 40 SCRs bei den irrelevanten Items.

Es zeigte sich eine signifikante Wechselwirkung, $F(4,198) = 7.733$, $\varepsilon = .585$, $p < .01$ (s. genauer Anhang F, Tabelle F.61). Diese ist in Abbildung 9 graphisch illustriert. Wie aus Abbildung 9 hervorgeht, manifestierten sich in den SCRs auf die Pufferitems (*Täterinnen*: M = 0.044, SD = 0.056; *Zeuginnen*: M = 0.025, SD = 0.047; *falsche Zeuginnen*: M = 0.028, SD = 0.031) sowie in den SCRs auf die irrelevanten Items (*Täterinnen*: M = 0.046, SD = 0.047; *Zeuginnen*: M = 0.037, SD = 0.073; *falsche Zeuginnen*: M = 0.037, SD = 0.036) keine nennenswerten Gruppenunterschiede. Dagegen zeigten bei den relevanten Items die *Täterinnen* (M = 0.133, SD = 0.120) und die *Zeuginnen* (M = 0.108, SD = 0.168) deutlich stärkere Reaktionen als die *falschen Zeuginnen* (M = 0.039, SD = 0.048). In Abschnitt 2.2.3.1 wurde dargelegt, daß sich aus den Grundannahmen des *GAT* gerichtete Hypothesen hinsichtlich der SCR-Magnituden der drei experimentellen Gruppen bei den relevanten und irrelevanten Items ableiten lassen, nämlich daß bei den relevanten Items die *Täterinnen* (*Schuldige*) stärker als die *Zeuginnen* (*Unschuldige mit Tatwissen*) und diese wiederum stärker als die *falschen Zeuginnen* (*Unschuldige ohne Tatwissen*) reagieren sollen, wohingegen bei den irrelevanten Items keine Gruppenunterschiede zu erwarten sind. Zur genaueren Analyse dieser Hypothesen wurden, getrennt für die relevanten und für die irrelevanten Items, jeweils drei paarweise t-Tests für unabhängige Stichproben gerechnet, wobei aufgrund des A-priori-Charakters der Hypothesen keine Adjustierung des Alpha-Niveaus vorgenommen wurde (Bortz, 1999, S. 262). Die Ergebnisse der t-Tests sind in den Tabellen F.62 bis F.67 im Anhang F dokumentiert. Bei den irrelevanten Items bestanden keine signifikanten Gruppenunterschiede. Bei den relevanten Items unterschieden sich die *falschen Zeuginnen* signifikant sowohl von den *Täterinnen*, $t(66) = 4.257$, $p < .01$, als auch von den *Zeuginnen*, $t(66) = 2.297$, $p < .05$, wohingegen die Differenz zwischen den beiden letztgenannten Gruppen nicht signifikant war.

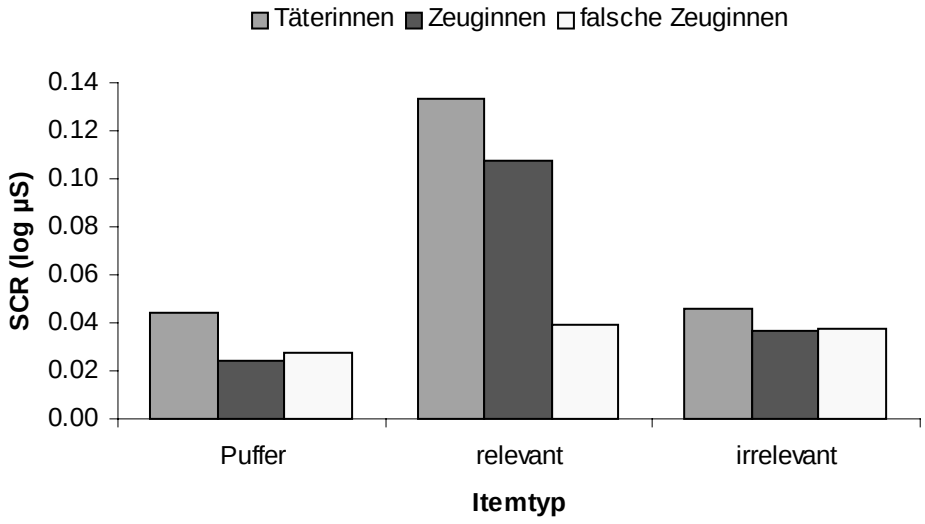


Abbildung 9. SCR-Magnituden, getrennt nach experimentellen Gruppen und Itemtypen (SCR-Quantifizierungsmethode A).

Neben der Interaktion erwies sich auch der Haupteffekt des Faktors Itemtyp als signifikant, $F(2,198) = 38.949$, $\varepsilon = .585$, $p < .01$. Anschlußtests für paarweise Vergleiche der Faktorstufen (geringste signifikante Differenz) ergaben, daß die SCRs bei den relevanten Items (M = 0.093, SD = 0.127) signifikant größer waren als bei den Pufferitems (M = 0.032; SD = 0.046) und bei den irrelevanten Items (M = 0.040; SD = 0.054), jeweils $p < .01$. Zudem waren die SCRs bei den Pufferitems signifikant kleiner als bei den irrelevanten Items, $p < .05$ (s. Anhang F, Tabelle F.68). Der Status der aussagenden Person übte keinen signifikanten Haupteffekt aus (s. Tabelle F.61).

Wie oben erläutert, wurden für die zweifaktorielle ANOVA pro Pb und pro Itemtyp jeweils die bei allen zehn *GAT*-Fragen gemessenen SCR-Amplituden gemittelt. Um nun zu überprüfen, ob die dabei resultierenden Effekte sich auch konsistent bei der isolierten Betrachtung der einzelnen *GAT*-Fragen feststellen ließen, wurden für die zehn *GAT*-Fragen getrennte zweifaktorielle ANOVAs mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Itemtyp gerechnet. In diese ANOVAs gingen also pro Pb der SCR-Amplitudenwert beim Pufferitem, der SCR-Amplitudenwert beim relevanten Item und die mittlere SCR-Amplitude bei den vier irrelevanten Items der entsprechenden *GAT*-Frage ein. Die Ergebnisse der zehn zweifaktoriellen ANOVAs sind in Tabelle 20 zusammengefaßt. Die ausführlichen ANOVA-Ergebnistabellen befinden sich im Anhang F, Tabellen F.69 bis F.78. Zum Vergleich sind in Tabelle 20 außerdem nochmals die Ergebnisse der zweifaktoriellen ANOVA der über alle *GAT*-Fragen gemittelten SCR-Amplitudenwerte (Gesamt-*GAT*) aufgeführt. Es fällt auf, daß der für die vorliegende Fragestellung in erster Linie interessierende Gruppe × Itemtyp-Interaktionseffekt nur in den ersten sieben *GAT*-Fragen konsistent auftrat bzw. in den letzten drei Fragen nicht mehr statistisch signifikant wurde.

Tabelle 20. Gegenüberstellung der Ergebnisse der zweifaktoriellen ANOVAs für den Gesamt-*GAT* und für die einzelnen *GAT*-Fragen (SCR-Quantifizierungsmethode A)

	Effekte		
ANOVA	Gruppe	Itemtyp	Gruppe × Itemtyp
Gesamt-*GAT*	n.s.	**	**
GAT-Frage 1	*	**	**
GAT-Frage 2	*	**	**
GAT-Frage 3	n.s.	**	*
GAT-Frage 4	*	**	**
GAT-Frage 5	n.s.	**	*
GAT-Frage 6	n.s.	**	**
GAT-Frage 7	n.s.	**	**
GAT-Frage 8	n.s.	**	n.s.
GAT-Frage 9	n.s.	**	n.s.
GAT-Frage 10	n.s.	**	n.s.

Anmerkung: * $p < .05$; ** $p < .01$; n.s. = nicht signifikant.

5.3.1.1.3 Herauspartialisierung der Kontrollvariablen

Um auszuschließen, daß die oben dargestellten Resultate der psychophysiologischen Glaubhaftigkeitsbeurteilung mit dem *GAT* auf einer Konfundierung des Gruppenfaktors mit den Kontrollvariablen (vgl. Abschnitte 4.3.2 und 5.1) beruhten, wurden zusätzlich Kovarianzanalysen vorgenommen. In Ergänzung zur einfaktoriellen ANOVA der numerischen *GAT*-Scores (s. Abschnitt 5.3.1.1.1) wurde eine einfaktorielle ANCOVA gerechnet. Weiterhin wurde ergänzend zu der auf den Gesamt-*GAT* bezogenen zweifaktoriellen ANOVA der SCR-Amplituden (s. Abschnitt 5.3.1.1.2) eine entsprechende zweifaktorielle ANCOVA durchgeführt. Als Kovariaten wurden jeweils die Variablen „Elektrodermale Labilität", „Motivation, im *GAT* einen unschuldigen Eindruck zu hinterlassen", „subjektiv eingeschätzte persönliche Erfolgswahrscheinlichkeit im *GAT*", „Manipulationsmaßnahmen im *GAT*", „Vorkenntnisse zur Glaubhaftigkeitsbeurteilung" sowie „frühere Teilnahme an Untersuchungen zur Glaubhaftigkeitsbeurteilung" in das statistische Design mit aufgenommen (s. Abschnitt 4.3.2). Die Kontrollvariable „freies Erinnern bzw. Wiedererkennen der kritischen Tatortdedails" (s. Abschnitt 4.3.2) wurde nicht als Kovariate berücksichtigt, da die diesbezüglich gefundenen Gruppenunterschiede (s. Abschnitt 5.1) ja gerade eine Grundbedingung für die Anwendung des *GAT* sind (s. Abschnitt 2.2.3.1) und dementsprechend auch keine Konfundierung darstellen können.

Der signifikante Effekt der experimentellen Gruppenzugehörigkeit auf die numerischen Scores blieb auch nach Herauspartialisierung der Kovariaten erhalten, $F(2,93) = 15.882$, $p < .01$, wobei sich die Effektstärke nur unwesentlich verminderte ($\eta^2 = .255$ vs. $\eta^2 = .260$; s. genauer Tabelle F.79 im Anhang F). Die auf den Gesamt-*GAT* bezogene zweifaktorielle ANCOVA der SCR-Amplituden erbrachte, ebenso wie die entsprechende zweifaktorielle ANOVA, einen signifikanten Interaktionseffekt der Faktoren Status der aussagenden Person und Itemtyp, $F(4,186) = 6.228$, $\varepsilon = .590$, $p < .01$. Die Effektstärke war gegenüber der ANOVA leicht reduziert ($\eta^2 = .118$ vs. $\eta^2 = .135$). Im Gegensatz zur ANOVA wurde der Itemtyp-Haupteffekt in der ANCOVA nicht signifikant (s. Tabelle F.80 im Anhang F). Abbildung 10 stellt den empirischen Zellenmittelwerten die kovarianzanalytisch geschätzten Durchschnittswerte gegenüber. Die Abbildung macht deutlich, daß die empirischen Werte sich nur unwesentlich von den kovarianzanalytischen Schätzwerten unterschieden.

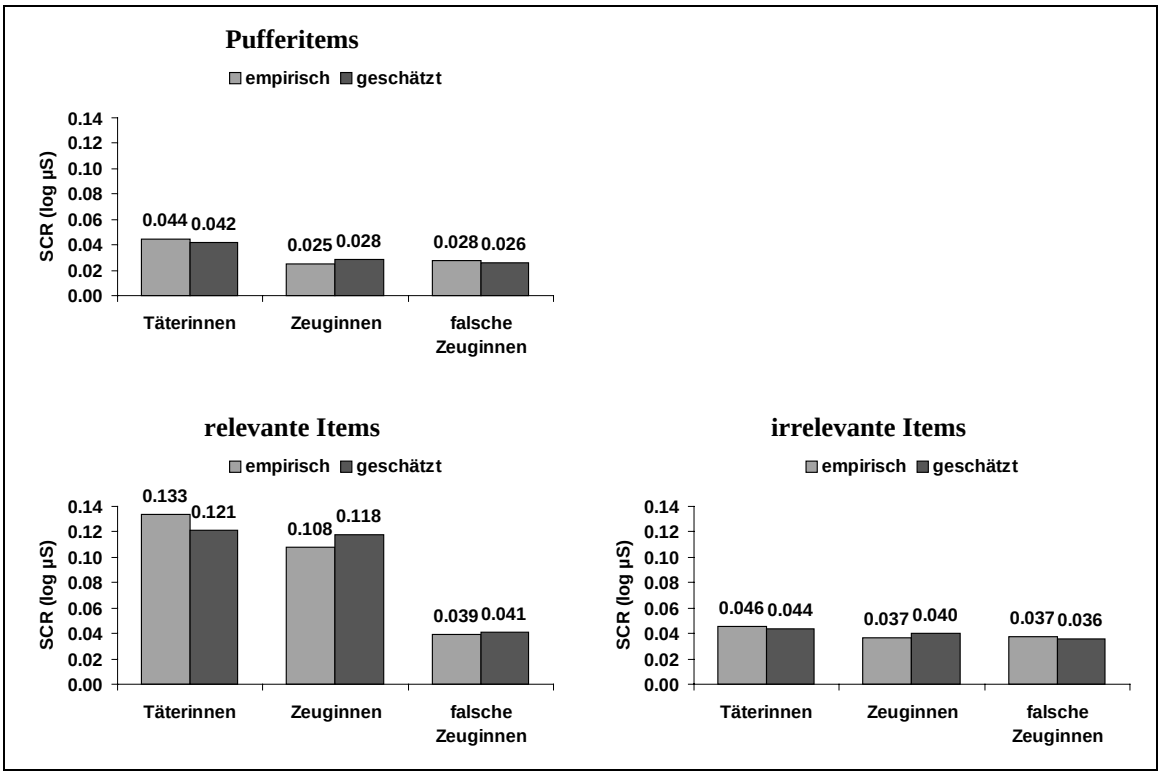


Abbildung 10. Empirische und kovarianzanalytisch geschätzte SCR-Magnituden, getrennt nach Itemtypen und experimentellen Gruppen (SCR-Quantifizierungsmethode A).

5.3.1.2 Treffsicherheit

5.3.1.2.1 Trefferquoten bei Zugrundelegung der A-priori-Entscheidungsregel für die numerischen Scores

In Anlehnung an die bisher durchgeführten Studien zum *GAT* (s. Abschnitt 2.2.3.2) wurde zunächst überprüft, welche Trefferquoten sich unter Verwendung des numerischen Auswertungssystems nach Lykken (1959; s. Abschnitt 2.2.2.1) erzielen ließen. Als Cuttoff-Wert für die diagnostische Urteilsbildung wurde die Punktzahl 10 festgelegt, d.h. die Diagnose „unschuldig“ wurde getroffen, wenn eine Pb bis zu 10 Punkte erzielte. Ergaben sich jedoch 11 oder mehr Punkte (also mehr als die Hälfte der möglichen Punkte; vgl. Abschnitt 2.2.2.1), so lautete die Diagnose auf „schuldig“. Tabelle 21 faßt die Treffer- bzw. Fehlerquoten zusammen. Es sei jedoch daran erinnert, daß diese – ursprünglich für den *TWT* entwickelte – numerische Auswertungsmethode der Logik des *GAT* grundsätzlich nicht gerecht wird (s. Abschnitt 2.2.3.3). Damit zusammenhängend sind auch die Diagnosebezeichnungen „unschuldig“ vs. „schuldig“ gewissermaßen willkürlich, da die durch den *GAT* intendierte dritte Diagnosekategorie „unschuldig mit Tatwissen“ übergangen wird. Läßt man diese Einschränkungen einmal außer acht, so fällt an Tabelle 21 auf, daß alle *Unschuldigen ohne Tatwissen* richtig klassifiziert wur-

den, daß jedoch die *Schuldigen* in etwa genauso oft falsch wie richtig eingeordnet wurden. Ferner ist zu beachten, daß die *Unschuldigen mit Tatwissen* in etwa genauso oft als „schuldig“ klassifiziert wurden wie die *Schuldigen*.

Tabelle 21. Treffsicherheit der anhand der numerischen Scores und auf Basis der A-priori-Entscheidungsregel vorgenommenen diagnostischen Urteile (SCR-Quantifizierungsmethode A)

	Diagnose	
	schuldig	unschuldig
Täterinnen (*Schuldige*)	16 (47.1%)	18 (52.9%)
Zeuginnen (*Unschuldige mit Tatwissen*)	15 (44.1%)	19 (55.9%)
falsche Zeuginnen (*Unschuldige ohne Tatwissen*)	– –	34 (100%)

Anmerkung: grau unterlegte Felder = Trefferquoten.

5.3.1.2.2 Diskriminanzanalytische Klassifikationsgenauigkeit bei Zugrundelegung der numerischen Scores

Die in Abschnitt 5.3.1.2.1 vorgenommene diagnostische Urteilsbildung anhand der A-priori-Entscheidungsregel sieht nur zwei Urteilskategorien vor. Dagegen sollten mit dem *GAT* grundsätzlich alle drei experimentellen Gruppenzugehörigkeiten diagnostizierbar sein. Um die Klassifikationsgenauigkeit bei Verwendung aller drei Urteilskategorien zu bestimmen, wurden die numerischen Scores einer diskriminanzanalytischen Klassifikationsprozedur unterzogen, wobei das vorherzusagende Kriterium der Gruppierungsvariable entsprach, also die drei Ausprägungen „schuldig“, „unschuldig mit Tatwissen“ und „unschuldig ohne Tatwissen“ enthielt.

Die Berechnung der diskriminanzanalytischen Klassifikationsregeln erfolgte anhand einer Zufallsstichprobe von 51 Pbn (jeweils 17 *Täterinnen* [*Schuldige*], *Zeuginnen* [*Unschuldige mit Tatwissen*] bzw. *falsche Zeuginnen* [*Unschuldige ohne Tatwissen*]). Hierbei handelte es sich um die gleiche Zufallsstichprobe, die auch bei den diskriminanzanalytischen Klassifikationsprozeduren im Rahmen der inhaltsorientierten Glaubhaftigkeitsbeurteilung herangezogen wurde (vgl. Abschnitt 5.2.3.1 bzw. 5.2.3.2). Erneut wurde die diskriminanzanalytische Klassifikation anhand der verbleibenden 51 Fälle kreuzvalidiert („Hold-out sample“-Methode; s. Abschnitt 5.2.3.1). Die Klassifizierungsergebnisse der Kreuzvalidierung sind in Tabelle 22 dargestellt. Eine detaillierte Erläuterung der Diskriminanzanalyse findet sich in Tabelle F.81 im Anhang F. Aus Tabelle 22 geht hervor, daß die *Unschuldigen ohne Tatwissen* am häufigsten richtig klassifiziert

wurden. Die Trefferquote lag hier deutlich über dem Zufallsniveau von 33.3%. Fehlklassifikationen in dieser Gruppe ergaben sich ausschließlich in Form der Diagnose „unschuldig mit Tatwissen“. Die Trefferquoten bei den beiden anderen Gruppen lagen jeweils unter dem Zufallsniveau von 33.3%. Bei den *Schuldigen* ergaben sich Fehlklassifikationen in erster Linie in Form der Diagnose „unschuldig mit Tatwissen“. Die *Unschuldigen mit Tatwissen* wurden vornehmlich als „unschuldig ohne Tatwissen“, häufig aber auch als „schuldig“ fehlklassifiziert.

Tabelle 22. Ergebnisse der Kreuzvalidierung der diskriminanzanalytischen Klassifikation der *Täterinnen*, *Zeuginnen* und *falschen Zeuginnen* als schuldig, unschuldig mit Tatwissen bzw. unschuldig ohne Tatwissen: numerischer Score als Prädiktor (SCR-Quantifizierungsmethode A)

	vorhergesagte Gruppenzugehörigkeit		
tatsächliche Gruppenzugehörigkeit	schuldig	unschuldig mit Tatwissen	unschuldig ohne Tatwissen
Täterinnen (*Schuldige*)	4 (23.5%)	10 (58.8%)	3 (17.6%)
Zeuginnen (*Unschuldige mit Tatwissen*)	5 (29.4%)	5 (29.4 %)	7 (41.2%)
falsche Zeuginnen (*Unschuldige ohne Tatwissen*)	– –	3 (17.6%)	14 (82.4%)

Anmerkung: grau unterlegte Felder = Trefferquoten. Gesamttrefferquote = 45.1%.

5.3.1.2.3 Diskriminanzanalytische Klassifikationsgenauigkeit bei Zugrundelegung der intraindividuellen Reaktionsstärkedifferenzen zwischen relevanten und irrelevanten Items

Die in Abschnitt 5.3.1.2.2 angewandte Klassifizierungsmethode berücksichtigt zwar alle drei Urteilskategorien; gleichwohl ist auch hieran zu bemängeln, daß die numerische Auswertungsmethode prinzipiell nicht geeignet ist, das Ausmaß der intraindividuellen Differenz zwischen den Reaktionsstärken auf die relevanten vs. irrelevanten *GAT*-Items zu quantifizieren. Damit zusammenhängend ist die numerische Auswertungsmethode auch nicht geeignet, etwaige differentielle physiologische Reaktionsweisen von *Schuldigen* vs. *Unschuldigen mit Tatwissen* abzubilden (s. Abschnitt 2.2.3.3). Letztlich hat dies zur Konsequenz, daß es grundsätzlich nicht möglich ist, auf der Grundlage der numerischen Scores präzise die Gruppenzugehörigkeit von *Schuldigen* vs. *Unschuldigen mit Tatwissen* vorherzusagen.

Um diese Unzulänglichkeit zu umgehen, wurde eine weitere diskriminanzanalytische Klassifikationsprozedur vorgenommen, in welcher die Gruppenzugehörigkeit der Pbn auf der Grundlage von deren intraindividuellen SCR-Magnituden-Differenzen bei den

relevanten vs. irrelevanten *GAT*-Items vorhergesagt wurde. Es wurde also pro Pb die intraindividuelle SCR-Magnitude bei den irrelevanten Items von der intraindividuellen SCR-Magnitude bei den relevanten Items subtrahiert. Der resultierende Differenzwert diente dann im Rahmen der Diskriminanzanalyse als Prädiktor der experimentellen Gruppenzugehörigkeit. Erneut wurden die diskriminanzanalytischen Klassifikationsregeln nach der „Hold-out sample"-Methode (vgl. Abschnitt 5.2.3.1) berechnet bzw. kreuzvalidiert. Die „Konstruktions-" und die „Klassifikationsstichprobe" waren mit denjenigen in Abschnitt 5.2.3.1 bzw. 5.2.3.2 bzw. 5.3.1.2.2 identisch. In Tabelle 23 sind die Treffer- bzw. Fehlerquoten der Kreuzvalidierung abgetragen. Nähere Angaben zur Diskriminanzanalyse finden sich in Tabelle F.82 im Anhang F. An Tabelle 23 fällt auf, daß keine Pb als „unschuldig mit Tatwissen" klassifiziert wurde. Damit einhergehend wurden sämtliche *Unschuldigen mit Tatwissen* fehlklassifiziert, und zwar überwiegend als „unschuldig ohne Tatwissen". Die höchste Trefferquote ergab sich für die *Unschuldigen ohne Tatwissen*. Auch bei den *Schuldigen* lag die Trefferquote deutlich über dem Zufallsniveau von 33.3%, allerdings wurde hier auch ein beträchtlicher Anteil als „unschuldig ohne Tatwissen" fehlklassifiziert. Hervorzuheben ist ferner, daß von den *Unschuldigen mit Tatwissen* deutlich weniger als „schuldig" klassifiziert wurden als von den *Schuldigen*.

Tabelle 23. Ergebnisse der Kreuzvalidierung der diskriminanzanalytischen Klassifikation der *Täterinnen*, *Zeuginnen* und *falschen Zeuginnen* als schuldig, unschuldig mit Tatwissen bzw. unschuldig ohne Tatwissen: intraindividuelle SCR-Magnituden-Differenz zwischen relevanten und irrelevanten Items als Prädiktor (SCR-Quantifizierungsmethode A)

	vorhergesagte Gruppenzugehörigkeit		
tatsächliche Gruppenzugehörigkeit	schuldig	unschuldig mit Tatwissen	unschuldig ohne Tatwissen
Täterinnen (*Schuldige*)	11 (64.7%)	– –	6 (35.3%)
Zeuginnen (*Unschuldige mit Tatwissen*)	6 (35.3%)	– –	11 (64.7%)
falsche Zeuginnen (*Unschuldige ohne Tatwissen*)	2 (11.8%)	– –	15 (88.2%)

Anmerkung: grau unterlegte Felder = Trefferquoten. Gesamttrefferquote = 51.0%.

5.3.2 Resultate bei Anwendung von SCR-Quantifizierungsmethode B

5.3.2.1 Differenzierung der experimentellen Gruppen

5.3.2.1.1 Gruppenunterschiede in den numerischen Scores

Auch bei Anwendung von SCR-Quantifizierungsmethode B ergab eine einfaktorielle ANOVA der numerischen Scores einen signifikanten Effekt des Faktors Status der aussagenden Person, $F(2,99) = 15.166$, $p < .01$ (s. genauer Tabelle F.83 im Anhang F). Die Gruppenunterschiede sind in Abbildung 11 veranschaulicht. Scheffé-Tests ergaben, daß sowohl die *Täterinnen* (M = 9.88, SD = 5.57) als auch die *Zeuginnen* (M = 9.29, SD = 5.84) signifikant höhere numerische Scores erzielten als die *falschen Zeuginnen* (M = 3.94, SD = 2.63), jeweils $p < .01$. Die Differenz zwischen den *Täterinnen* und *Zeuginnen* war nicht statistisch bedeutsam (vgl. Anhang F, Tabelle F.84).

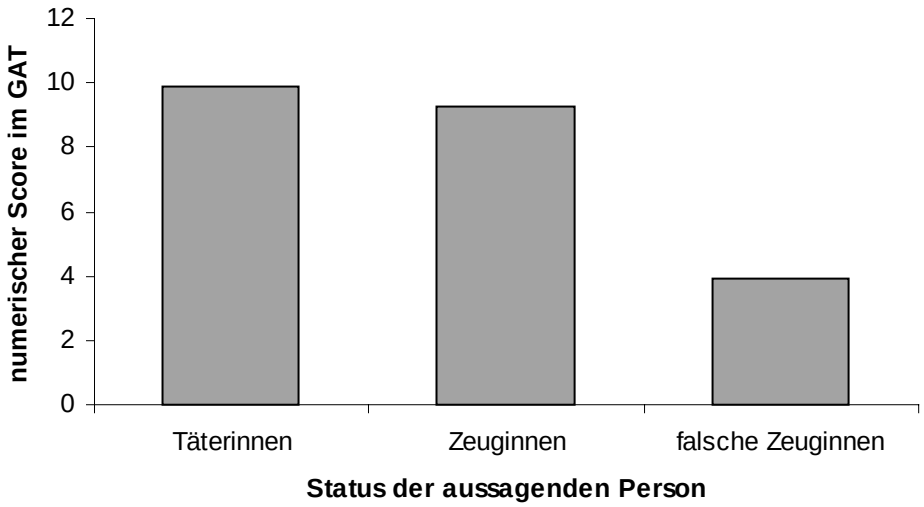


Abbildung 11. Durchschnittliche numerische *GAT*-Scores in den experimentellen Gruppen (SCR-Quantifizierungsmethode B).

5.3.2.1.2 Gruppenunterschiede hinsichtlich der Stärke der Hautleitfähigkeitsreaktionen auf die relevanten und irrelevanten Items

Analog zu der in Abschnitt 5.3.1.1.2 beschriebenen Vorgehensweise wurden auch die mit Quantifizierungsmethode B bestimmten SCR-Amplitudenwerte einer zweifaktoriellen ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Itemtyp unterzogen. Hierbei zeigte sich ebenfalls eine signifikante Interaktion beider Faktoren, $F(4,198) = 8.567$, $\varepsilon = .592$, $p < .01$ (s. genauer Anhang F, Tabelle F.85), die in Abbildung 12 graphisch illustriert ist. Aus Abbildung 12 geht her-

vor, daß bei den Pufferitems (*Täterinnen*: M = 0.061, SD = 0.061; *Zeuginnen*: M = 0.043, SD = 0.071; *falsche Zeuginnen*: M = 0.042, SD = 0.047) und bei den irrelevanten Items (*Täterinnen*: M = 0.060, SD = 0.058; *Zeuginnen*: M = 0.048, SD = 0.083; *falsche Zeuginnen*: M = 0.050, SD = 0.045) keine deutlichen Unterschiede zwischen den experimentellen Gruppen bestanden, daß sich jedoch bei den relevanten Items die *Täterinnen* (M = 0.140, SD = 0.122) und die *Zeuginnen* (M = 0.112, SD = 0.167) klar von den *falschen Zeuginnen* (M = 0.049, SD = 0.053) abhoben. Auch hier wurden analog zu der Vorgehensweise in Abschnitt 5.3.1.1.2 t-Tests zur Überprüfung der a priori formulierten Einzelvergleichshypothesen berechnet (s. Tabellen F.86 bis F.91 im Anhang F). Bei den irrelevanten Items zeigten sich keine signifikanten Gruppenunterschiede. Bei den relevanten Items unterschieden sich die *falschen Zeuginnen* sowohl von den *Täterinnen* als auch von den *Zeuginnen* signifikant, $p < .01$ bzw. $p < .05$, während der Unterschied zwischen den beiden letzteren Gruppen keine statistische Bedeutsamkeit erlangte.

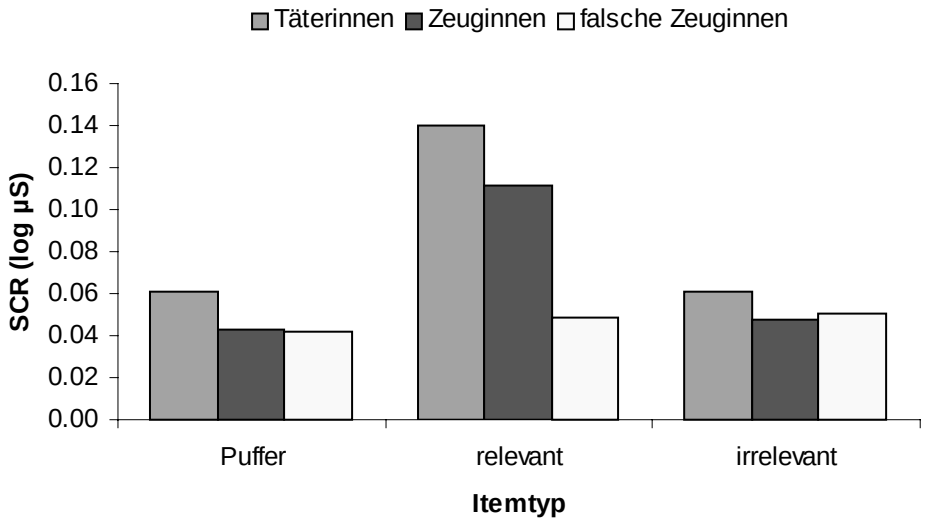


Abbildung 12. SCR-Magnituden, getrennt nach experimentellen Gruppen und Itemtypen (SCR-Quantifizierungsmethode B).

Auch hier zeigte sich neben dem Wechselwirkungseffekt ein signifikanter Haupteffekt des Meßwiederholungsfaktors, $F(2,198) = 37.486$, $\varepsilon = .592$, $p < .01$, wohingegen der Haupteffekt des Gruppenfaktors nicht signifikant war (s. Anhang F, Tabelle F.85). Anschlußtests (geringste signifikante Differenz) zeigten, daß die SCRs bei den relevanten Items (M = 0.100; SD = 0.128) signifikant größer waren als bei den Pufferitems (M = 0.049; SD = 0.060) und den irrelevanten Items (M = 0.053; SD = 0.063), jeweils $p < .01$. Zwischen Pufferitems und irrelevanten Items bestand kein signifikanter Unterschied (s. Anhang F, Tabelle F.92).

Analog zu der in Abschnitt 5.3.1.1.2 beschriebenen Vorgehensweise wurden auch die mit Quantifizierungsmethode B ermittelten SCR-Amplituden separaten zweifaktoriellen

ANOVAs für die zehn *GAT*-Fragen unterzogen, um zu überprüfen, ob die auf den Gesamt-*GAT* bezogenen Effekte sich auch konsistent bei isolierter Betrachtung der einzelnen *GAT*-Fragen manifestierten. Tabelle 24 stellt den Effekten aus dem Gesamt-*GAT* diejenigen aus den einzelnen *GAT*-Fragen gegenüber. Die ausführlichen Ergebnistabellen zu den auf einzelne *GAT*-Fragen bezogenen ANOVAs sind im Anhang F, Tabellen F.93 bis F.102 nachzulesen. Tabelle 24 verdeutlicht, daß die im vorliegenden Zusammenhang primär interessierende Wechselwirkung zwischen Gruppe und Itemtyp sich nur in den ersten sechs *GAT*-Fragen konsistent manifestierte. Dagegen wurde die Interaktion in den Fragen 7 bis 10 nicht mehr statistisch bedeutsam.

Tabelle 24. Gegenüberstellung der Ergebnisse der zweifaktoriellen ANOVAs für den Gesamt-*GAT* und für die einzelnen *GAT*-Fragen (SCR-Quantifizierungsmethode B)

	Effekte		
ANOVA	Gruppe	Itemtyp	Gruppe × Itemtyp
Gesamt-*GAT*	n.s.	**	**
GAT-Frage 1	*	**	**
GAT-Frage 2	*	**	**
GAT-Frage 3	n.s.	**	**
GAT-Frage 4	*	**	**
GAT-Frage 5	n.s.	**	**
GAT-Frage 6	n.s.	**	**
GAT-Frage 7	n.s.	**	n.s.
GAT-Frage 8	n.s.	*	n.s.
GAT-Frage 9	n.s.	**	n.s.
GAT-Frage 10	n.s.	**	n.s.

Anmerkung: * $p < .05$; ** $p < .01$; n.s. = nicht signifikant.

5.3.2.1.3 Herauspartialisierung der Kontrollvariablen

Auch die unter Anwendung von SCR-Quantifizierungsmethode B erzielten Resultate wurden auf mögliche Konfundierungen hin überprüft, indem mittels kovarianzanalytischer Methoden die Einflüsse der Kontrollvariablen herauspartialisiert wurden (vgl. Abschnitt 5.3.1.1.3). Die in Ergänzung zur einfaktoriellen ANOVA der numerischen *GAT*-Scores (s. Abschnitt 5.3.2.1.1) vorgenommene ANCOVA erbrachte ebenfalls einen signifikanten Gruppeneffekt, $F(2,93) = 15.452$, $p < .01$. Die Effektstärke wurde durch die Eliminierung der Kovariaten-Effekte leicht erhöht ($\eta^2 = .249$ vs. $\eta^2 = .235$; s. genauer Tabelle F.103 im Anhang F). Auch die Signifikanz des Itemtyp × Gruppe-Wechselwirkungseffekts auf die SCR-Amplituden blieb nach der Eliminierung der Kovariateneffekte in der zweifaktoriellen ANCOVA erhalten, $F(4,186) = 6.707$, $\varepsilon = .599$, $p < .01$. Dabei war die Effektstärke in der ANCOVA gegenüber derjenigen in der ANOVA leicht reduziert ($\eta^2 = .126$ vs. $\eta^2 = .148$). Im Gegensatz zur zweifaktoriellen

ANOVA erwies sich der Haupteffekt des Itemtyps in der zweifaktoriellen ANCOVA als nicht signifikant (s. Tabelle F.104 im Anhang F). Abbildung 13 bietet eine graphische Gegenüberstellung der empirischen und der kovarianzanalytisch geschätzten Zellenmittelwerte. Es wird deutlich, daß die empirischen Magnitudenwerte sich nur unwesentlich von den kovarianzanalytischen Schätzungen unterschieden.

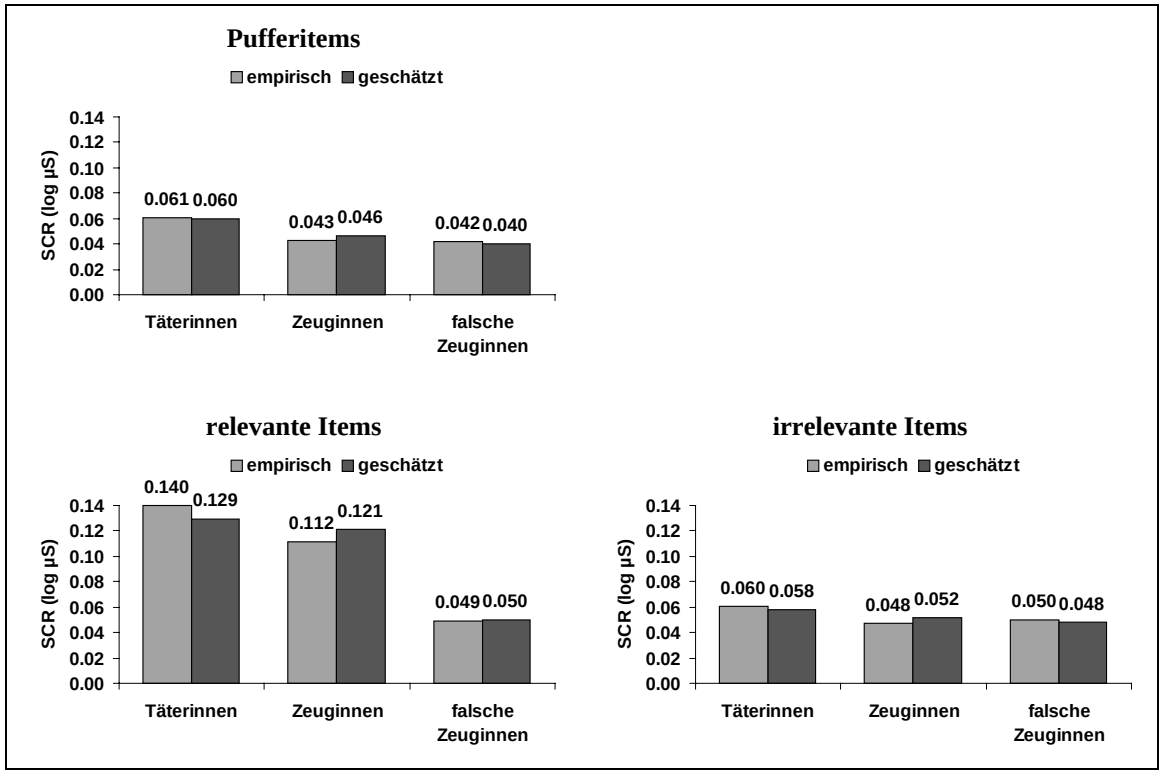


Abbildung 13. Empirische und kovarianzanalytisch geschätzte SCR-Magnituden, getrennt nach Itemtypen und experimentellen Gruppen (SCR-Quantifizierungsmethode B).

5.3.2.2 Treffsicherheit

Die Analyse der Treffsicherheit des *GAT* bei Zugrundelegung von SCR-Quantifizierungsmethode B erfolgte analog zur Treffsicherheitsanalyse bei Zugrundelegung von Quantifizierungsmethode A (vgl. Abschnitt 5.3.1.2.1, 5.3.1.2.2 und 5.3.1.2.3).

5.3.2.2.1 Trefferquoten bei Zugrundelegung der A-priori-Entscheidungsregel für die numerischen Scores

Die Treffer- bzw. Fehlerquoten, die sich bei Anwendung der A-priori-Entscheidungsregel für die numerischen Scores ergaben, sind in Tabelle 25 zusammengefaßt. Wie aus Tabelle 25 hervorgeht, wurden alle *Unschuldigen ohne Tatwissen* korrekt diagnostiziert.

Bei den *Unschuldigen mit Tatwissen* bewegte sich die Trefferquote auf dem Zufallsniveau, bei den *Schuldigen* sogar deutlich darunter. Bemerkenswert ist, daß die *Unschuldigen mit Tatwissen* häufiger als „schuldig“ diagnostiziert wurden als die *Schuldigen*.

Tabelle 25. Treffsicherheit der anhand der numerischen Scores und auf Basis der A-priori-Entscheidungsregel vorgenommenen diagnostischen Urteile (SCR-Quantifizierungsmethode B)

	Diagnose	
	schuldig	unschuldig
Täterinnen (*Schuldige*)	13 (38.2%)	21 (61.8%)
Zeuginnen (*Unschuldige mit Tatwissen*)	16 (47.1%)	18 (52.9%)
falsche Zeuginnen (*Unschuldige ohne Tatwissen*)	– –	34 (100%)

Anmerkung: grau unterlegte Felder = Trefferquoten.

5.3.2.2.2 Diskriminanzanalytische Klassifikationsgenauigkeit bei Zugrundelegung der numerischen Scores

Tabelle 26 gibt die diskriminanzanalytische Klassifikationsgenauigkeit (Kreuzvalidierung nach der „Hold-out sample“-Methode; vgl. Abschnitt 5.2.3.1) wieder, wenn der numerische Score als Prädiktor der Gruppenzugehörigkeit verwendet wurde. Die entsprechende Diskriminanzanalyse ist in Tabelle F.105 im Anhang F genauer erläutert.

Tabelle 26. Ergebnisse der Kreuzvalidierung der diskriminanzanalytischen Klassifikation der *Täterinnen*, *Zeuginnen* und *falschen Zeuginnen* als schuldig, unschuldig mit Tatwissen bzw. unschuldig ohne Tatwissen: numerischer Score als Prädiktor (SCR-Quantifizierungsmethode B)

	vorhergesagte Gruppenzugehörigkeit		
tatsächliche Gruppenzugehörigkeit	schuldig	unschuldig mit Tatwissen	unschuldig ohne Tatwissen
Täterinnen (*Schuldige*)	5 (29.4%)	7 (41.2%)	5 (29.4%)
Zeuginnen (*Unschuldige mit Tatwissen*)	7 (41.2%)	3 (17.6%)	7 (41.2%)
falsche Zeuginnen (*Unschuldige ohne Tatwissen*)	– –	4 (23.5%)	13 (76.5%)

Anmerkung: grau unterlegte Felder = Trefferquoten. Gesamttrefferquote = 41.2%.

Aus Tabelle 26 geht hervor, daß nur bei den *Unschuldigen ohne Tatwissen* eine hohe Trefferquote erzielt wurde. Fehlklassifikationen traten in dieser Gruppe nur in Form der

Zuordnung zur Gruppe der *Unschuldigen mit Tatwissen* auf. Die Trefferquoten in bezug auf die *Schuldigen* und die *Unschuldigen mit Tatwissen* lagen jeweils unter dem Zufallsniveau von 33.3%. Die *Schuldigen* wurden hauptsächlich als „unschuldig mit Tatwissen", oft aber auch als „unschuldig ohne Tatwissen" fehlklassifiziert. Die *Unschuldigen mit Tatwissen* wurden gleichermaßen sowohl als „schuldig" als auch als „unschuldig ohne Tatwissen" fehlklassifiziert.

5.3.2.2.3 Diskriminanzanalytische Klassifikationsgenauigkeit bei Zugrundelegung der intraindividuellen Reaktionsstärkedifferenzen zwischen relevanten und irrelevanten Items

Tabelle 27 gibt die diskriminanzanalytische Klassifikationsgenauigkeit (Kreuzvalidierung nach der „Hold-out sample"-Methode, s. Abschnitt 5.2.3.1) wieder, wenn die intraindividuelle SCR-Magnituden-Differenz zwischen den relevanten vs. irrelevanten *GAT*-Items als Prädiktor der Gruppenzugehörigkeit eingesetzt wurde (vgl. auch Abschnitt 5.3.1.2.3). Genauere Angaben zur entsprechenden Diskriminanzanalyse finden sich in Tabelle F.106 im Anhang F. Aus Tabelle 27 geht hervor, daß nur eine von insgesamt 51 Pbn der Kreuzvalidierungsstichprobe als „unschuldig mit Tatwissen" (fehl-) klassifiziert wurde. Sämtliche *Unschuldigen mit Tatwissen* wurden fehlklassifiziert, und zwar in etwa zu gleichen Anteilen als „schuldig" bzw. „unschuldig ohne Tatwissen". Bei den *Unschuldigen ohne Tatwissen* traten kaum Fehlklassifizierungen auf. Auch bei den *Schuldigen* lag die Trefferquote deutlich über dem Zufallsniveau von 33.3%, allerdings wurde auch ein beträchtlicher Anteil als „unschuldig ohne Tatwissen" fehlklassifiziert. Zudem sei darauf hingewiesen, daß bei den *Unschuldigen mit Tatwissen* der Anteil der als „schuldig" klassifizierten geringer war als in der Gruppe der *Schuldigen*.

Tabelle 27. Ergebnisse der Kreuzvalidierung der diskriminanzanalytischen Klassifikation der *Täterinnen*, *Zeuginnen* und *falschen Zeuginnen* als schuldig, unschuldig mit Tatwissen bzw. unschuldig ohne Tatwissen: intraindividuelle SCR-Magnituden-Differenz zwischen relevanten und irrelevanten Items als Prädiktor (SCR-Quantifizierungsmethode B)

	vorhergesagte Gruppenzugehörigkeit		
tatsächliche Gruppenzugehörigkeit	schuldig	unschuldig mit Tatwissen	Unschuldig ohne Tatwissen
Täterinnen (*Schuldige*)	11 (64.7%)	– –	6 (35.3%)
Zeuginnen (*Unschuldige mit Tatwissen*)	8 (47.1%)	– –	9 (52.9%)
falsche Zeuginnen (*Unschuldige ohne Tatwissen*)	2 (11.8%)	1 (5.9%)	14 (82.4%)

Anmerkung: grau unterlegte Felder = Trefferquoten. Gesamttrefferquote = 49.0%.

5.4 Resultate der naiven Glaubhaftigkeitsbeurteilung

5.4.1 Beschreibung der Beurteilerstichprobe

Die insgesamt 32 naiven Rater, die die beiden Zufallsstichproben von jeweils 15 experimentellen Aussagen beurteilten (vgl. Abschnitt 4.4.3) waren zwischen 19 und 32 Jahren alt (M = 24 Jahre, SD = 3.6 Jahre, Md = 23 Jahre). Unter den Beurteilern befanden sich 13 Psychologie-Studierende, ferner 11 Studierende anderer Fachrichtungen (Sozialpädagogik, Sozialarbeit [4], Sport [2], Betriebswirtschaftslehre, Jura [2], Publizistik). Bei den nichtstudentischen Ratern handelte es sich um einen Juristen, einen Programmierer, drei Soldaten und drei Kaufleute. Die Beurteilerstichprobe war zur Hälfte weiblich bzw. männlich. Die Frage, ob sie irgendwelche Vorkenntnisse zum Thema „Glaubwürdigkeitsbeurteilung“ besäßen, wurde von sämtlichen Beurteilern verneint.

5.4.2 Differenzierung der experimentellen Gruppen und Ratereffekte

Die Einflüsse der experimentellen Gruppenzugehörigkeit der Aussagen sowie etwaiger differentieller Beurteilungsstile der naiven Beurteiler auf die Höhe der eingeschätzten Glaubhaftigkeit der Aussagen wurden überprüft, indem für die beiden Zufallsstichproben von jeweils 15 experimentellen Aussagen (5 *Täterinnen*, 5 *Zeuginnen*, 5 *falsche Zeuginnen*; s. Abschnitt 4.4.3) jeweils eine zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem 16-stufigen Meßwiederholungsfaktor Rater gerechnet wurde. Die Ergebnisse der beiden ANOVAs sind in den Tabellen F.107 bzw. F.108 im Anhang F dokumentiert. Für keine der beiden Aussagenstichproben ergab sich ein signifikanter Haupteffekt der experimentellen Gruppenzugehörigkeit auf die Höhe der eingeschätzten Glaubhaftigkeit. Ebenso wurde die Wechselwirkung zwischen Gruppenzugehörigkeit und Rater weder in Aussagenstichprobe I noch in Stichprobe II statistisch bedeutsam. Der einzige signifikante Effekt war der Rater-Haupteffekt in Stichprobe I, $F(15,180) = 3.341$, $\varepsilon = .422$, $p < .01$. Allerdings ließ dieser sich in Stichprobe II nicht „replizieren“.

Die Ergebnisse der oben dargestellten statistischen Analysen sind jedoch aufgrund der geringen Zellenbesetzungen (jeweils n = 5) nur eingeschränkt interpretierbar. Um aussagekräftigere Informationen über den Einfluß der experimentellen Gruppenzugehörigkeit der Aussagen auf die naive Glaubhaftigkeitsbeurteilung zu erhalten, wurden die beiden Aussagenstichproben bzw. die entsprechenden Glaubhaftigkeitsbeurteilungen durch die jeweils 16 naiven Rater zusammengefaßt. Somit konnte eine weitere zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem

Meßwiederholungsfaktor Rater gerechnet werden, in welche die doppelte Anzahl von Aussagen einging (jeweils 10 Aussagen von *Täterinnen*, *Zeuginnen* bzw. *falschen Zeuginnen*). Natürlich muß betont werden, daß etwaige (Wechselwirkungs-) Effekte des Meßwiederholungsfaktors (Rater) in diesem statistischen Design nicht interpretierbar sind, da es sich nicht bei allen 32 Aussagen um die gleichen 16 Rater handelte. Auch in dieser ANOVA erwies sich jedoch der Haupteffekt der experimentellen Gruppenzugehörigkeit nicht als statistisch bedeutsam (s. genauer Tabelle F.109 im Anhang F), d.h. den Aussagen der *Täterinnen* (M = 4.49, SD = 2.02), *Zeuginnen* (M = 4.64, SD = 2.25) und *falschen Zeuginnen* (M = 4.27, SD = 2.12) wurde von den naiven Beurteilern alles in allem die gleiche Glaubhaftigkeit zugesprochen.

Um zu überprüfen, ob es unter den insgesamt 32 naiven Beurteilern nicht doch einige gab, die in ihren Ratings die drei experimentellen Aussagegruppen korrekt im Sinne ihrer tatsächlichen Glaubhaftigkeit differenzierten, wurde pro Beurteiler eine einfaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person gerechnet. In jede dieser insgesamt 32 ANOVAs gingen also die Beurteilungen von 5 *Täterinnen*-Aussagen, von 5 *Zeuginnen*-Aussagen sowie von 5 Schilderungen *falscher Zeuginnen* ein. Die Ergebnisse der 32 ANOVAs sind im Anhang F tabelliert (Tabelle F.110). Abbildung 14 stellt für jeden naiven Beurteiler dessen durchschnittliche Ratings bei den drei Aussagebedingungen vergleichend gegenüber. Aus der Abbildung geht hervor, daß immerhin zwölf naive Beurteiler (Rater 1, 16, 17, 19, 21, 22, 23, 24, 25, 28, 29, 30) jeweils die erlebnisbasierenden Aussagen der *Zeuginnen* im Durchschnitt für glaubhafter hielten als die erfundenen Schilderungen der *Täterinnen* und der *falschen Zeuginnen*. Allerdings erwiesen sich diese Unterschiede zwischen den experimentellen Aussagegruppen bei keinem der genannten zwölf Beurteiler als statistisch bedeutsam (s. Tabelle F.110 im Anhang F). Dabei ist jedoch zu betonen, daß das Ausbleiben statistischer Signifikanz in einigen Fällen auch allein durch die geringe Zellenbesetzung (n = 5) bedingt sein kann (hierauf wird in Abschnitt 6.3 näher eingegangen). Der einzige signifikante Effekt zeigte sich bei Rater 5. Dieser hielt die Aussagen der *Täterinnen* für glaubhafter als die Aussagen der beiden anderen experimentellen Gruppen (s. Abbildung 14).

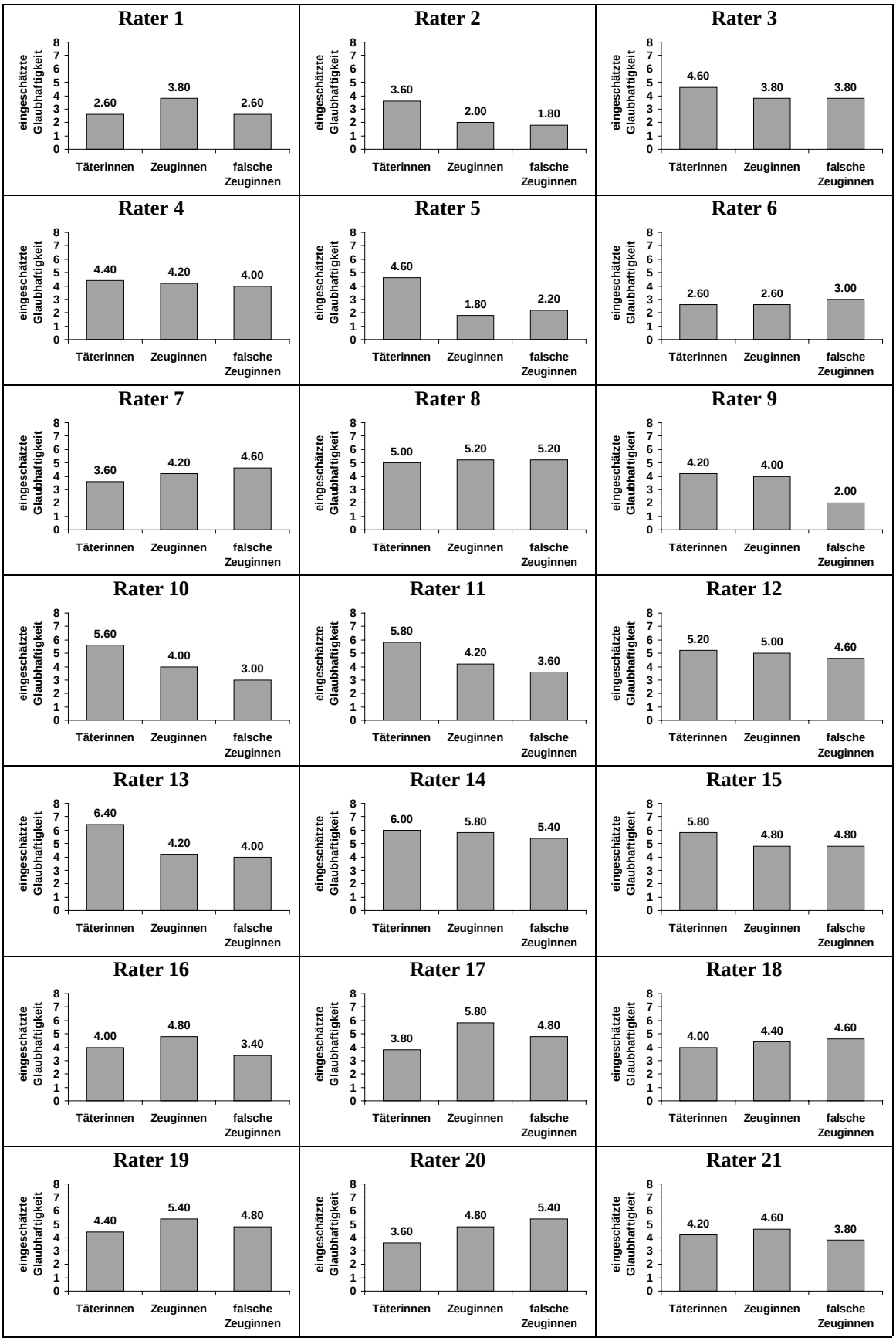


Abbildung 14. Naive Beurteilung der Glaubhaftigkeit, getrennt nach Ratern und experimentellen Aussagegruppen.

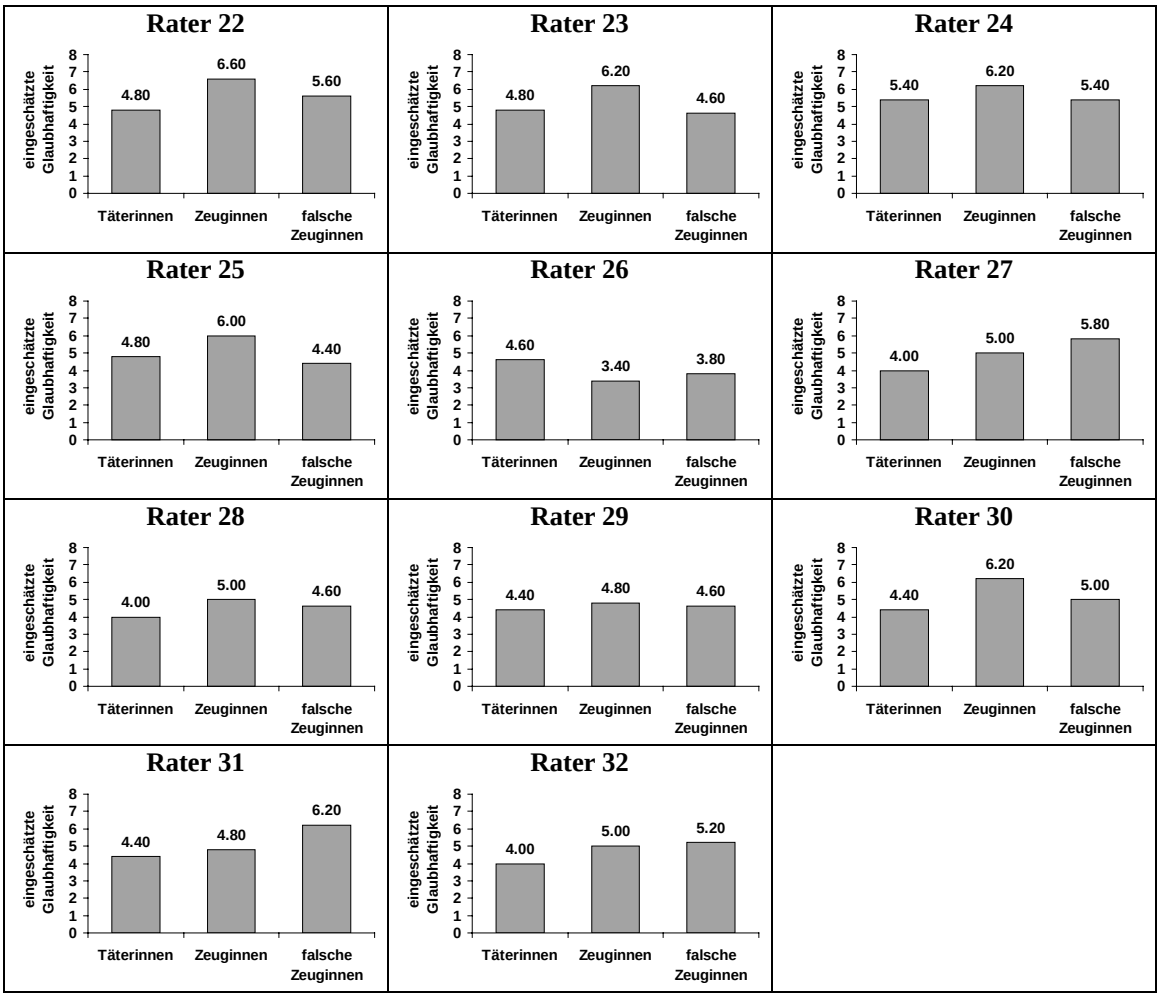


Abbildung 14 (Fortsetzung). Naive Beurteilung der Glaubhaftigkeit, getrennt nach Ratern und experimentellen Aussagegruppen.

5.4.3 Treffsicherheit

Zur Bestimmung der Treffsicherheit der naiven Glaubhaftigkeitsbeurteilung wurden zunächst die auf der achtstufigen Skala (vgl. Anhang C) vorgenommenen Ratings dichotomisiert. Die Skalenstufen 1 (*äußerst unglaubwürdig*) bis 4 (*eher unglaubwürdig*) wurden zu dem Urteil „unglaubhaft“ zusammengefaßt; die Skalenstufen 5 (*eher glaubwürdig*) bis 8 (*äußerst glaubwürdig*) repräsentierten das Urteil „glaubhaft“. Die dichotomisierten Urteile wurden dann zur tatsächlichen Glaubhaftigkeit der Aussagen in Beziehung gesetzt. Tabelle 28 gibt sowohl getrennt nach den einzelnen Auswertern als auch zusammenfassend an, wie häufig die dichotomisierten naiven Beurteilungen mit der tatsächlichen Glaubhaftigkeit der Aussagen übereinstimmten bzw. dieser widersprachen. Wie aus der untersten Zeile von Tabelle 28 hervorgeht, bewegte sich die Treffsicherheit der naiven Glaubhaftigkeitsbeurteilung alles in allem auf dem Zufallsniveau von 50%. Dies galt gleichermaßen für alle drei Aussagebedingungen. Allerdings geht aus der rechten äußeren Spalte der Tabelle auch hervor, daß bei zehn naiven Ratern

(Rater 1, 4, 16, 17, 21, 22, 23, 25, 29, 30) die jeweilige basisratenkorrigierte Gesamttrefferquote immerhin 60% oder mehr betrug, also über dem Zufallsniveau lag. Fünf dieser zehn Rater (Rater 4, 16, 17, 21, 29) diagnostizierten sowohl die glaubhaften Aussagen der *Zeuginnen* als auch die unglaubhaften Aussagen der *Täterinnen* und der *falschen Zeuginnen* jeweils in mehr als der Hälfte der Fälle zutreffend.

Tabelle 28. Treffsicherheit der naiven Glaubhaftigkeitsbeurteilung

	Status der aussagenden Person													
	Täterinnen (unglaubhaft)				*Zeuginnen* (glaubhaft)				*falsche Zeuginnen* (unglaubhaft)					
	Diagnose				Diagnose				Diagnose				Gesamt-trefferquote	
	unglaubh.		glaubhaft		unglaubh.		glaubhaft		unglaubh.		glaubhaft			
Rater	n	%	n	%	n	%	n	%	n	%	n	%	%	%*
1	4	80	1	20	3	60	2	40	4	80	1	20	66.7	60.0
2	4	80	1	20	5	100	0	0	5	100	0	0	60.0	45.0
3	3	60	2	40	4	80	1	20	3	60	2	40	46.7	40.0
4	3	60	2	40	2	40	3	60	3	60	2	40	60.0	60.0
5	2	40	3	60	5	100	0	0	5	100	0	0	46.7	35.0
6	4	80	1	20	4	80	1	20	3	60	2	40	53.3	45.0
7	4	80	1	20	3	60	2	40	2	40	3	60	53.3	50.0
8	2	40	3	60	2	40	3	60	2	40	3	60	46.7	50.0
9	2	40	3	60	3	60	2	40	5	100	0	0	60.0	55.0
10	1	20	4	80	3	60	2	40	4	80	1	20	46.7	45.0
11	2	40	3	60	3	60	2	40	3	60	2	40	46.7	45.0
12	2	40	3	60	2	40	3	60	2	40	3	60	46.7	50.0
13	1	20	4	80	3	60	2	40	3	60	2	40	40.0	40.0
14	1	20	4	80	1	20	4	80	1	20	4	80	40.0	50.0
15	2	40	3	60	3	60	2	40	2	40	3	60	40.0	40.0
16	3	60	2	40	2	40	3	60	4	80	1	20	66.7	65.0
17	3	60	2	40	1	20	4	80	3	60	2	40	66.7	70.0
18	3	60	2	40	3	60	2	40	2	40	3	60	46.7	45.0
19	2	40	3	60	2	40	3	60	2	40	3	60	46.7	50.0
20	3	60	2	40	2	40	3	60	2	40	3	60	53.3	55.0
21	4	80	1	20	2	40	3	60	4	80	1	20	73.3	70.0
22	3	60	2	40	1	20	4	80	1	20	4	80	53.3	60.0
23	2	40	3	60	0	0	5	100	2	40	3	60	60.0	70.0
24	1	20	4	80	1	20	4	80	1	20	4	80	40.0	50.0
25	2	40	3	60	0	0	5	100	3	60	2	40	66.7	75.0
26	3	60	2	40	4	80	1	20	3	60	2	40	46.7	40.0
27	3	60	2	40	2	40	3	60	1	20	4	80	46.7	50.0
28	4	80	1	20	3	60	2	40	3	60	2	40	60.0	55.0
29	3	60	2	40	2	40	3	60	3	60	2	40	60.0	60.0
30	3	60	2	40	1	20	4	80	1	20	4	80	53.3	60.0
31	3	60	2	40	3	60	2	40	1	20	4	80	40.0	40.0
32	3	60	2	40	2	40	3	60	2	40	3	60	53.3	55.0
gesamt		53.1		46.9		48.1		51.9		53.1		46.9	52.7	52.5

Anmerkung: grau unterlegte Felder = Trefferquoten; * Gesamttrefferquote bei Korrektur der Basisraten von glaubhaften und unglaubhaften Aussagen.

6 Diskussion

Ziel der vorliegenden Untersuchung war der direkte empirische Vergleich inhaltsorientierter und psychophysiologischer Methoden der forensischen Glaubhaftigkeitsbeurteilung. Die inhaltsorientierte Glaubhaftigkeitsbeurteilung erfolgte anhand der *Kriterienorientierten Inhaltsanalyse* (Steller & Köhnken, 1989), wobei die ersten 18 der insgesamt 19 Glaubhaftigkeitskriterien zur Anwendung kamen. Die psychophysiologische Glaubhaftigkeitsbeurteilung erfolgte anhand des *GAT*. Neben den beiden genannten Beurteilungsmethoden wurde als Kontrollbedingung auch noch eine naive Einschätzung der Glaubhaftigkeit vorgenommen. Der Vergleich der drei diagnostischen Ansätze wurde im Rahmen einer experimentellen Simulationsstudie durchgeführt, an welcher drei Gruppen von Pbn teilnahmen. Die Pbn der ersten Gruppe begingen ein Scheinverbrechen (*Täterinnen* bzw. *Schuldige*). Anschließend stritten sie die Täterschaft wahrheitswidrig ab und bezichtigten in einer erfundenen „Zeugenaussage“ eine andere Person der Täterschaft. Die Pbn der zweiten Gruppe beobachteten das Delikt (*Zeuginnen* bzw. *Unschuldige mit Tatwissen*). Anschließend legten sie eine wahrheitsgemäße Zeugenaussage zum Tathergang ab und stritten die Täterschaft wahrheitsgemäß ab. Die Pbn der dritten Gruppe begingen weder das Scheinverbrechen noch beobachteten sie es. Anschließend legten sie eine erfundene „Zeugenaussage“ zum Tathergang ab, stritten jedoch die Täterschaft wahrheitsgemäß ab (*falsche Zeuginnen* bzw. *Unschuldige ohne Tatwissen*). Die Glaubhaftigkeit der vermeintlichen Zeugenaussage wurde jeweils mit Hilfe der *Kriterienorientierten Inhaltsanalyse* beurteilt. Die Glaubhaftigkeit der Täterschaftsabstreitung beurteilte man jeweils psychophysiologisch, mit dem *GAT*. Mit diesem Verfahren soll grundsätzlich differenzierbar sein, ob es sich bei einer verdächtigen Person um den Täter (Schuldigen), einen Unschuldigen mit Tatwissen oder einen Unschuldigen ohne Tatwissen handelt. Die naive Glaubhaftigkeitsbeurteilung erfolgte jeweils, indem Personen ohne theoretische oder methodische Vorkenntnisse auf dem Gebiet der Glaubhaftigkeitsbeurteilung Videoaufzeichnungen der vermeintlichen Zeugenaussagen vorgespielt wurden, welche von den naiven Beurteilern dann auf einer standardisierten Glaubhaftigkeitsskala einzustufen waren. Im Rahmen dieses experimentellen Designs konnte somit überprüft werden, wie gut die verschiedenen Beurteilungsmethoden im intendierten Sinne zwischen den experimentellen Gruppen zu differenzieren vermochten.

6.1 Zum diagnostischen Potential der inhaltsorientierten Glaubhaftigkeitsbeurteilung in der vorliegenden Studie

Wie in Abschnitt 2.1.5 erläutert wurde, hat die Validierung der inhaltsorientierten Glaubhaftigkeitsbeurteilung in zwei Schritten zu erfolgen. Der Ermittlung der diagnostischen Treffsicherheit sollte die Überprüfung der theoretischen Grundannahme („Undeutsch-Hypothese“) vorausgehen. In diesem Abschnitt wird zunächst auf die Ergebnisse zur Gültigkeit der „Undeutsch-Hypothese“ eingegangen. Anschließend wird die diagnostische Treffsicherheit der inhaltsorientierten Glaubhaftigkeitsbeurteilung in der vorliegenden Untersuchung diskutiert. Beginnen soll die Diskussion jedoch mit einigen ergänzenden Überlegungen zur Auswertungsobjektivität der *Kriterienorientierten Inhaltsanalyse* in der vorliegenden Studie.

6.1.1 Ergänzende Bemerkungen zur Auswertungsobjektivität

In Abschnitt 5.2.1 wurde ausführlich auf die Objektivität bzw. Interrater-Reliabilität der inhaltsanalytischen Auswertung der experimentellen Aussagen eingegangen, wobei auch bereits die kritische Würdigung der diesbezüglichen Ergebnisse vorgezogen wurde (s. Abschnitt 5.2.1.6). An dieser Stelle sollen lediglich die wichtigsten Ergebnisse und Schlußfolgerungen nochmals in Erinnerung gerufen sowie zu den Resultaten anderer Studien in Beziehung gesetzt werden.

Die Analyse der Verteilungseigenschaften der von den drei geschulten Auswertern vorgenommenen Ratings ergab, daß die Produkt-Moment-Korrelation, der gewichtete Kappa-Koeffizient und die Intraklassenkorrelation im vorliegenden Fall nicht als Kennwerte für die Auswertungsobjektivität geeignet waren. Daher stützte sich die Bewertung der Auswertungsobjektivität in erster Linie auf die – pro Glaubhaftigkeitskriterium berechneten – **Finn-Koeffizienten**, ferner auf die **erweiterten prozentualen Übereinstimmungen** zwischen den Ratern. Auf der Grundlage dieser beiden Indizes konnte die **Auswertungsobjektivität von 17 Kriterien** als **mindestens akzeptabel** eingestuft werden. Lediglich die Auswertungsobjektivität von **Kriterium 9 (*Nebensächliches*)** erwies sich als **unzureichend**. Für die Validitätsanalysen bedeutete dies, daß man bei den 17 Kriterien mit akzeptabler Auswertungsobjektivität jeweils die Ratings der drei Auswerter durch **arithmetische Mittelung** zusammenfassen konnte. Dagegen konnte eine Mittelung der Ratings der drei Auswerter bei Kriterium 9 (*Nebensächliches*) nur unter Vorbehalt erfolgen.

Bislang wurden in Fachzeitschriften zwei empirische Studien publiziert, die sich ausführlich mit der Frage der Auswertungsobjektivität bzw. Interrater-Reliabilität der *Kriterienorientierten Inhaltsanalyse* auseinandersetzten (Anson, Golding & Gully, 1993; Horowitz, Lamb, Esplin, Boychuk, Krispin & Reiter-Lavery, 1997). In der Untersuchung von Anson et al. (1993) wurde für neun Kriterien eine „adäquate" („Adequate reliability", S. 336), für weitere vier Kriterien eine „marginale Reliabilität" („Marginal reliability", S. 336) festgestellt. Die Auswertungsobjektivität von sechs Kriterien (2. *Unordnung*; 5. *Interaktionen*; 11. *Indirektes*; 12. *Eigenseelisches*; 15. *Erinnerungslükken*; 19. *Deliktspezifisches*) wurde als unzureichend („inadequate reliability", S. 336) eingestuft. Als entscheidendes statistisches Reliabilitätsmaß zogen die Autoren Maxwell´s RE-Koeffizienten heran. Letzterer ist nur für dichotome Ratings konzipiert, was auch der Grund dafür ist, daß er in der vorliegenden Studie nicht als Objektivitätskennwert verwendet werden konnte. Kriterium 9 (*Nebensächliches*), für welches in der vorliegenden Studie als einziges keine ausreichende Auswertungsobjektivität resultierte, erreichte in der Studie von Anson et al. (1993) „marginale Reliabilität". Die durchschnittliche prozentuale Übereinstimmung der Rater über die Kriterien 1 bis 18 betrug bei Anson et al. (1993) 76.5%. In der vorliegenden Studie betrug die durchschnittliche erweiterte prozentuale Gesamtübereinstimmung aller drei Rater über alle 18 Kriterien 89.2%, bei Zugrundelegung der mittleren paarweisen Übereinstimmungen zwischen den Ratern sogar 94.7% (s. Tabelle F.21 im Anhang F). Dabei ist zu berücksichtigen, daß die Rater bei Anson et al. (1993) lediglich dichotome Urteile abgeben mußten (Kriterium vorhanden vs. nicht vorhanden), wohingegen die Rater in der vorliegenden Studie vierstufige Skalen verwendeten. Insofern ist die prozentuale Raterübereinstimmung in der vorliegenden Studie im Vergleich zu Anson et al. (1993) als deutlich höher zu bewerten. Allerdings ist die Interpretierbarkeit bzw. Vergleichbarkeit der Ergebnisse von Anson et al. (1993) insbesondere deshalb eingeschränkt, weil die Ratings dort auf der Grundlage von Aussagevideos erfolgten, wohingegen standardgemäß Aussagetranskripte zu verwenden sind, wie es auch in der vorliegenden Studie praktiziert wurde. Ferner ist an der Studie von Anson et al. (1993) die geringe Aussagenstichprobe von nur N = 23 zu bemängeln.

Einen besseren Vergleichsmaßstab als die Studie von Anson et al. (1993) bietet die Untersuchung von Horowitz et al. (1997). Drei geschulte Rater analysierten jeweils 100 Aussagetranskripte auf das Vorhandensein bzw. Fehlen der 19 Kriterien der *Kriterienorientierten Inhaltsanalyse*. Als entscheidender statistischer Kennwert für die Interrater-Reliabilität der dichotomen Ratings wurde auch hier Maxwell´s RE-Koeffizient herangezogen. Bei fünf Kriterien (8. *Ausgefallenes*; 9. *Nebensächliches*; 11. *Indirektes*; 14. *Verbesserungen*; 15. *Erinnerungslücken*) wurde die Auswertungsobjektivität als unzureichend eingestuft. Die durchschnittliche prozentuale Übereinstimmung zwischen den

Ratern bei den Kriterien 1 bis 18 betrug 82.1%[30], verglichen mit der erweiterten prozentualen Übereinstimmung von 89.2% bzw. 94.7% in der vorliegenden Studie. Gemessen an der prozentualen Übereinstimmung, wurde in der vorliegenden Studie also auch eine höhere Auswertungsobjektivität erzielt als bei Horowitz et al. (1997). Interessant ist, daß sich die Auswertungsobjektivität von Kriterium 9 (*Nebensächliches*) sowohl in der vorliegenden Studie als auch bei Horowitz et al. (1997) als unzureichend erwies, wobei die für dieses Kriterium resultierende prozentuale Übereinstimmung bei Horowitz und Kollegen (70%) noch geringer war als die erweiterte prozentuale Übereinstimmung bei Kriterium 9 in der vorliegenden Studie (74.2%; s. Tabelle F.21).

In Relation zu den beiden publizierten Reliabilitätsstudien läßt sich die **Interrater-Reliabilität der vorliegenden Untersuchung** also **insgesamt** als **zufriedenstellend** bewerten. Zum einen erwiesen sich in der vorliegenden Studie – gemessen an den erweiterten prozentualen Übereinstimmungen und insbesondere an den Finn-Koeffizienten – mehr Kriterien als hinreichend reliabel als in den beiden anderen Studien, in denen allerdings ein anderer statistischer Kennwert (Maxwell´s RE-Koeffizient) als Entscheidungsgrundlage diente. Zum anderen war die Gesamtobjektivität über die Kriterien 1 bis 18 in der vorliegenden Studie höher als in den beiden anderen Untersuchungen, sofern man sich an der (erweiterten) prozentualen Raterübereinstimmung orientiert. Es ist allerdings zu bedenken, daß in den Studien von Anson et al. (1993) sowie Horowitz et al. (1997) kindliche Aussagen aus forensischen Realfällen (mutmaßliche Sexualdelikte) als Analysematerial verwendet wurden, wohingegen es sich in der vorliegenden Untersuchung um Aussagen erwachsener Frauen zu einem experimentell simulierten Diebstahl handelte. Es ist offen, inwiefern hierdurch die Vergleichbarkeit der jeweiligen Reliabilitätsbefunde eingeschränkt wird.

Auf jeden Fall sprechen die Ergebnisse der vorliegenden Untersuchung dafür, daß bei adäquater Schulung der Rater eine befriedigende Auswertungsobjektivität erzielt werden kann. Somit sind die vorliegenden Befunde auch als **Bestätigung für die Effizienz des *Kieler Trainingsprogramms zur Beurteilung der Glaubwürdigkeit von Zeugenaussagen*** (*KTBG*; Krause, 1997; Petersen, 1997; vgl. auch Höfer et al., 1999) anzusehen, an welchem sich die Raterschulung in der vorliegenden Studie orientierte. Man kann sogar davon ausgehen, daß die vorliegenden Objektivitätsbefunde die wirkliche Effektivität des *KTBG* eher unterschätzen, da das *KTBG* bei der vorliegenden Raterschulung nur als grober Leitfaden diente bzw. die einzelnen *KTBG*-Trainingselemente hier zeitlich sehr stark gestrafft wurden (vgl. Anhang E). Bei einer originalgetreuen Umsetzung des *KTBG* wäre die Auswertungsobjektivität in der vorliegenden Untersu-

[30] Die Aussagen wurden von den Ratern zweimal ausgewertet, um auch die Retest-Reliabilität bestimmen zu können. Der Wert von 82.1% ist der Durchschnitt aus den beiden Auswertungsdurchgängen.

chung möglicherweise noch höher ausgefallen. Zudem ist es denkbar, daß bei einer originalgetreuen Umsetzung des *KTBG* sich auch für Kriterium 9 (*Nebensächliches*) eine hinreichende Auswertungsobjektivität ergeben hätte. In einem unveröffentlichten Manuskript über eine empirische Evaluation des *KTBG* (Krause & Petersen, keine Jahresangabe) wird jedenfalls berichtet, daß bezüglich Kriterium 9 sowohl die prozentuale Raterübereinstimmung als auch Maxwell´s RE-Koeffizient eine „adäquate“ Höhe erreicht haben. Insofern kommt als Ursache für die mangelhafte Auswertungsobjektivität von Kriterium 9 in der vorliegenden Studie in erster Linie eine unzureichende Unterweisung der Rater in bezug auf dieses Kriterium in Betracht.

Abschließend sei noch angemerkt, daß in manchen Untersuchungen neben der Interrater-Reliabilität bzw. Auswertungsobjektivität auch noch die **Übereinstimmung der Rater mit einem Experten** – als Maß für die Validität der Ratings – erfaßt wird (vgl. Steller et al., 1992). Als Experten gelten im Optimalfall Personen, die an der Entwicklung der *Kriterienorientierten Inhaltsanalyse* beteiligt waren. In der vorliegenden Untersuchung konnte keine Überprüfung der Rater-Experten-Übereinstimmung erfolgen, da kein Experte im oben genannten Sinne zur Verfügung stand. Ebensowenig konnte in der vorliegenden Studie auf den Aspekt der **Retest-Reliabilität** (zeitliche Stabilität) der *Kriterienorientierten Inhaltsanalyse* eingegangen werden, wie dies etwa bei Horowitz et al. (1997) getan wurde. Der Grund hierfür liegt darin, daß die drei geschulten Rater nur drei Monate zur Verfügung standen; das Zeitintervall zwischen den mindestens zwei erforderlichen Auswertungen einer Aussage durch ein und denselben Rater müßte aber schon mindestens drei Monate betragen (vgl. Horowitz et al., 1997). Es **muß** also **offen bleiben**, ob die Ratings der vorliegenden Untersuchung reliabel im Sinne zeitlicher Stabilität waren und wie hoch ihre Validität im Sinne einer Rater-Experten-Übereinstimmung war.

6.1.2 Gültigkeit der „Undeutsch-Hypothese“ in der vorliegenden Untersuchung

Die Grundannahme der inhaltsorientierten Glaubhaftigkeitsbeurteilung besagt, daß erlebnisbezogene Schilderungen im intraindividuellen Vergleich eine höhere inhaltliche Qualität aufweisen als erfundene Aussagen. Die inhaltliche Aussagequalität wurde in der vorliegenden Studie anhand der *Kriterienorientierten Inhaltsanalyse* operationalisiert. Gemäß der Grundannahme war zu erwarten, daß die *Zeuginnen* im Durchschnitt eine höhere inhaltliche Aussagequalität erzielen würden als die *Täterinnen* und als die *falschen Zeuginnen*, d.h. daß sich die 18 inhaltlichen Glaubhaftigkeitskriterien in den erlebnisbezogenen Aussagen der *Zeuginnen* häufiger bzw. in stärkerer Ausprägung ma-

nifestieren würden als in den konfabulierten Schilderungen der beiden anderen experimentellen Gruppen.

Die „Undeutsch-Hypothese“ erfuhr in der vorliegenden Studie **grundsätzliche Bestätigung**. Dies kam am unmittelbarsten darin zum Ausdruck, daß die über alle 18 Glaubhaftigkeitskriterien aufsummierten **Gesamtscores bei den *Zeuginnen*** im Durchschnitt **signifikant höher** waren **als bei den *Täterinnen* und bei den *falschen Zeuginnen***, wobei sich **zwischen den beiden falschaussagenden Gruppen kein signifikanter Unterschied** ergab.

Die signifikante Differenzierung der experimentellen Gruppen anhand der Gesamtscores steht im Einklang mit den Resultaten sämtlicher Studien, in denen ein entsprechendes Gesamtmaß gebildet bzw. statistisch analysiert wurde. In drei Felduntersuchungen (Craig et al., 1999; Esplin et al., 1988, zitiert nach Raskin & Esplin, 1991a, S. 160ff.; Lamb et al., 1997) und in drei experimentellen Studien (Landry & Brigham, 1992; Vrij et al., 2000; Winkel & Vrij, 1995) hat man entsprechende Gesamtscores über die jeweils untersuchten Glaubhaftigkeitskriterien berechnet (vgl. Abschnitt 2.1.6.1). In allen genannten Studien ergaben sich signifikante Unterschiede dergestalt, daß die Gesamtscores der erlebnisbezogenen Aussagen höher waren als die der erfundenen Schilderungen. Dabei gilt es jedoch zu beachten, daß die hinter den statistischen Signifikanzen stehenden Unterschiede zwischen erlebnisbezogenen und erfundenen Aussagen in den einzelnen Studien unterschiedlich stark ausgeprägt waren. So gab es in der Feldstudie von Esplin et al. (1988, zitiert nach Raskin & Esplin, 1991a, S. 160ff.), in welcher Gesamtscores von 0 bis 38 erzielt werden konnten, keine Überschneidung der Häufigkeitsverteilungen für die Gesamtscores in den erlebnisbezogenen vs. erfundenen Aussagen. Die Gesamtscores der erfundenen Aussagen lagen in einem Bereich von 0 bis 10 (M = 3.6), diejenigen der erlebnisbezogenen Aussagen lagen in einem Bereich von 16 bis 34 (M = 24.8). Demgegenüber war die Differenzierung etwa in der Felduntersuchung von Lamb et al. (1997) deutlich schwächer. Bei einem maximal erreichbaren Gesamtscore von 14 betrug der Mittelwert der erlebnisbezogenen Aussagen 6.74, derjenige der erfundenen Aussagen 4.85 (zu den Streuungen beider Verteilungen werden leider keine Angaben gemacht). Beim Vergleich der Ergebnisse dieser beiden Feldstudien muß allerdings berücksichtigt werden, daß bei Esplin et al. (1988, zitiert nach Raskin & Esplin, 1991a, S. 160ff.) alle 19 Kriterien auf jeweils dreistufigen Skalen (0, 1, 2) ausgewertet wurden, während bei Lamb et al. (1997) nur 14 Kriterien berücksichtigt und zudem jeweils nur dichotom (0, 1) kodiert wurden. Daher ist die größere Mittelwertdifferenz zwischen erlebnisbezogenen und erfundenen Aussagen bei Esplin et al. (1988, zitiert nach Raskin & Esplin, 1991a, S. 160ff.) teilweise auch durch die größere Skalenbreite der Gesamtscores (0 – 38, verglichen mit 0 – 14 bei Lamb et al., 1997)

bedingt. In der dritten Feldstudie, in welcher Gesamtscores gebildet wurden (Craig et al., 1999), ergab sich ebenfalls ein eher moderater (wenngleich signifikanter) Unterschied zwischen erlebnisbezogenen und erfundenen Aussagen. Wie bei Lamb et al. (1997) wurden nur 14 Glaubhaftigkeitskriterien berücksichtigt, so daß bei dichotomer Kodierung (0, 1) der Gesamtscore von null bis 14 variieren konnte. Der Mittelwert bei den erlebnisbezogenen Aussagen betrug 7.2, der bei den erfundenen Aussagen 5.7. Die Standardabweichungen in den beiden Aussagegruppen (2.2 bzw. 3.2) zeigen an, daß die Gesamtscore-Häufigkeitsverteilungen von erlebnisbezogenen und erfundenen Aussagen sich deutlich überschnitten.

Auch in den drei genannten Experimentalstudien war die (signifikante) Differenzierung der Aussagegruppen anhand der Gesamtscores eher mäßig ausgeprägt. In der Studie von Vrij et al. (2000), in welcher bei dichotomer Kodierung (0, 1) von zwölf Glaubhaftigkeitskriterien ein Gesamtscore von null bis zwölf erzielt werden konnte, betrugen die durchschnittlichen Gesamtscores der erlebnisbezogenen und der erfundenen Aussagen 4.33 bzw. 2.67. Den entsprechenden Standardabweichungen (1.73 bzw. 1.10) ist zu entnehmen, daß die Häufigkeitsverteilungen der Gesamtscores beider Aussagegruppen sich deutlich überschnitten. Bei Landry und Brigham (1992) wurden die Gesamtscores pro Aussagebedingung (erlebnisbezogen vs. erfunden) berechnet, indem man die Scores für die einzelnen Kriterien nicht nur über alle Kriterien, sondern auch über sämtliche Aussagen einer Bedingung aufsummierte. Bei Verwendung dreistufiger Ratingskalen (0, 1, 2), 14 berücksichtigten Glaubhaftigkeitskriterien und sechs Aussagen pro Bedingung konnten somit Gesamtscores von null bis 168 erzielt werden. Die Gesamtscores der erlebnisbezogenen und der erfundenen Aussagen betrugen 57.69 bzw. 51.49 und lagen somit sehr dicht beieinander. Bei Winkel und Vrij (1995) diente pro Aussage der Mittelwert der Scores auf den elf verwendeten Einzelkriterien als Gesamtscore. Da vermutlich vierstufige Ratingskalen (0, 1, 2, 3) verwendet wurden (dies geht aus der Publikation nicht genau hervor), ist anzunehmen, daß die Gesamtscores im Bereich von 0 bis 3 liegen konnten. Für die erlebnisbezogenen Aussagen ergab sich ein durchschnittlicher Gesamtscore von 1.68, bei den frei erfundenen Aussagen betrug der mittlere Gesamtscore 1.21. Zu den Streuungen werden keine Angaben gemacht; gleichwohl lassen die bloßen Mittelwerte vermuten, daß die Differenzierung zwischen den beiden Aussagegruppen nicht sehr deutlich war.

Auch in der vorliegenden Untersuchung ergaben sich zwar für die erlebnisbezogenen Aussagen der *Zeuginnen* signifikant höhere Gesamtscores als für die erfundenen Schilderungen der *Täterinnen* und der *falschen Zeuginnen*; gleichwohl waren die signifikanten **Gruppenunterschiede – rein numerisch betrachtet – nicht sehr deutlich ausgeprägt**. Da für die 18 berücksichtigten Glaubhaftigkeitskriterien jeweils vierstufige Ra-

tingskalen (0, 1, 2, 3) verwendet wurden, konnte der Gesamtscore einer Aussage 0 bis maximal 54 betragen. Vor diesem Hintergrund unterschied sich der durchschnittliche Gesamtscore der erlebnisbezogenen Aussagen der *Zeuginnen* (14.49) nicht sehr von den mittleren Gesamtscores der erfundenen Aussagen der *Täterinnen* und der *falschen Zeuginnen* (12.43 bzw. 11.98). Die Standardabweichungen (*Zeuginnen*: 3.07; *Täterinnen*: 3.85; *falsche Zeuginnen*: 2.98) lassen zudem erkennen, daß sich die Gesamtscore-Häufigkeitsverteilungen der drei experimentellen Gruppen weitgehend überlagerten. Insofern liegt die vorliegende Untersuchung im Trend der meisten anderen empirischen Studien, in denen Gesamtscores analysiert wurden. Die Trennung erlebnisbezogener und erfundener Aussagen anhand der Gesamtscores war zwar statistisch überzufällig, numerisch jedoch eher moderat ausgeprägt. Lediglich in der Feldstudie von Esplin et al. (1988, zitiert nach Raskin & Esplin, 1991a, S. 160ff.) war die Differenzierung zwischen erlebnisbasierenden und konfabulierten Aussagen anhand der Gesamtscores wesentlich deutlicher.

Daß die **Differenzierung zwischen erlebnisbezogenen und erfundenen Aussagen anhand der Gesamtscores** in der vorliegenden Studie eher mäßig ausgeprägt war, kam auch darin zum Ausdruck, daß sie stark **von der Beurteilung von Kriterium 9 (*Nebensächliches*) abhing**. Da für dieses Glaubhaftigkeitskriterium keine ausreichende Auswertungsobjektivität sichergestellt werden konnte, war nicht auszuschließen, daß die Höhe der Gesamtscores und die Differenzierung der experimentellen Gruppen anhand der Gesamtscores möglicherweise mit der differentiellen Beurteilung von Kriterium 9 durch die drei Rater zusammenhingen. Zur genaueren Überprüfung dieser Vermutung wurden drei zusätzliche ANOVAs gerechnet, in welchen anstelle des Mittelwerts der drei Rater bei Kriterium 9 jeweils nur der von Rater A bzw. Rater B bzw. Rater C quantifizierte Ausprägungsgrad dieses Kriteriums in den Gesamtscore einfloß. Ging das Rating von Auswerter B bei Kriterium 9 oder dasjenige von Auswerter C bei Kriterium 9 in den Gesamtscore ein, so zeigte sich jeweils der gleiche Effekt wie in der ursprünglichen ANOVA, in welcher das mittlere Rating aller drei Auswerter im Gesamtscore berücksichtigt wurde, d.h. die Gesamtscores der *Zeuginnen* waren im Durchschnitt signifikant höher als die der beiden falschaussagenden Gruppen. Wurde jedoch das mittlere Rating aller drei Auswerter bei Kriterium 9 durch das diesbezügliche Rating von Auswerter A ersetzt, so differenzierte der resultierende Gesamtscore nur noch marginal zwischen den Aussagen der *Zeuginnen* und der *falschen Zeuginnen*. (Auf die Unterschiede zwischen den drei Auswertern bezüglich der Einschätzung von Kriterium 9 wird weiter unten noch genauer eingegangen.)

Bei der Gültigkeitsüberprüfung der „Undeutsch-Hypothese“ auf der Ebene einzelner Glaubhaftigkeitskriterien fiel zunächst auf, daß letztere in der vorliegenden Untersu-

chung nicht gänzlich als voneinander unabhängige Indikatoren des Erlebnisbezugs bzw. der Glaubhaftigkeit angesehen werden konnten. So erwiesen sich 40 der insgesamt 153 paarweisen Interkorrelationen zwischen den Glaubhaftigkeitskriterien als statistisch signifikant. Diese **Interdependenz der Kriterien** entspricht nicht der ursprünglichen theoretischen Annahme von Steller et al. (1992), welche postulierten: „Jedes Realkennzeichen besitzt eine eigene Definition und ist unabhängig von den Definitionen der übrigen, d.h., die 19 Realkennzeichen der Kriterienorientierten Aussageanalyse sind als voneinander unabhängige Instrumente zur Beurteilung der Glaubhaftigkeit einer Aussage anzusehen [...]“ (S. 164). Die in der vorliegenden Untersuchung festgestellte Abhängigkeit der Kriterien korrespondiert jedoch mit Befunden anderer empirischer Studien. So erwiesen sich z.B. in der Analoguntersuchung von Wolf und Steller (1997) sogar 70 der 153 paarweisen Interkorrelationen zwischen den ersten 18 Kriterien der *Kriterienorientierten Inhaltsanalyse* als statistisch signifikant – also fast doppelt so viele wie in der vorliegenden Untersuchung. Letztendlich ist die wechselseitige Abhängigkeit der Kriterien nicht verwunderlich, bedenkt man, daß alle Kriterien dasselbe indizieren sollen, nämlich den Erlebnisbezug von Aussagen.

Angesichts der festgestellten Interdependenz wurden etwaige Gruppenunterschiede in den Ausprägungsgraden der 18 Glaubhaftigkeitskriterien zunächst mittels einer **multivariaten Varianzanalyse** (MANOVA) überprüft. Der simultane Effekt der experimentellen Gruppenzugehörigkeit auf die Ausprägungsgrade der 18 Kriterien erwies sich als statistisch bedeutsam. Drei Anschluß-MANOVAs, in denen jeweils nur zwei Gruppen berücksichtigt wurden, ergaben, daß **die *Zeuginnen* sich sowohl von den *Täterinnen* als auch von den *falschen Zeuginnen* signifikant unterschieden**, während der Unterschied zwischen den beiden falschaussagenden Gruppen die (korrigierte) Signifikanzgrenze knapp verfehlte. Die Ergebnisse der MANOVAs stehen also im Einklang mit dem oben erörterten Effekt der experimentellen Gruppenzugehörigkeit auf die Höhe der Gesamtscores.

Zur genaueren Aufklärung der multivariaten Gruppenunterschiede wurde zum einen eine **Diskriminanzanalyse** gerechnet. Zum anderen wurden separate **univariate Varianzanalysen** (ANOVAs) für die einzelnen Glaubhaftigkeitskriterien durchgeführt. Beide statistischen Vorgehensweisen führten zu **äquivalenten Resultaten**. Sowohl den diskriminanzanalytischen als auch den univariat-varianzanalytischen Berechnungen zufolge unterschieden sich die experimentellen Gruppen in der Ausprägung der Glaubhaftigkeitskriterien 3 (*Details*), 4 (*Verknüpfungen*), 6 (*Gespräche*), 8 (*Ausgefallenes*) und 9 (*Nebensächliches*).

Die Anschlußtests für die signifikanten ANOVA-Effekte ergaben, daß die **Kriterien 3 (*Details*) und 6 (*Gespräche*)** jeweils **in den Aussagen der *Zeuginnen* signifikant stärker ausgeprägt** waren **als in den Aussagen der *Täterinnen* und der *falschen Zeuginnen***. Bezüglich Kriterium 3 (*Details*) betrug der durchschnittliche Score der *Zeuginnen* auf der vierstufigen Skala 2.31 und lag somit zwischen den beiden Skalenstufen *mittel vorhanden* und *stark vorhanden*. Dagegen beliefen sich die durchschnittlichen Scores der *Täterinnen* und der *falschen Zeuginnen* auf 1.51 bzw. 1.58 und lagen somit jeweils zwischen den Skalenstufen *schwach vorhanden* und *mittel vorhanden*. Kriterium 6 (*Gespräche*) war in den Aussagen der *Zeuginnen mittel* bis *stark vorhanden* (durchschnittlicher Score: 2.27); dagegen war dieses Kriterium in den Schilderungen der *Täterinnen* (M = 1.48) und der *falschen Zeuginnen* (M = 1.40) jeweils *schwach* bis *mittel vorhanden*. Was **Kriterium 9 (*Nebensächliches*)** betrifft, so wurde hier **nur der Unterschied zwischen den *Zeuginnen* und den *falschen Zeuginnen* signifikant**. Während das Kriterium in den Aussagen der *Zeuginnen* immerhin noch *schwach vorhanden* war (M = 1.07), lag der durchschnittliche Score der *falschen Zeuginnen* (M = 0.40) nur zwischen den Skalenstufen *nicht vorhanden* und *schwach vorhanden*. Was **Kriterium 8** angeht, so **berichteten neben den *Zeuginnen* auch die *Täterinnen* signifikant mehr *Ausgefallenes* als die *falschen Zeuginnen***. Während das Kriterium in den Aussagen der *Täterinnen* (M = 0.45) und der *Zeuginnen* (M = 0.57) im Durchschnitt jeweils *nicht* bis *schwach vorhanden* war, entsprach der durchschnittliche Score der *falschen Zeuginnen* (0.17) der Skalenstufe *nicht vorhanden*. Die statistische Signifikanz bei **Kriterium 4** beruhte darauf, daß **in den Aussagen der *falschen Zeuginnen* signifikant mehr *Verknüpfungen*** auftraten **als in den Schilderungen der *Zeuginnen***. Dabei lag der durchschnittliche Score der *falschen Zeuginnen* (1.42) zwischen den Skalenstufen *schwach vorhanden* und *mittel vorhanden*, während der durchschnittliche Score der *Zeuginnen* (0.86) der Skalenstufe *schwach vorhanden* entsprach.

Die **Gültigkeit der „Undeutsch-Hypothese“** konnte in der vorliegenden Untersuchung also **in erster Linie an den Kriterien 3 (*Details*) und 6 (*Gespräche*) festgemacht** werden. Diese waren die einzigen Kriterien, die in den erfundenen Aussagen sowohl der *Täterinnen* als auch der *falschen Zeuginnen* schwächer ausgeprägt waren als in den erlebnisbezogenen Aussagen der *Zeuginnen*. Während *Details* ebenso wie *Gespräche* in den erlebnisbezogenen Aussagen (*Zeuginnen*) *mittel* bis *stark vorhanden* waren, waren sie in den erfundenen Schilderungen (*Täterinnen* und *falsche Zeuginnen*) nur *schwach* bis *mittel vorhanden*. Allerdings waren die signifikanten **Gruppenunterschiede – rein numerisch betrachtet – eher moderat ausgeprägt**. Zum einen betrugen die Mittelwertsunterschiede zwischen den wahr- und falschaussagenden Gruppen hinsichtlich der Kriterien 3 und 6 jeweils weniger als eine Skalenstufe, was bei einer vierstufigen Skala relativ wenig ist. Zum anderen ist an den Standardabweichungen (s. Tabelle 16 in Ab-

schnitt 5.2.2.1) zu erkennen, daß die Häufigkeitsverteilungen der *Zeuginnen* bezüglich der Ausprägungsgrade der Kriterien 3 und 6 sich jeweils deutlich mit den entsprechenden Häufigkeitsverteilungen der falschaussagenden Gruppen überlagerten.

Neben den Kriterien 3 und 6 verhielt sich auch noch **Kriterium 9 (*Nebensächliches*)** teilweise im Sinne der „Undeutsch-Hypothese“. Allerdings war dieses Kriterium nur in den erfundenen Aussagen der *falschen Zeuginnen* signifikant schwächer ausgeprägt als in den erlebnisbasierenden Aussagen der *Zeuginnen*. Dagegen blieb der gemäß der „Undeutsch-Hypothese“ zu erwartende signifikante Unterschied zwischen *Täterinnen* und *Zeuginnen* aus. Auch die **signifikante Differenzierung** zwischen *Zeuginnen* und *falschen Zeuginnen* war – **numerisch gesehen – eher mäßig ausgeprägt**. So betrug der Unterschied zwischen den Gruppenmittelwerten kaum mehr als eine halbe Skalenstufe; und die Standardabweichungen (s. Tabelle 16 in Abschnitt 5.2.2.1) verweisen auf eine weitgehende Überlappung der Häufigkeitsverteilungen beider Gruppen. Allerdings ist die **Interpretierbarkeit** der Ergebnisse zu Kriterium 9 dadurch **eingeschränkt**, daß die **Auswertungsobjektivität** dieses Kriteriums **unzureichend** war. Um die Auswirkungen der mangelhaften Interrater-Reliabilität statistisch zu kontrollieren, wurde eine ANOVA gerechnet, in welcher neben der Gruppenzugehörigkeit der Aussagen auch noch die drei Rater als zusätzlicher Faktor berücksichtigt wurden. Dabei resultierte neben dem oben beschriebenen Haupteffekt der experimentellen Gruppenzugehörigkeit der Aussagen auch ein Rater-Haupteffekt, welcher darauf basierte, daß Rater C Kriterium 9 in den Aussagen insgesamt signifikant stärker ausgeprägt sah als Rater B. Letzterer wiederum sprach den Aussagen signifikant mehr *Nebensächliches* zu als Rater A. Entscheidend ist jedoch, daß sich neben den beiden Haupteffekten auch eine signifikante Wechselwirkung manifestierte. Diese beruhte darauf, daß die Rater B und C Kriterium 9 ganz im Sinne der „Undeutsch-Hypothese“ quantifizierten, während die von Rater A vorgenommenen Einschätzungen bezüglich dieses Kriteriums erwartungskonträr ausfielen. Das heißt, die Rater B und C erkannten jeweils in den Aussagen der *Zeuginnen* mehr *Nebensächliches* als in den Schilderungen der beiden anderen experimentellen Gruppen, wohingegen nach Einschätzung von Rater A die *Täterinnen* mehr *Nebensächliches* berichteten als die *Zeuginnen* und die *falschen Zeuginnen*. Oben wurde die Vermutung geäußert, daß die unzureichende Auswertungsobjektivität von Kriterium 9 auf eine inadäquate Unterweisung der drei Rater bezüglich dieses Kriteriums zurückzuführen war. Die aufgefundene Gruppenzugehörigkeit × Rater-Interaktion könnte Anlaß geben zu der weitergehenden Vermutung, daß möglicherweise insbesondere Rater A die Operationalisierungsvorschriften zu Kriterium 9 falsch aufgefaßt hat, wohingegen bei den beiden anderen Ratern die Unterweisung in der Handhabung dieses Kriteriums besser gelang. Gegen diese Vermutung spricht allerdings die genauere Betrachtung der Kennwerte für die Auswertungsobjektivität. Wäre die Unterweisung der Rater B und C bes-

ser gelungen als die von Rater A, so müßte auch die Urteilsübereinstimmung zwischen den Ratern B und C bei Kriterium 9 höher sein als die paarweisen Übereinstimmungen zwischen den Ratern A und B bzw. A und C. Die höchste erweiterte prozentuale Übereinstimmung bezüglich Kriterium 9 bestand jedoch zwischen den Ratern A und B, gefolgt von den Raterpaaren B-C und A-C (s. Tabelle F.21 im Anhang F). Insgesamt können die Resultate zu Kriterium 9 allenfalls als schwacher Hinweis auf die Validität der „Undeutsch-Hypothese“ gewertet werden, wobei nur spekuliert werden kann, ob sich dieses Kriterium bei einer besseren diesbezüglichen Raterschulung bzw. einer höheren Auswertungsobjektivität als validerer Indikator des Erlebnisbezugs erwiesen hätte.

Auch die signifikanten Ergebnisse zu **Kriterium 8 (*Ausgefallenes*)** können mit Einschränkungen noch als **Bestätigung der „Undeutsch-Hypothese“** angesehen werden, und zwar insofern, als dieses Kriterium in den wahren Aussagen der *Zeuginnen* signifikant stärker ausgeprägt war als in den Falschaussagen der *falschen Zeuginnen*. Allerdings wurde auch hier der im Sinne der „Undeutsch-Hypothese“ zu erwartende Unterschied zwischen *Täterinnen* und *Zeuginnen* nicht signifikant. Stattdessen wiesen die Falschaussagen der *Täterinnen* signifikant mehr *Ausgefallenes* auf als die Falschaussagen der *falschen Zeuginnen*. Allerdings muß auch hier betont werden, daß sowohl der unerwartete Unterschied zwischen *Täterinnen* und *falschen Zeuginnen* als auch der hypothesenkonforme Unterschied zwischen den *Zeuginnen* und *falschen Zeuginnen* – **rein numerisch betrachtet** – **eher moderat ausgeprägt** waren. Die Differenzen zwischen den Gruppenmittelwerten betrugen jeweils weniger als eine halbe Skalenstufe; und die Häufigkeitsverteilungen der Gruppen überlagerten sich jeweils deutlich.

Das einzige Kriterium, das sich – vordergründig – im Widerspruch zur „Undeutsch-Hypothese“ verhielt, war **Kriterium 4**. Erwartungswidrig zeichneten sich die konfabulierten Aussagen der *falschen Zeuginnen* in stärkerem Maße durch ***Verknüpfungen*** aus als die wahren Geschichten der *Zeuginnen*. Auch hier ist jedoch einschränkend zu berücksichtigen, daß der Mittelwertsunterschied kaum mehr als eine halbe Skalenstufe betrug und daß die Häufigkeitsverteilungen beider Gruppen bezüglich der Ausprägungsgrade dieses Kriteriums sich sehr stark überlappten. Entscheidend kommt jedoch hinzu, daß der Effekt der experimentellen Gruppenzugehörigkeit auf die Ausprägung von Kriterium 4 nicht mehr signifikant wurde, wenn die Einflüsse der potentiellen konfundierenden Variablen herauspartialisiert wurden. Dieser Punkt wird nachfolgend noch näher erläutert. Somit kann festgehalten werden, daß das Ergebnis zu Kriterium 4 in letzter Konsequenz **nicht** als **hypothesenkonträr** eingestuft werden kann.

Als **potentielle konfundierende Variablen**, die die Ergebnisse zur Gültigkeit der „Undeutsch-Hypothese“ hätten verfälschen können, wurden die Motivation der Pbn,

beim Ablegen der Zeugenaussage glaubhaft zu erscheinen, die subjektiv eingeschätzte persönliche Erfolgswahrscheinlichkeit beim Ablegen der Zeugenaussage, etwaige Manipulationsmaßnahmen beim Ablegen der Zeugenaussage, die intellektuelle Begabung der Pbn, etwaige Vorerfahrungen mit einem Diebstahl, Vorkenntnisse zum Thema Glaubhaftigkeitsbeurteilung sowie etwaige frühere Teilnahmen an Untersuchungen zur Glaubhaftigkeitsbeurteilung in Betracht gezogen (zur Begründung s. Abschnitt 4.3.2). Um potentielle Konfundierungen des experimentellen Gruppenfaktors Status der aussagenden Person mit den genannten Variablen zu kontrollieren, wurden letztere im Rahmen des Versuchsablaufs miterhoben und dann im Rahmen ergänzender kovarianzanalytischer Verfahren als Kovariaten berücksichtigt. Sowohl der simultane multivariate Effekt der experimentellen Gruppenzugehörigkeit auf die Ausprägungsgrade der 18 Glaubhaftigkeitskriterien als auch der univariate Effekt des Gruppenfaktors auf die Höhe der Gesamtscores blieben nach der Herauspartialisierung der Kovariaten weiterhin signifikant. Die kovarianzanalytisch geschätzten Gruppenmittelwerte der Gesamtscores unterschieden sich kaum von den empirischen Gruppenmittelwerten. Ebenso deckten sich die Ergebnisse separater univariater ANCOVAs für die einzelnen Glaubhaftigkeitskriterien weitestgehend mit den Resultaten der entsprechenden ANOVAs, d.h. die oben erläuterten Gruppenunterschiede bezüglich der Kriterien 3 (*Details*), 6 (*Gespräche*), 8 (*Ausgefallenes*) und 9 (*Nebensächliches*) erwiesen sich auch nach der Eliminierung der Kovariaten-Effekte noch als statistisch bedeutsam, wobei die empirischen Gruppenmittelwerte der Kriterien nur minimal von den entsprechenden kovarianzanalytischen Schätzwerten abwichen. Lediglich der Gruppen-Effekt bei Kriterium 4 (*Verknüpfungen*) verfehlte in der ANCOVA – im Gegensatz zur ANOVA – die korrigierte Signifikanzgrenze. Zusammenfassend sprechen die Resultate der kovarianzanalytischen Berechnungen dafür, daß die UV Status der aussagenden Person **nicht** mit den berücksichtigten Kontrollvariablen **konfundiert** war. Die einzige **Ausnahme** diesbezüglich ist das Ergebnis zu **Kriterium 4**.

Bezüglich der Kontrollvariable „intellektuelle Begabung“ kann man davon ausgehen, daß alle Pbn um ein möglichst gutes Testergebnis bemüht waren, so daß die Unverfälschtheit der Daten außer Zweifel steht. Zu den übrigen Kontrollvariablen ist allerdings kritisch anzumerken, daß sie sämtlich über einfache Befragungen der Pbn erfaßt wurden. Es ist kaum abzuschätzen, inwiefern die retrospektiven Angaben der Pbn etwa zur eigenen Motivation, zu durchgeführten Manipulationsmaßnahmen oder zu Vorerfahrungen mit Diebstählen absichtlichen oder auch irrtümlichen Verfälschungen unterlagen. Zur Kontrollvariable „subjektiv eingeschätzte persönliche Erfolgswahrscheinlichkeit beim Ablegen der Zeugenaussage“ ist ferner kritisch anzumerken, daß hierüber indirekt die subjektiven Annahmen der Pbn über die Zuverlässigkeit der inhaltsorientierten Glaubhaftigkeitsbeurteilung erfaßt werden sollten. Die retrospektiv eingeschätzte

Wahrscheinlichkeit, beim Ablegen der Zeugenaussage überzeugend gewirkt zu haben, läßt aber kaum Rückschlüsse auf die *vor* der „Vernehmung“ bestehenden Annahmen hinsichtlich der Zuverlässigkeit der Glaubhaftigkeitsbeurteilung zu. Auch hinsichtlich der Kontrollvariable „intellektuelle Begabung“ sei nochmals einschränkend in Erinnerung gerufen, daß das hierfür herangezogene Maß, der MWTB-Score, sehr grob und eigentlich nicht für individualdiagnostische Zwecke geeignet ist. Eine adäquatere Erfassung der Kontrollvariablen war jedoch angesichts des ohnehin sehr komplexen und zeitintensiven Versuchsablaufs nicht praktikabel, da dies für die Pbn mit einem kaum zumutbaren Mehraufwand verbunden gewesen wäre.

Insgesamt kann festgehalten werden, daß die „Undeutsch-Hypothese“ nur anhand der Kriterien 3 (*Details*) und 6 (*Gespräche*) uneingeschränkt verifiziert werden konnte, wobei allerdings die signifikante Differenzierung zwischen den erlebnisbezogenen und erfundenen Aussagen – rein numerisch betrachtet – eher mäßig ausgeprägt war. Auch die Ergebnisse zu Kriterium 8 (*Ausgefallenes*) lassen sich – mit Einschränkungen – als Evidenz für die Validität der „Undeutsch-Hypothese“ bewerten insofern, als dieses Kriterium in den erlebnisbezogenen Aussagen der *Zeuginnen* signifikant stärker ausgeprägt war als in den erfundenen Aussagen der *falschen Zeuginnen*. Diese Differenzierung war allerdings numerisch noch schwächer ausgeprägt als bei den Kriterien 3 und 6. Die Tatsache, daß bei Kriterium 8 der erwartete signifikante Unterschied zwischen den *Zeuginnen* und *Täterinnen* ausblieb und daß es in den Falschaussagen der *Täterinnen* signifikant stärker vorhanden war als in den Falschaussagen der *falschen Zeuginnen*, ist wohl dadurch zu erklären, daß die *Täterinnen* z.T. doch auf eine Erlebnisgrundlage zurückgreifen konnten, indem sie selber bei der Tatbegehung ausgefallene Einzelheiten bemerkten und diese anschließend in ihre Falschaussagen integrierten. Kriterium 9 (*Nebensächliches*) war zwar ebenfalls in den erlebnisbezogenen Aussagen der *Zeuginnen* signifikant – wenngleich numerisch nur geringfügig – stärker ausgeprägt als in den erfundenen Schilderungen der *falschen Zeuginnen*; dieser Befund darf jedoch in Anbetracht der defizitären Auswertungsobjektivität, welche vermutlich auf eine mangelhafte Unterweisung der Rater in der Handhabung dieses Kriteriums verweist, nicht als Evidenz für die Validität des Kriteriums bzw. für die Gültigkeit der „Undeutsch-Hypothese“ betrachtet werden. Der einzige hypothesenkonträre Befund, nämlich die signifikant stärkere Ausprägung von Kriterium 4 (*Verknüpfungen*) bei den *falschen Zeuginnen* im Vergleich zu den *Zeuginnen* – erwies sich als Resultat einer Konfundierung der UV Status der aussagenden Person mit den berücksichtigten Kontrollvariablen. Zudem war der empirisch vorgefundene Mittelwertsunterschied minimal.

Setzt man die **Ergebnisse der einzelnen Glaubhaftigkeitskriterien in Relation zu den Befunden anderer empirischer Studien**, so erscheinen die Resultate der vorlie-

genden Untersuchung auf den ersten Blick vergleichsweise ernüchternd. In den meisten anderen empirischen Studien konnten mehr als nur drei Kriterien in ihrer Validität gestützt werden. In der methodisch saubersten Feldstudie (Lamb et al., 1997) gelang die Bestätigung von fünf Einzelkriterien. In den Feldstudien von Boychuk (1991, zitiert nach Greuel et al., 1998, S. 137) sowie Esplin et al. (1988, zitiert nach Raskin & Esplin, 1991a, S. 160ff.) wurden sogar zwölf bzw. 16 Einzelkriterien in ihrer Validität untermauert. Lediglich in der Felduntersuchung von Krahé und Kundrotas (1992) ergaben sich für nur zwei Kriterien hypothesenkonforme Befunde, denen zudem ein erwartungswidriger Befund gegenüberstand (vgl. Tabelle 2 in Abschnitt 2.1.6.1). Allerdings ist die letztgenannte Studie unter methodischen Gesichtspunkten in erheblichem Maße anfechtbar (s. Abschnitt 2.1.6.1). Auch in den meisten Experimentalstudien konnten mehr validitätsstützende Befunde auf der Ebene einzelner Kriterien erzielt werden als in der vorliegenden Arbeit. Hier ist an erster Stelle die Untersuchung von Wolf und Steller (1997) mit elf bestätigten Einzelkriterien zu nennen (s. Tabelle 3 in Abschnitt 2.1.6.1). Ferner sind hier exemplarisch die Studien von Steller et al. (1992) sowie Landry und Brigham (1992) zu erwähnen, in denen neun bzw. zehn Einzelkriterien bestätigt wurden, wobei allerdings auch ein bzw. zwei hypothesenwidrige Befunde auftraten. Allerdings gab es auch einige Experimentalstudien, in denen ähnlich wenige Einzelkriterien bestätigt wurden wie in der vorliegenden Untersuchung. So konnten Vrij et al. (2000) nur vier Einzelkriterien in ihrer Validität untermauern; und bei Köhnken et al. (1995) sowie Porter und Yuille (1996) betrug die Anzahl der bestätigten Einzelkriterien jeweils nur drei.

Bei näherer Betrachtung wird jedoch deutlich, daß die **Resultate der vorliegenden Arbeit der Grundtendenz der empirischen Gesamtbefundlage entsprechen**. Wie in Abschnitt 2.1.6.1 erläutert wurde, ist Kriterium 3 (*Details*) das einzige Kriterium, dessen Validität sowohl in den Feld- als auch in den Experimentalstudien nahezu konsistent bestätigt wurde. Auch in der vorliegenden Untersuchung ergaben sich für Kriterium 3 hypothesenkonforme Unterschiede zwischen glaubhaften und unglaubhaften Aussagen. Während in den experimentellen Untersuchungen kein weiteres Kriterium regelmäßig Bestätigung fand, gelang in den Feldstudien für vier weitere Kriterien (Kriterien 2, 4, 5, 6) regelmäßig die Stützung der Validität. Eines dieser vier Kriterien, nämlich Kriterium 6 (*Gespräche*), erwies sich auch in der vorliegenden Untersuchung als valide. Letzterer Befund steht allerdings im Kontrast dazu, daß Kriterium 6 in drei der fünf bisherigen Experimentalstudien keine Bestätigung fand. Kriterium 8 (*Ausgefallenes*) konnte weder in den Feld- noch in den Experimentalstudien konsistent bestätigt werden. Zwei von vier Feldstudien sprechen für die Validität dieses Kriteriums, wobei allerdings auch ein hypothesenwidriger Befund zu vermerken ist (vgl. Tabelle 2 in Abschnitt 2.1.6.1). In den acht Experimentalstudien, in denen dieses Kriterium über-

prüft wurde, ergaben sich jedoch immerhin fünf validitätsstützende Befunde, denen nur ein hypothesenkonträres Resultat gegenübersteht (vgl. Tabelle 3 in Abschnitt 2.1.6.1). Insofern entspricht das Ergebnis der vorliegenden Untersuchung bezüglich Kriterium 8 der Tendenz der diesbezüglichen experimentellen Befundlage.

Die Tatsache, daß in der vorliegenden Untersuchung neben den Kriterien 3 (*Details*), 6 (*Gespräche*) und 8 (*Ausgefallenes*) keine weiteren Kriterien im Sinne der „Undeutsch-Hypothese“ zwischen erlebnisbezogenen und erfundenen Aussagen differenzierten, steht also durchaus nicht in krassem Widerspruch zur sonstigen empirischen Gesamtbefundlage. Dennoch sollen nachfolgend **potentielle Gründe für die mangelnde Differenzierungsfähigkeit der einzelnen Kriterien** in der vorliegenden Studie diskutiert werden, und zwar unter der hypothetischen Annahme, daß diese Kriterien grundsätzlich sehr wohl im Sinne der „Undeutsch-Hypothese“ zwischen erlebnisbezogenen und erfundenen Schilderungen zu differenzieren vermögen.

Kriterium 1 (*Konsistenz*) war in allen drei experimentellen Gruppen maximal ausgeprägt. Hier lag also gewissermaßen ein Deckeneffekt vor, der möglicherweise auf das einfache Aussagethema zurückzuführen ist. Offensichtlich war es für die falschaussagenden Gruppen nicht schwierig, auch aus der Phantasie heraus einen einfachen Gelddiebstahl zu schildern, ohne sich dabei in logische Widersprüche zu verwickeln. Hinsichtlich der mangelnden Differenzierung bei **Kriterium 2 (*Unordnung*)** kann man umgekehrt argumentieren. Der diesbezüglich vorgefundene Bodeneffekt (alle drei Gruppen erzielten sehr niedrige Werte auf diesem Kriterium) könnte damit zusammenhängen, daß das aussagerelevante Geschehen (die experimentelle Simulation eines Gelddiebstahls) vom Komplexitätsgrad und von der zeitlichen Ausdehnung her eher karg war. Dies hat möglicherweise dazu geführt, daß die *Zeuginnen* in ihrer Beschreibung des Tathergangs weitestgehend nicht von einer strukturierten und chronologischen Darstellungsweise abweichen mußten. Wie oben dargestellt wurde, zeigte sich bei **Kriterium 4** eine hypothesenwidrige Differenzierung zwischen den *Zeuginnen* und den *falschen Zeuginnen*. Letztere produzierten in ihren Aussagen mehr ***Verknüpfungen***, was sich allerdings als Ergebnis einer Konfundierung herausstellte. Jedoch selbst, wenn der hypothesenkonträre Befund nicht auf einer Effektvermengung beruht hätte, wäre das höhere Ausmaß an *Verknüpfungen* bei den *falschen Zeuginnen* im Vergleich zu den *Zeuginnen* leicht erklärbar gewesen. Kriterium 4 bezieht sich darauf, ob die Kernhandlung der Aussage mit bestimmten örtlichen oder zeitlichen Gegebenheiten, bestimmten eigenen Gewohnheiten oder Gewohnheiten von Personen aus dem sozialen Umfeld der aussagenden Person verflochten ist. Die Kernhandlung in den Aussagen der *Zeuginnen* (beobachteter Gelddiebstahl) basierte zwar auf einer Erlebnisgrundlage. Dieses zugrundeliegende Erlebnis war jedoch völlig losgelöst bzw. isoliert von den tatsächlichen all-

gemeinen Lebensumständen der *Zeuginnen*. So wurden die *Zeuginnen* ja per Instruktion aufgefordert, sich mental in die Lage einer „Reinigungskraft im Psychologischen Institut“ hineinzuversetzen, um dann bei der Ausübung dieser Rolle *Zeugin* eines Diebstahls zu werden. Insofern geschah die Kernhandlung in einer artifiziellen Situation, die logischerweise nicht mit den Umständen aus dem „wirklichen Leben“ der *Zeuginnen* verwoben sein konnte. Entsprechende *Verknüpfungen* mit den allgemeinen Lebensumständen konnten von den *Zeuginnen* allenfalls erfunden werden. Im Gegensatz dazu waren die *falschen Zeuginnen* bei der Ausschmückung ihrer erfundenen Aussagen nicht so sehr durch das tatsächliche Geschehen (die Diebstahlsimulation) festgelegt, zumal sie nicht explizit aufgefordert wurden, sich in eine bestimmte Rolle hineinzuversetzen, die von ihren wirklichen allgemeinen Lebensumständen abgekoppelt war. Folglich war es für die *falschen Zeuginnen* einfacher als für die *Zeuginnen*, die Schilderung des Diebstahls mit den eigenen Lebensumständen zu verflechten.

Kriterium 5 (*Interaktionen*), welches nicht zwischen den experimentellen Gruppen differenzierte, war in allen drei Aussagegruppen nur in sehr geringem Maß vorhanden. Der Ablauf der experimentellen Simulation war jedoch gezielt so konzipiert, daß es zu *Interaktionen* zwischen *Täterinnen* und *Zeuginnen* kommen konnte. Allerdings wurde im Rahmen des Versuchsablaufs nicht miterfaßt, inwiefern während der Diebstahlsimulation tatsächlich Ketten sich gegenseitig bedingender bzw. sich aufeinander beziehender Handlungen von *Täterinnen* und *Zeuginnen* auftraten. Deren Erfassung hätte idealerweise erfolgen können, indem man die Diebstahlsimulation jeweils mittels einer versteckten Videokamera aufgezeichnet und später auf das Auftreten von *Interaktionen* hin analysiert hätte (von ethischen und datenschutzrechtlichen Bedenken sei hier einmal abgesehen). Hiermit ist ein Punkt angesprochen, der nicht nur für die Auswertung von Kriterium 5, sondern prinzipiell für die Auswertung aller Glaubhaftigkeitskriterien relevant ist. Ob sich die einzelnen Kriterien in erlebnisbasierenden Schilderungen manifestieren können, hängt in entscheidendem Maße auch davon ab, ob die Kriterien in dem der Aussage zugrundeliegenden Geschehen auch tatsächlich realisiert waren. Treten bei einem Tathergang beispielsweise keine Interaktionen, Komplikationen oder ungewöhnlichen Einzelheiten auf, so kann auch der aufrichtigste und zuverlässigste Zeuge in seiner Aussage nicht die entsprechenden Glaubhaftigkeitskriterien (im Beispiel die Kriterien 5, 7, 8) produzieren. Diese Problematik wurde in den bisherigen empirischen Untersuchungen nicht ausreichend berücksichtigt. Im Hinblick auf die Feldforschung ist dies verständlich, da es hier äußerst schwierig ist, den tatsächlichen Tathergang genau zu rekonstruieren. Gleiches gilt für experimentelle Untersuchungen, die sich des autobiographischen Paradigmas (vgl. Abschnitt 2.1.6.1) bedienen. Im Rahmen des experimentellen Scheinverbrechen-Paradigmas ist es jedoch grundsätzlich möglich, den tatsächlichen Geschehensablauf – z.B. mittels Videoaufzeichnung – mitzuerfassen. Im

experimentellen Film-Paradigma (vgl. Abschnitt 2.1.6.1) liegt das aussagerelevante Ereignis von vorneherein in standardisierter Form (Filmaufzeichnung) vor, braucht also gar nicht eigens miterfaßt zu werden. Zukünftige experimentelle Forschung sollte ihr Augenmerk vermehrt darauf richten zu kontrollieren, inwiefern die einzelnen Kriterien im aussagerelevanten Geschehensablauf auch tatsächlich realisiert waren. Letztlich läßt sich ein Kriterium nämlich nur dann als invalider Indikator des Erlebnisbezugs einstufen, wenn sein Fehlen in erlebnisbezogenen Schilderungen nicht auf ein Fehlen des Kriteriums im tatsächlichen Geschehensablauf zurückgeführt werden kann. Im vorliegenden Fall ist es jedoch denkbar, daß Interaktionen zwischen *Täterinnen* und *Zeuginnen* primär auf der verbalen Ebene stattfanden, so daß sie unter Kriterium 6 (*Gespräche*) zu kodieren waren. Wie oben dargestellt wurde, differenzierte Kriterium 6 hypothesenkonform zwischen den experimentellen Gruppen.

Kriterium 7 (*Komplikationen*) liegt vor, wenn es in der Schilderung für den Täter durch plötzlich erscheinende Personen oder andere Umstände nötig wird, die Tat zu unterbrechen oder abzubrechen. Die Simulation war jedoch so angelegt, daß der Diebstahl relativ problemlos bewältigt werden konnte. Die einzige potentielle Komplikation für die *Täterinnen* war die Anwesenheit der *Zeuginnen* am Tatort. Allerdings wurden die *Täterinnen* per Instruktion explizit auf die Gegenwart der *Zeuginnen* vorbereitet, so daß letztere kein unvorhergesehenes Hindernis darstellten. Insofern ist es nachvollziehbar, daß die erlebnisbasierenden Aussagen der *Zeuginnen* ebenso wie die Falschaussagen der beiden anderen Gruppen nahezu keine *Komplikationen* aufwiesen.

Die fehlende Differenzierung bei **Kriterium 10 (*Unverstandenes*)** ist nicht ohne weiteres nachträglich erklärbar. Die Schilderungen der *Zeuginnen* enthielten ebenso wie die der *Täterinnen* und der *falschen Zeuginnen* fast nichts *Unverstandenes*. Man hätte jedoch erwarten können, daß das von den *Täterinnen* an den Tag gelegte Suchverhalten nach der Diebesbeute von den *Zeuginnen* zwar richtig beschrieben, jedoch offensichtlich nicht in seiner genauen Bedeutung durchschaut wurde, daß also Kriterium 10 in den Aussagen der *Zeuginnen* zum Tragen kam. So hatten die *Täterinnen* ja eine Art „Schnitzeljagd“ nach den Ziffern der angeblichen Zahlenkombination für das Schloß der Geldkassette zu bewältigen, indem sie an verschiedenen Stellen des Büros nach Notizzetteln suchten und sich die darauf befindlichen Angaben notierten. Es ist kaum denkbar, daß sich den *Zeuginnen* der genauere Sinn dieses Verhaltens erschloß, zumal sich an der Geldkassette letztlich gar kein Zahlenschloß befand, für welches die *Täterinnen* die notierten Zahlen hätten verwenden können. Insofern ist es verwunderlich, daß die *Zeuginnen* keine höheren Scores auf Kriterium 10 erzielten. Als nächstliegende Erklärung drängt sich eine mangelhafte Raterschulung hinsichtlich Kriterium 10 auf. Hiergegen spricht jedoch die ausgezeichnete Auswertungsobjektivität dieses Kriteriums

(vgl. insbesondere Tabelle F.21 und F.26 im Anhang F). Die hohe Übereinstimmung zwischen den Ratern läßt allerdings die Möglichkeit offen, daß alle drei Rater die Definition bzw. Handhabung von Kriterium 10 gleichermaßen mißverstanden haben, was letztlich nur durch eine – in der vorliegenden Studie nicht vorgenommene – Überprüfung der Rater-Experten-Übereinstimmung geklärt werden könnte. Es sei jedoch darauf hingewiesen, daß die Validität von Kriterium 10 noch in keiner Feldstudie und in nur einer von drei experimentellen Untersuchungen gestützt werden konnte (vgl. Abschnitt 2.1.6.1). Insofern wird der diagnostische Wert dieses Kriteriums also auch durch die sonstige empirische Befundlage erheblich in Frage gestellt.

Kriterium 11 (*Indirektes*) kam in den Aussagen der *Zeuginnen* ebenso wie in den Schilderungen der falschaussagenden Gruppen so gut wie gar nicht vor. Das Kriterium ist erfüllt, wenn Ereignisse berichtet werden, die dem inkriminierten Tatgeschehen zwar ähneln, sich jedoch zu anderer Zeit und mit anderen Personen zugetragen haben. In Abschnitt 5.1 wurde gesagt, daß signifikant mehr *Zeuginnen* als *falsche Zeuginnen* Vorerfahrungen mit einem Diebstahl hatten. Gerade vor diesem Hintergrund hätte man eigentlich erwarten können, daß die Aussagen der *Zeuginnen* mehr indirekt handlungsbezogene Schilderungen in dieser Richtung beinhalteten als die der *falschen Zeuginnen*. Andererseits war jedoch auch bei den *Zeuginnen* der Anteil mit Diebstahl-Vorerfahrungen sehr gering (7 von 34), so daß der geringe Ausprägungsgrad von Kriterium 11 in dieser Gruppe nicht unbedingt überrascht. Im übrigen läßt sich hier ähnlich argumentieren wie bei Kriterium 4 (*Verknüpfungen*). Die Diebstahlsimulation und insbesondere die eigene Rolle als „Reinigungskraft im Psychologischen Institut“ stellte für die *Zeuginnen* eine artifizielle, vom wirklichen Leben abgetrennte Situation dar. Daher dürfte es ihnen relativ schwer gefallen sein, Parallelen zu autobiographischen Begebenheiten zu erkennen und in die Beschreibung des Tathergangs einfließen zu lassen.

Kriterium 12 (*Eigenseelisches*), mit welchem erfaßt wird, ob im Zusammenhang mit dem Kerngeschehen aufgetretene eigene Gedanken, Gefühle bzw. entsprechende körperliche Begleiterscheinungen beschrieben werden, war bei den *Zeuginnen* nicht stärker ausgeprägt als bei den falschaussagenden Gruppen. Der Durchschnittswert der *Zeuginnen* auf diesem Kriterium entsprach der Skalenstufe *schwach vorhanden* und lag sogar tendenziell (nicht signifikant) unter den Mittelwerten der falschaussagenden Gruppen. Für diesen Befund läßt sich kaum eine plausible Erklärung anführen. Es ist durchaus davon auszugehen, daß die experimentelle Simulation für die Beteiligten emotional und gedanklich involvierend war, zumal man ihnen vorher per Instruktion mitgeteilt hatte, daß die jeweils andere beteiligte „Versuchsperson“ später belastend gegen sie aussagen würde. Allerdings entzieht sich diese Annahme jeglicher Überprüfung, da die Emotionen, Gedanken und körperlichen Begleiterscheinungen der Pbn zum Zeitpunkt der

Diebstahlsimulation jetzt nicht mehr rekonstruierbar sind. Man kann spekulieren, daß die fehlende Differenzierung anhand von Kriterium 9 eventuell mit einer unausgereiften Befragungstechnik der beiden Interviewerinnen in der vorliegenden Studie zusammenhängt. Diese waren vor der Untersuchung nicht eigens in der Interviewführung geschult worden und verstanden es deshalb möglicherweise nicht, durch ihre Fragen die für Kriterium 12 relevanten Gedächtnisinhalte der *Zeuginnen* in adäquater Weise zu aktivieren. Die Problematik der fehlenden Interviewerschulung läßt sich nicht nur in bezug auf Kriterium 12, sondern prinzipiell auch für alle anderen (insbesondere die nicht bestätigten) Kriterien anführen. Auf diesen Punkt wird weiter unten noch näher eingegangen. Was Kriterium 12 betrifft, so fanden sich jedenfalls in anderen empirischen Studien einige Hinweise auf dessen Validität. So konnte es immerhin in drei von vier Feldstudien und in drei von fünf experimentellen Untersuchungen bestätigt werden (allerdings auch ein hypothesenkonträrer experimenteller Befund; s. Abschnitt 2.1.6.1).

Auch **Kriterium 13 (*Fremdseelisches*)** differenzierte nicht zwischen den glaubhaften und unglaubhaften Aussagen, wobei die Mittelwerte aller drei experimentellen Gruppen zwischen den Skalenstufen *nicht vorhanden* und *schwach vorhanden* lagen. Der nächstliegende Grund für die niedrigen Scores der *Zeuginnen* auf diesem Kriterium ist, daß es sich bei der anderen beteiligten „Versuchsperson“ (*Täterin*) in Wirklichkeit jeweils um eine „Strohfrau“ handelte. Es ist denkbar, daß bei der „Strohfrau“, die immerhin 36 mal die Rolle der *Täterin* spielte, nach einigen Versuchsdurchgängen eine gewisse Routine einkehrte. Dies könnte dazu geführt haben, daß sie bei der Ausübung ihrer Rolle relativ unemotional wirkte. In diesem Zusammenhang ist auch zu erwähnen, daß es sich bei der „Strohfrau“ nicht um eine professionell ausgebildete Schauspielerin handelte. Sie war zwar insofern geübt, als sie vor der vorliegenden Studie im Rahmen eines experimentellen Praktikums bereits mehrfach die Rolle der *Täterin* gespielt hatte; gleichwohl wurde die Authentizität ihrer Darbietung nie von Schauspiel-Sachverständigen evaluiert. Auch im Rahmen der Versuchsdurchführung wurde nicht erfaßt, ob die Pbn in der „Strohfrau“ wirklich eine andere naive Versuchsperson sahen. Zumindest machte jedoch keine Versuchsteilnehmerin im abschließenden Aufklärungsgespräch diesbezüglich spontan Zweifel geltend. Im übrigen wurde Kriterium 13 auch in anderen empirischen Arbeiten nicht sehr oft bestätigt. Nur in einer von vier Feldstudien und in zwei von fünf Experimentalstudien resultierten hypothesenkonforme Befunde (außerdem ein erwartungskonträres experimentelles Resultat, vgl. Abschnitt 2.1.6.1).

Die **Kriterien 14 (*Verbesserungen*)** und **15 (*Erinnerungslücken*)** wurden in der vorliegenden Studie nicht als valide eingestuft. Es muß jedoch erwähnt werden, daß diese negativen Resultate teilweise auch durch die konservative inferenzstatistische Vorgehensweise in der vorliegenden Arbeit bedingt sind. Wie oben erläutert, wurde aufgrund

der festgestellten Interdependenz zwischen den einzelnen Kriterien bei den statistischen Entscheidungen in den 18 separaten ANOVAs bzw. ANCOVAs jeweils ein korrigiertes Alpha-Niveau zugrunde gelegt ($\alpha' = .05/18$). Dieses Vorgehen kann man als sehr konservativ betrachten. Hätte man auf die Alpha-Adjustierung verzichtet, wie dies z.B. Steller et al. (1992) taten, so wären die ANOVA- bzw. ANCOVA-Effekte bezüglich der Kriterien 14 und 15 signifikant geworden (s. Tabelle F.33 und F.55 im Anhang F), und die Ergebnisse der Anschlußtests an die ANOVAs (Scheffé-Tests) hätten jeweils auf statistisch bedeutsame hypothesenkonforme Gruppenunterschiede zwischen den *Zeuginnen* und *falschen Zeuginnen* verwiesen. Kriterium 15 (*Erinnerungslücken*) war sowohl bei den *Zeuginnen* als auch bei den *falschen Zeuginnen* so gut wie gar nicht vorhanden (M = 0.27 bzw. 0.09), so daß der (auf dem unkorrigierten Alpha-Niveau) signifikante Gruppenunterschied – numerisch betrachtet – wenig substantiell ist. Dagegen war Kriterium 14 (*Verbesserungen*) in den erlebnisbezogenen Aussagen der *Zeuginnen* immerhin *schwach* bis *mittel vorhanden* (M = 1.46), während es in den konfabulierten Schilderungen der *falschen Zeuginnen* nur *schwach vorhanden* war (M = 1.08). Allerdings kann auch das Ausmaß dieser Mittelwertsdifferenz allenfalls als moderat eingestuft werden, zumal die Häufigkeitsverteilungen der beiden Gruppen bezüglich der Ausprägungsgrade von Kriterium 14 sich weitgehend überschnitten. Die sonstige empirische Befundlage zu den Kriterien 14 und 15 ist durchwachsen. Kriterium 14 wurde jeweils durch die Hälfte der Feld- und Experimentalstudien in seiner Validität untermauert. Hinsichtlich Kriterium 15 ergaben sich in einer von drei Felduntersuchungen und in vier von acht Laborstudien validitätsstützende Befunde (s. Abschnitt 2.1.6.1).

Die **Kriterien 16 (*Selbsteinwände*)** und **17 (*Eigenbelastung*)** waren jeweils in den Aussagen aller drei Gruppen praktisch nicht vorhanden und konnten dementsprechend auch nicht zwischen glaubhaften und unglaubhaften Bekundungen differenzieren. Das Fehlen beider Kriterien in den Schilderungen der *Zeuginnen* dürfte in erster Linie auf den Aussagegegenstand (simulierter Diebstahl) und die Aussageumstände zurückzuführen sein. Die äußeren Wahrnehmungsbedingungen für die *Zeuginnen* zum Tatzeitpunkt waren nahezu optimal, ihr Erregungsniveau zum Tatzeitpunkt dürfte eher moderat gewesen sein, und zwischen der Tatbeobachtung und der Zeugenaussage lag nur eine kurze Zeitspanne. Insofern war das Entstehen von Wahrnehmungs- und Gedächtnisunsicherheiten seitens der *Zeuginnen* eher unwahrscheinlich, so daß letztere von daher möglicherweise gar keine Veranlassung hatten, Zweifel hinsichtlich der Richtigkeit der eigenen Aussage (Kriterium 16) zu äußern. Bei dem Tathergang selbst waren die *Zeuginnen* in erster Linie Beobachter. Zwar bestand durchaus die Möglichkeit, daß es zu Interaktionen zwischen *Täterin* und *Zeugin* kam; die Wahrscheinlichkeit, daß die *Zeuginnen* durch ihr Verhalten das Gelingen des Diebstahls in irgendeiner Weise begünstigten, war jedoch minimal, insbesondere weil den *Zeuginnen* ja vorher mitgeteilt wurde, daß die andere

„Versuchsperson“ etwas zu stehlen beabsichtige und daß die *Zeugin* später selber der Täterschaft verdächtigt würde. Insofern bestand für die *Zeuginnen* kaum Veranlassung, in selbstkritischer Weise Unvorteilhaftes über das eigene Verhalten (Kriterium 17) zu berichten. Im Gegenteil waren sie ja vorher gezielt instruiert worden, mit ihrer Zeugenaussage auch zur eigenen Entlastung beizutragen. Zudem übernahmen die *Zeuginnen* im Rahmen der Simulation eine künstliche Rolle („Reinigungskraft im Psychologischen Institut“), die losgelöst von ihrer wirklichen Autobiographie war. Daher war es für die *Zeuginnen* auch nicht naheliegend, Unvorteilhaftes über die eigene Persönlichkeit oder Beschreibungen früheren Fehlverhaltens (Kriterium 17) in die Beschreibung des simulierten Diebstahls einfließen zu lassen. Anderweitige validitätsstützende Befunde zu den Kriterien 16 und 17 sind rar. Während für Kriterium 16 (*Selbsteinwände*) erst ein einziger validitätsstützender Befund (aus einer Experimentalstudie) vorliegt, konnte die Gültigkeit von Kriterium 17 (*Eigenbelastung*) noch gar nicht bestätigt werden (allerdings zwei hypothesenkonträre experimentelle Befunde; vgl. Abschnitt 2.1.6.1).

Die fehlende Differenzierung der experimentellen Gruppen anhand von **Kriterium 18 (*Fremdentlastung*)** ist wohl durch die experimentelle Hintergrundgeschichte begründet. Wie bei Kriterium 17 ist auch hier anzuführen, daß die *Zeuginnen* per Instruktion gewissermaßen in eine kompetitive Situation hineinversetzt wurden, in welcher es darauf ankam, die *Täterin* möglichst stark zu belasten, um so zur eigenen Entlastung beizutragen. Daher ist es nicht verwunderlich, daß in den Aussagen der *Zeuginnen* Entlastungen der Angeschuldigten ebenso selten vorkamen wie in den Bekundungen der beiden anderen Gruppen. Allerdings ist es bislang auch in anderen experimentellen Studien noch nicht gelungen, die Validität von Kriterium 18 nachzuweisen; und in nur einer von drei Feldstudien wurde das Kriterium bestätigt (vgl. Abschnitt 2.1.6.1).

Abschließend soll noch auf einige **problematische Aspekte** der vorliegenden Untersuchung eingegangen werden, nämlich den **Komplexitätsgrad des aussagerelevanten Geschehensablaufs**, die Art der **Interviewführung**, die **Länge der Aussagen** und das verwendete **interindividuelle Versuchsdesign**. Die inhaltsorientierte Glaubhaftigkeitsbeurteilung ist nur sinnvoll anwendbar, wenn die Aussage sich auf einen einigermaßen **komplexen Geschehensablauf** bezieht. Die Aussagen in forensischen Realfällen richten sich zumeist auf inzestuöse Beziehungen, welche sich nicht selten über mehrere Monate oder gar Jahre erstreckten. Verglichen damit war der aussagerelevante Geschehensablauf in der vorliegenden Studie – der simulierte Gelddiebstahl – eher karg strukturiert. Insofern kann man argumentieren, daß die insgesamt ernüchternden Befunde der vorliegenden Studie zur Gültigkeit der „Undeutsch-Hypothese“ bzw. zur Validität der *Kriterienorientierten Inhaltsanalyse* teilweise durch die relative Kargheit des experimentell simulierten Geschehensablaufs bedingt sind bzw. daß sich bei einem komplexe-

ren Tathergang möglicherweise eine deutlichere Differenzierung zwischen den erlebnisbezogenen und erfundenen Aussagen ergeben hätte. Mit der Problematik der Komplexität des Geschehensablaufs ist zugleich die grundsätzliche Frage der externen Validität der vorliegenden Untersuchung tangiert. Wie in Abschnitt 2.1.6.1 erörtert wurde, lassen sich in bezug auf die inhaltsorientierte Glaubhaftigkeitsbeurteilung vier Forschungsansätze unterscheiden. Neben der Feldforschung gibt es drei experimentelle Ansätze, nämlich das Film-Paradigma, das autobiographische Paradigma und das Scheinverbrechen-Paradigma. Die optimistischsten Befunde zur Gültigkeit der „Undeutsch-Hypothese“ bzw. zur Validität der *Kriterienorientierten Inhaltsanalyse* ergaben sich – abgesehen von den Feldstudien – im Rahmen des autobiographischen Paradigmas. In Abschnitt 2.1.6.1 wurde argumentiert, daß dies möglicherweise mit einer höheren externen Validität des autobiographischen Paradigmas im Vergleich zum Film- und zum Scheinverbrechen-Paradigma zusammenhänge. Prinzipiell ist es jedoch nicht zulässig, die externe Validität einer Untersuchung daran zu messen, ob sie hypothesenkonforme Ergebnisse erbringt. Der vorliegenden Studie ist jedenfalls zu gute zu halten, daß die psychologische Situation der *Zeuginnen* während der Diebstahlsimulation zumindest ansatzweise durch Eigenbeteiligung und negative emotionale Tönung geprägt war, also durch zwei jener drei Charakteristika (3. Merkmal: Kontrollverlust), die z.B. Raskin und Esplin (1991a) sowie Steller et al. (1992) als typisch für die psychologische Situation von Opferzeugen und somit als entscheidend im Hinblick auf die externe Validität experimenteller Studien zur inhaltsorientierten Glaubhaftigkeitsbeurteilung erachten. So waren die *Zeuginnen* der vorliegenden Studie direkt in das aussagerelevante Geschehen involviert (Eigenbeteiligung); und sie gingen bereits während der Verbrechenssimulation davon aus, daß sie von der *Täterin* zu Unrecht des Diebstahls bezichtigt würden und daß ihnen aufgrund dessen der Verlust der in Aussicht gestellten finanziellen Belohnung drohe (negative emotionale Tönung). Die Komponente des Kontrollverlusts konnte selbstverständlich nicht simuliert werden, da aus ethischen Gründen den Pbn die Möglichkeit eingeräumt werden mußte, das Experiment jederzeit abzubrechen.

Die inhaltsorientierte Glaubhaftigkeitsbeurteilung setzt auch einen gewissen **Mindestumfang der Aussage** voraus. Dieser Mindestumfang ist allerdings nicht näher spezifiziert. In der vorliegenden Studie lag die durchschnittliche Länge der Aussagen bei 686 Worten (SD = 314). Leider finden sich in Publikationen anderer empirischer Validitätsstudien kaum Angaben über die Länge der analysierten Aussagen, die man als Vergleichsmaßstab heranziehen könnte. Lediglich Ruby und Brigham (1998) sowie Winkel und Vrij (1995) machen diesbezüglich genaue Angaben, indem sie den durchschnittlichen Umfang ihrer experimentell gewonnenen Aussagen auf 255 bzw. 214 Worte beziffern. In Relation dazu ist die durchschnittliche Aussagelänge in der vorlie-

genden Untersuchung also durchaus groß, wobei allerdings zu berücksichtigen ist, daß die Aussagen bei Ruby und Brigham (1998) in Englisch, bei Winkel und Vrij (1995) in Holländisch erfolgten. Weitere, jedoch wesentlich unpräzisere Angaben finden sich ferner bei Krahé und Kundrotas (1992) sowie Landry und Brigham (1992), in deren Publikationen von durchschnittlich drei Schreibmaschinenseiten bzw. eineinhalb bis zwei Minuten Aussagelänge die Rede ist.

In der vorliegenden Untersuchung bestanden überzufällige **Gruppenunterschiede in bezug auf den Aussageumfang**. Eine einfaktorielle ANOVA ergab einen signifikanten Effekt der experimentellen Gruppenzugehörigkeit auf die Wortzahl der Aussagen, $F(2,99) = 7.979$, $p < .01$ (s. genauer Tabelle F.111 im Anhang F). Scheffé-Tests (s. Tabelle F.112 im Anhang F) zufolge beruhte dieser Effekt darauf, daß die Aussagen der *Zeuginnen* mit durchschnittlich 825 Worten (SD = 303) signifikant länger waren als die der *falschen Zeuginnen* (M = 540, SD = 215), $p < .01$. Der Aussageumfang hing auch mit der Art der Interviewführung zusammen. Aus organisatorischen Gründen wurden in der vorliegenden Untersuchung zwei Personen als Versuchsleiter 2 (die angebliche „gerichtspsychologische Expertin“) eingesetzt, wobei die Zuteilung der beiden Versuchsleiterinnen zu den Pbn nach dem Zufallsprinzip erfolgte. Eine einfaktorielle ANOVA ergab, daß die beiden **Interviewerinnen sich signifikant bezüglich der Länge der evozierten Aussagen unterschieden**, $F(1,100) = 4.736$, $p < .05$ (s. genauer Tabelle F.113 im Anhang F). Bei Interviewerin A, die 54 Befragungen durchführte, betrug die durchschnittliche Aussagelänge 749 Worte (SD = 333); dagegen resultierte in den 48 von Interviewerin B durchgeführten Befragungen ein mittlerer Aussageumfang von nur 616 Worten (SD = 278). Die **Wechselwirkung** der experimentellen **Gruppenzugehörigkeit der Aussagen** und des Faktors **Interviewerin** (A vs. B) **auf die Länge der Aussagen** erwies sich in einer zweifaktoriellen ANOVA als **nicht signifikant** (s. genauer Tabelle F.114 im Anhang F), d.h. bei beiden Interviewerinnen fielen die Aussagen der *Zeuginnen* signifikant länger aus als die der *falschen Zeuginnen*, und alle drei experimentellen Gruppen machten bei Interviewerin A umfangreichere Aussagen als bei Interviewerin B.

Die unterschiedlichen Aussageumfänge bei den beiden Interviewerinnen hängen möglicherweise damit zusammen, daß in der vorliegenden Studie **kein intensives Interviewertraining** stattfand. Der Schwerpunkt der semistandardisierten Befragung lag auf der Evozierung eines möglichst umfangreichen freien Berichts über den gesamten Tathergang. (Zu diesem Zweck wurden die Pbn vor dem Interview in einer schriftlichen Instruktion gezielt darauf hingewiesen, daß dem freien Bericht besondere Bedeutung hinsichtlich der Beurteilung ihrer Glaubhaftigkeit zukomme.) Die Aufgabe der Interviewerinnen bestand in erster Linie darin, nachzufragen, sofern sich Unklarheiten im

Spontanbericht ergaben oder die Pbn Schwierigkeiten hatten, der Aufforderung zum freien Erzählen nachzukommen. Ferner sollten die Interviewerinnen im Anschluß an den freien Bericht zunächst möglichst offene Fragen zum Aussehen der *Täterin* und zur Beschaffenheit des Tatorts stellen. Schließlich sollten sie den Pbn noch die Möglichkeit zur Ergänzung ihrer Bekundungen geben. Bei der Betrachtung der Interviewtranskripte stellte sich heraus, daß Interviewerin B die Pbn gegen Ende der jeweiligen Befragungen seltener zu etwaigen Aussageergänzungen aufforderte als Interviewerin A. Dies dürfte der Grund für die alles in allem kürzeren Aussagen bei Interviewerin B sein.

Der Befragungsstil beeinflußt jedoch prinzipiell nicht nur den Aussageumfang, sondern auch die Anzahl bzw. Ausprägung der produzierten Glaubhaftigkeitskriterien und möglicherweise sogar die diesbezügliche Differenzierung zwischen erlebnisbezogenen und erfundenen Aussagen (Lamb, 1998). So fanden Hershkowitz, Lamb, Sternberg und Esplin (1997) ebenso wie Craig et al. (1999) bei der Analyse von Befragungen kindlicher Zeugen in forensischen Realfällen heraus, daß offene Fragen alles in allem längere, detailreichere und kriterienhaltigere Aussagen nach sich ziehen als gezielte Fragen. Lamb, Sternberg, Esplin, Hershkowitz und Orbach (1997) führen überdies Untersuchungsbefunde aus dem eigenen Arbeitskreis an, die darauf hindeuten, daß in erlebnisbezogenen Aussagen offene Fragen mehr Glaubhaftigkeitskriterien evozieren als gezielte Fragen, wohingegen dieser Effekt in erfundenen Schilderungen nicht auftreten soll. Mit Ausnahme der Bedingung *falsche Zeuginnen* waren die experimentellen Gruppen der vorliegenden Untersuchung im Hinblick auf die beiden Interviewerinnen nicht „ausbalanciert“. Bei den *Täterinnen* wurden die Befragungen 18 mal von Interviewerin A und 16 mal von Interviewerin B durchgeführt. Bei den *Zeuginnen* kam 19 mal Interviewerin A und 15 mal Interviewerin B zum Einsatz. Nur bei den *falschen Zeuginnen* war das Verhältnis mit 17:17 ausgeglichen. Um auszuschließen, daß die Befunde zur Differenzierungsfähigkeit der Glaubhaftigkeitskriterien bzw. Gültigkeit der „Undeutsch-Hypothese“ möglicherweise durch eine Konfundierung des Gruppenfaktors mit dem Faktor Interviewerin (A vs. B) verfälscht waren, wurde letztgenannte Variable im Rahmen kovarianzanalytischer Verfahren als Kovariate berücksichtigt. Es wurden eine einfaktorielle MANCOVA über alle 18 Glaubhaftigkeitskriterien, eine einfaktorielle ANCOVA der Gesamtscores sowie 18 separate einfaktorielle ANCOVAs für die einzelnen Glaubhaftigkeitskriterien gerechnet. Nach der Herauspartialisierung der Variable Interviewerin ergaben sich die gleichen signifikanten Effekte wie in den ursprünglichen statistischen Analysen, und auch die Effektstärken waren nahezu identisch (s. Tabelle F.115, F.116 und F.117 im Anhang F). Zudem unterschieden sich die kovarianzanalytischen Schätzwerte fast überhaupt nicht von den empirisch vorgefundenen Gruppenmittelwerten der 18 Glaubhaftigkeitskriterien und des Gesamtscores.

Auch wenn eine **Konfundierung mit dem Faktor Interviewerin** somit **auszuschließen** ist, könnte man doch den geringen Trainingsstand beider Interviewerinnen bemängeln. So wurden etwa die Interviewer in dem Scheinverbrechen-Experiment von Porter und Yuille (1996) intensiv in einer in der nordamerikanischen forensischen Praxis verbreiteten Befragungstechnik (sog. „Step-Wise Interview", S. 448) geschult, bevor sie die experimentellen Interviews durchführten. Die Befragungstechnik bei Porter und Yuille (1996) war jedoch letztlich der in der vorliegenden Studie sehr ähnlich – der Schwerpunkt lag auf der Evozierung eines freien Berichts. Die Tatsache, daß die *Kriterienorientierte Inhaltsanalyse* bei Porter und Yuille (1996) nicht besser zwischen erlebnisbezogenen und erfundenen Aussagen differenzierte als in der vorliegenden Untersuchung, deutet darauf hin, daß die relativ dürftige Interviewerschulung in der vorliegenden Studie sich nicht negativ auf die resultierenden Validitätsbefunde auswirkte.

In Abschnitt 2.1.1 wurde gesagt, daß die „Undeutsch-Hypothese" sich grundsätzlich auf den intraindividuellen Vergleich erlebnisbezogener und erfundener Schilderungen bezieht. In der vorliegenden Studie wurde die Gültigkeit der Grundannahme der inhaltsorientierten Glaubhaftigkeitsbeurteilung jedoch anhand des Vergleichs unabhängiger Gruppen von wahren und falschen Aussagen, also in einem **interindividuellen Versuchsdesign**, überprüft. Streng genommen könnte man daher argumentieren, daß die schwachen Validitätsbefunde auf der Unangemessenheit des interindividuellen Designs beruhen. In sämtlichen Feldstudien und in der Mehrzahl der experimentellen Arbeiten zur Validierung der „Undeutsch-Hypothese" bediente man sich allerdings ebenfalls eines interindividuellen Designs und erzielte dennoch mitunter sehr positive Resultate ganz im Sinne der „Undeutsch-Hypothese" (vgl. Abschnitt 2.1.6.1). Die einzigen Studien, die auf einem intraindividuellen Design basierten, sind die von Steller et al. (1992) sowie Ruby und Brigham (1998). Während Steller et al. (1992) recht eindeutige Ergebnisse im Sinne der „Undeutsch-Hypothese" erzielten, waren die Ergebnisse von Ruby und Brigham (1998) durchwachsen – der Bestätigung von sieben Glaubhaftigkeitskriterien standen fünf hypothesenkonträre Befunde gegenüber. Die Untersuchung von Ruby und Brigham (1998) ist jedoch in methodischer Hinsicht stark kritisierbar. So war die Aussagenstichprobe sehr gering (12 erlebnisbezogene und 12 erfundene Geschichten), die Auswerter hatten nur eine sehr knappe Unterweisung in der Handhabung der Glaubhaftigkeitskriterien erhalten, und der tatsächliche Wahrheitsstatus der Aussagen (erlebnisbezogen [autobiographische Begebenheit] vs. erfunden) wurde überhaupt nicht verifiziert. Ein Vergleich des intraindividuellen und des interindividuellen Ansatzes läßt sich am besten anhand der Studien von Steller et al. (1992) sowie Wolf und Steller (1997) vollziehen. In beiden Untersuchungen wurde das autobiographische Paradigma verwendet. Beide Studien wurden zudem in der gleichen Arbeitsgruppe durchgeführt, so daß davon auszugehen ist, daß sie sich in bezug auf methodische Feinheiten (Hand-

habung der Glaubhaftigkeitskriterien etc.) nicht wesentlich unterschieden. In diesem Zusammenhang ist auch zu betonen, daß an beiden Untersuchungen mit Steller ein Urheber der *Kriterienorientierten Inhaltsanalyse* beteiligt war. Die Resultate beider Untersuchungen sprechen recht eindeutig für die Validität der „Undeutsch-Hypothese“, wobei im Rahmen des interindividuellen Designs (Wolf & Steller, 1997) sogar noch mehr Einzelkriterien bestätigt wurden als im Rahmen der intraindividuellen Bedingungsvariation. Alles in allem spricht die empirische Befundlage also nicht dagegen, die Validierung der „Undeutsch-Hypothese“ bzw. der *Kriterienorientierten Inhaltsanalyse* mit einer interindividuellen Bedingungsvariation anzugehen.

6.1.3 Treffsicherheit der inhaltsorientierten Glaubhaftigkeitsbeurteilung in der vorliegenden Untersuchung

Für die inhaltsorientierte Glaubhaftigkeitsbeurteilung existieren keine verbindlichen diagnostischen Entscheidungsrichtlinien. Statt dessen erfolgt die Urteilsbildung in der Praxis auf klinisch-intuitivem Weg und ist abhängig von den individuellen Besonderheiten des konkreten Begutachtungsfalles. Auch in der vorliegenden Untersuchung wurde eine **klinisch-intuitive Gesamtbeurteilung** der Aussagen vorgenommen, d.h. die drei Auswerter gaben auf einer standardisierten, zehnstufigen Skala an, für wie glaubhaft bzw. unglaubhaft sie die Aussagen hielten. Dabei blieb es den Auswertern überlassen, in welcher Weise sie die Ergebnisse der Inhaltsanalyse in ihr Urteil einfließen ließen. Zunächst wurde überprüft, ob die klinisch-intuitiven Ratings bei den erlebnisbezogenen Aussagen höher ausfielen als bei den erfundenen. Da die klinisch-intuitive Beurteilung per definitionem nicht objektivierbar ist und somit nicht auszuschließen war, daß die drei Rater sich systematisch in ihrem diesbezüglichen Urteilsverhalten unterschieden, wurden neben der experimentellen Gruppenzugehörigkeit der Aussagen auch noch die drei Rater als varianzanalytischer Faktor berücksichtigt. Es zeigte sich in der Tat ein signifikanter Haupteffekt der experimentellen Gruppenzugehörigkeit dergestalt, daß die erlebnisbezogenen Aussagen der *Zeuginnen* von den drei Auswertern insgesamt für glaubhafter gehalten wurden als die konfabulierten Schilderungen der *Täterinnen* und der *falschen Zeuginnen*. Den Standardabweichungen war jedoch zu entnehmen, daß die Häufigkeitsverteilung der *Zeuginnen* bezüglich der Höhe der klinisch-intuitiven Beurteilung sich deutlich mit den entsprechenden Häufigkeitsverteilungen der falschaussagenden Gruppen überschnitt, so daß die Mittelwertsunterschiede im Hinblick auf eine etwaige Individualdiagnostik nicht als sehr deutlich ausgeprägt einzustufen sind. Noch entscheidender ist jedoch, daß neben dem beschriebenen Haupteffekt auch die **Wechselwirkung zwischen den Faktoren Gruppenzugehörigkeit und Rater** signifikant wurde. Nur zwei Auswerter erachteten die Aussagen der

Zeuginnen als glaubhafter als die Aussagen beider falschaussagenden Gruppen, während der dritte Auswerter nur die *Täterinnen* für weniger glaubhaft hielt als die *Zeuginnen*, den *falschen Zeuginnen* jedoch ebensoviel Glaubhaftigkeit zusprach wie den *Zeuginnen*. Überdies zeigte sich auch noch ein Rater-Haupteffekt; die Rater A und C hielten die Aussagen insgesamt für signifikant glaubhafter als Rater B. Dabei ist auch noch anzumerken, daß alle drei Rater jeweils in ihrem durchschnittlichen Urteil über alle drei Gruppen hinweg eher in Richtung glaubhaft tendierten, was angesichts der doppelt so hohen Basisrate unglaubhafter im Vergleich zu glaubhaften Aussagen umso bemerkenswerter ist.

Um die diagnostischen Treffer- bzw. Fehlerquoten bestimmen zu können, wurden die auf der zehnstufigen Skala vorgenommenen Ratings dichotomisiert. Die resultierenden Diagnosen („glaubhaft“ vs. „unglaubhaft“) wurden dann zur tatsächlichen Glaubhaftigkeit der Aussagen in Beziehung gesetzt. Die erlebnisbezogenen Aussagen der *Zeuginnen* wurden von allen drei Auswertern mit einer deutlich überzufälligen Häufigkeit zutreffend als glaubhaft diagnostiziert. Im Gegensatz dazu diagnostizierten alle drei Auswerter jeweils die erfundenen Schilderungen der *falschen Zeuginnen* häufiger unzutreffend als zutreffend. Insbesondere Rater A stufte fast alle *falschen Zeuginnen* unzutreffend als glaubhaft ein. Auch die Aussagen der *Täterinnen* wurden alles in allem öfter falsch als richtig klassifiziert. Allerdings lag dies nur an den Auswertern A und C, wohingegen Auswerter B die *Täterinnen* mit einer deutlich überzufälligen Häufigkeit korrekt als unglaubhaft diagnostizierte. Die basisratenkorrigierten Gesamttrefferquoten der Rater B und C lagen jeweils deutlich über dem Zufallsniveau, die von Rater A lag hingegen nur knapp über 50%. Insgesamt sprechen die Treffer- bzw. Fehlerraten für eine deutliche **Fehlertendenz in Richtung falsch positiver Entscheidungen**, d.h. unglaubhafte Aussagen wurden im Vergleich zu glaubhaften Aussagen viel eher fehldiagnostiziert. Diese Fehlertendenz korrespondiert mit dem oben angesprochenen Befund, daß die Urteile aller drei Rater auf der zehnstufigen Skala insgesamt in Richtung glaubhaft tendierten.

Die vorgefundene Fehlertendenz in Richtung falsch positiver Entscheidungen steht im Einklang mit den Ergebnissen anderer empirischer Studien, in denen die Treffsicherheit der inhaltsorientierten Glaubhaftigkeitsbeurteilung überprüft wurde. In den in Abschnitt 2.1.6.2 zitierten Untersuchungen mit klinisch-intuitiver Urteilsbildung wurden die erfundenen Aussagen durchschnittlich zu 48% fehldiagnostiziert (falsch positive), gegenüber durchschnittlich 22% Fehldiagnosen bei den erlebnisbasierenden Aussagen (falsch negative). Im Vergleich dazu war die diagnostische Urteilsverzerrung der klinisch-intuitiven Beurteilung in der vorliegenden Studie sogar noch viel deutlicher ausgeprägt,

mit 61.3% falsch positiven (Mittelwert der *Täterinnen* und *falschen Zeuginnen*) gegenüber 11.8% falsch negativen Entscheidungen.

Die (basisratenkorrigierte) Gesamttrefferquote der klinisch-intuitiven Diagnosen in der vorliegenden Studie (63.5%) entsprach in etwa der durchschnittlichen Gesamttrefferquote, die in den in Abschnitt 2.1.6.2 zitierten Studien zur klinisch-intuitiven Urteilsbildung erreicht wurde (65%), insbesondere wenn man bedenkt, daß letztere leicht nach oben verzerrt ist, insofern, als es in zwei der drei Studien auch noch die Urteilskategorie „unentscheidbar“ gab, die „unentscheidbaren“ Fälle (in beiden Studien jeweils ca. 10%) jedoch nicht bei der Berechnung der Trefferquoten berücksichtigt wurden. Die Frage, warum in der vorliegenden Untersuchung die Trefferquote bei den glaubhaften Aussagen höher und bei den unglaubhaften Aussagen niedriger war als im Durchschnitt der oben zitierten empirischen Arbeiten, läßt viel Raum für Spekulation. So waren z.B. die drei Auswerter in der vorliegenden Studie nicht über die tatsächlichen Häufigkeiten von wahren und falschen Aussagen informiert. Möglicherweise gingen die Rater implizit von gleichen Basisraten beider Aussagetypen aus und tendierten deshalb bei den Falschaussagen häufiger zu „Glaubhaft“-Diagnosen als sie es getan hätten, wenn sie von der in Wirklichkeit höheren Basisrate erfundener Aussagen gewußt hätten. Die irrtümliche Annahme, in etwa genauso viele „Glaubhaft“- wie „Unglaubhaft“-Diagnosen vergeben zu müssen, könnte schließlich zu einer artifiziellen Erhöhung der Rate valide positiver Entscheidungen bzw. Senkung der Rate valide negativer Entscheidungen geführt haben. Allerdings wurde nicht erfragt, ob die Auswerter tatsächlich von gleichen Basisraten erlebnisbezogener und erfundener Aussagen ausgingen.

Auffallend ist auch, daß die Falschaussagen der *Täterinnen* von den drei Auswertern häufiger entlarvt wurden als die Falschaussagen der *falschen Zeuginnen*. Dabei wäre eher zu erwarten gewesen, daß es den *Täterinnen* leichter fallen würde, ihren Aussagen einen Anschein von Erlebnisbezogenheit zu verleihen. Im Gegensatz zu den *falschen Zeuginnen* hatten sie ja immerhin die Möglichkeit, auf eigene reale Erfahrungen beim Tathergang zurückzugreifen bzw. die eigene Tatbegehung in der Beschreibung des Tathergangs auf die *Zeugin* zu übertragen. Insofern spricht das Ergebnis der klinisch-intuitiven Beurteilung dafür, daß die inhaltsorientierte Glaubhaftigkeitsbeurteilung für die Begutachtung von Tatverdächtigen, die sich durch Falschbezichtigung einer anderen Person selbst entlasten wollen, mindestens genauso geeignet ist wie für die Begutachtung von unaufrichtigen „Zeugen“, die nicht selber in den Tathergang involviert waren und somit keinerlei Erlebnisbezug zum Tathergang aufweisen. Die höhere diagnostische Treffsicherheit bei den *Täterinnen* im Vergleich zu den *falschen Zeuginnen* ist allerdings nicht damit erklärbar, daß die Auswerter die inhaltlichen Glaubhaftigkeitskriterien in den Aussagen der *Täterinnen* schwächer ausgeprägt gesehen hätten als in den Aussa-

gen der *falschen Zeuginnen*. Wie oben erläutert wurde, ergaben sich in den Kriterienratings keine signifikant höheren Werte seitens der *falschen Zeuginnen*; Kriterium 8 wurde sogar den *Täterinnen* in stärkerem Ausmaß zugesprochen als den *falschen Zeuginnen*. Selbst in der durchschnittlichen Höhe der auf der zehnstufigen Skala vorgenommenen klinisch-intuitiven Glaubhaftigkeitsurteile unterschied sich die Gruppe der *Täterinnen* nicht signifikant von der Gruppe der *falschen Zeuginnen*. An der Vorgehensweise in der vorliegenden Studie ist zu bemängeln, daß die Auswerter nicht aufgefordert wurden, ihre klinisch-intuitiven Urteile jeweils zu begründen. Somit muß leider offen bleiben, was letzlich ausschlaggebend dafür war, daß die Auswerter die Aussagen der *Täterinnen* häufiger zutreffend diagnostizierten als die der *falschen Zeuginnen*.

Es muß jedoch betont werden, daß das **Ergebnis der klinisch-intuitiven Glaubhaftigkeitsbeurteilung** in der vorliegenden Studie **nur sehr eingeschränkt interpretierbar** ist. So standen den Auswertern **neben dem bloßen Wortlaut der Aussagen keine weiteren diagnoserelevanten Informationen** zur Verfügung. Eine adäquate klinisch-intuitive Beurteilung muß jedoch so erfolgen, daß die anhand der Glaubhaftigkeitskriterien festgestellte inhaltliche Aussagequalität an weiteren diagnostisch relevanten Informationen (Ergebnisse von Persönlichkeits- und Motivanalysen etc.) relativiert wird, woraus sich letztlich ein Wahrscheinlichkeitsurteil hinsichtlich des Erlebnisbezugs der Aussage ergibt. Insofern könnte man argumentieren, daß die hier erzielten Hitraten die tatsächliche Treffsicherheit der klinisch-intuitiven Beurteilung unterschätzen. Allerdings wurden bis dato noch keine kontrollierten Untersuchungen publiziert, in welchen den Auswertern sämtliche diagnostisch relevanten Informationen zur Verfügung gestanden hätten. Das Vorgehen in den drei Untersuchungen mit klinisch-intuitiver Urteilsbildung (s. Abschnitt 2.1.6.2) war ähnlich wie das in der vorliegenden Untersuchung, d.h. den jeweiligen Auswertern stand für die Urteilsbildung ausschließlich der Wortlaut der Aussagen zur Verfügung. Somit entbehrt die Annahme, daß bei Vorhandensein sämtlicher diagnoserelevanter Daten eine höhere Treffsicherheit erzielt werden kann als in der vorliegenden bzw. in den in Abschnitt 2.1.6.2 angeführten Studien, bislang noch jeglicher empirischer Grundlage.

Die Aussagekraft der vorliegenden klinisch-intuitiven Trefferquoten wird auch noch durch einen weiteren Aspekt entscheidend eingeschränkt. Das zentrale Element der inhaltsorientierten Glaubhaftigkeitsbeurteilung ist die Feststellung der inhaltlichen Aussagequalität anhand der verschiedenen Glaubhaftigkeitskriterien. Insofern war zu erwarten, daß die klinisch-intuitiven Diagnosen auf der zehnstufigen Glaubhaftigkeitsskala in engem, positivem Zusammenhang standen mit den vom jeweiligen Auswerter eingeschätzten Ausprägungsgraden der 18 Kriterien, zumal den Auswertern neben dem Wortlaut der Aussagen keine weiteren Informationen zur Verfügung standen. Um dies

zu überprüfen, wurden statistische Reanalysen durchgeführt. Pro Aussage und pro Auswerter wurden die eingeschätzten Ausprägungsgrade aller 18 Glaubhaftigkeitskriterien zu einem Gesamtscore aufsummiert. Anschließend wurde für jeden der drei Auswerter die Produkt-Moment-Korrelation zwischen den von ihm vergebenen Gesamtscores einerseits und seinen klinisch-intuitiven Diagnosen (auf der zehnstufigen Skala) andererseits berechnet. Bei Auswerter A betrug die Korrelation .705, bei Auswerter B .790 und bei Auswerter C .569 (jeweils $p < .01$). Diese Koeffizienten sprechen immerhin für moderate bis deutliche Zusammenhänge zwischen den vorgenommenen Inhaltsanalysen und den abschließenden klinisch-intuitiven Glaubhaftigkeitsbeurteilungen. Des weiteren hätte man bei einer adäquaten klinisch-intuitiven Beurteilung jedoch auch erwarten können, daß diesbezüglich zwischen den drei Auswertern eine hohe Urteilskonkordanz bestand. Die oben erläuterte Gruppe × Rater-Wechselwirkung auf die Höhe der klinisch-intuitiven Urteile spricht allerdings dagegen. Um die Auswerterübereinstimmung weiter zu analysieren, wurden post hoc die paarweisen Produkt-Moment-Korrelationen zwischen den Auswertern hinsichtlich der Urteile auf der zehnstufigen Glaubhaftigkeitsskala berechnet. Bei den Auswerterpaaren A – B und B – C betrugen die Koeffizienten immerhin .557 bzw. .421 (jeweils $p < .01$). Dagegen bestand zwischen den Urteilen der Auswerter A und C praktisch kein Zusammenhang ($r = .147$; nicht signifikant)[31]. An dieser Stelle sei erwähnt, daß die paarweisen Korrelationen zwischen den Auswertern hinsichtlich der Höhe der vergebenen inhaltsanalytischen Gesamtscores wesentlich höher waren (Auswerterpaar A – B: $r = .719$; Auswerterpaar A – C: $r = .737$; Auswerterpaar B – C: $r = .778$; jeweils $p < .01$), was erneut die insgesamt zufriedenstellende Interrater-Reliabilität bzw. Auswertungsobjektivität der *Kriterienorientierten Inhaltsanalyse* in der vorliegenden Studie unterstreicht.

Insgesamt deuten also auch die Resultate der statistischen Reanalysen darauf hin, daß die **klinisch-intuitive Urteilsbildung in der vorliegenden Studie nicht in einer adäquaten Weise erfolgte**. Ansonsten hätte zumindest die diesbezügliche Übereinstimmung zwischen den drei Auswertern höher ausfallen müssen. Oben wurde bereits bemängelt, daß die Auswerter ihre Diagnosen nicht begründen mußten, so daß die Frage letztlich unbeantwortet bleiben muß, welche Faktoren für die Urteilsdiskrepanzen zwischen den Auswertern ausschlaggebend waren. Zusammenfassend ist festzuhalten, daß die klinisch-intuitiven Trefferquoten der vorliegenden Untersuchung keine angemesse-

[31] Im Zusammenhang mit der Analyse der Auswertungsobjektivität der Glaubhaftigkeitskriterien wurde darauf hingewiesen, daß niedrige Produkt-Moment-Korrelationen auch durch die Verteilungseigenschaften der beteiligten Variablen bedingt sein können, was dann zu einer Unterschätzung des tatsächlichen Zusammenhangs führt (vgl. Abschnitt 5.2.1.3). Diese Begründung gilt jedoch nicht für die geringe Korrelation zwischen den Auswertern A und C, die beide bei der klinisch-intuitiven Beurteilung die gesamte Breite der zehnstufigen Skala nutzten. Der niedrige Koeffizient reflektiert somit wirklich eine mangelhafte Urteilskonkordanz.

nen Kennwerte für die diagnostische Treffsicherheit der inhaltsorientierten Glaubhaftigkeitsbeurteilung darstellen.

Neben den klinisch-intuitiven Diagnosen wurde die Treffsicherheit der inhaltsorientierten Glaubhaftigkeitsbeurteilung in der vorliegenden Studie auch **diskriminanzanalytisch** überprüft. In Abschnitt 2.1.6.2 wurde heftige Kritik an den bis dato publizierten Studien geübt, in denen die Treffsicherheit der inhaltsorientierten Beurteilung diskriminanzanalytisch bestimmt wurde. Es wurde in erster Linie der Post-hoc-Charakter der Trefferquoten bemängelt. So wurden die diskriminanzanalytisch berechneten Klassifikationsregeln bzw. die damit erzielten Hitraten in den meisten Studien nicht an unabhängigen Aussagestichproben kreuzvalidiert (Ausnahmen: Ruby & Brigham, 1998; Köhnken et al., 1995). Werden jedoch die gleichen Aussagen klassifiziert, anhand derer auch die Klassifikationsregeln berechnet wurden, so überschätzen die resultierenden Trefferquoten normalerweise die für die Population gültigen Hitraten. Darüber hinaus wurde kritisiert, daß die diskriminanzanalytisch bestimmte Gewichtungsstruktur der Kriterien nicht theoretisch fundiert ist und zudem zwischen einzelnen Studien stark variierte.

Die genannten **Schwachpunkte** wurden in der vorliegenden Arbeit **vermieden**. Um der Tatsache Rechnung zu tragen, daß keine verbindlichen, theoretisch begründeten Richtlinien für die diagnostische Gewichtung der einzelnen Glaubhaftigkeitskriterien existieren, wurde – in Anlehnung an die Vorgehensweise von Vrij et al. (2000) – eine diskriminanzanalytische Klassifikationsprozedur vorgenommen, in welcher der über alle 18 Kriterien aufsummierte **Gesamtscore als einziger Prädiktor** der Glaubhaftigkeit diente. Die Klassifikationsregeln wurden anhand einer Zufallsstichprobe von 51 Aussagen berechnet, welche aus dem Gesamtpool von 102 experimentellen Aussagen gezogen wurde. Die Treffer- bzw. Fehlerquoten wurden bestimmt, indem die Klassifikationsregeln auf die verbleibenden 51 Fälle angewandt wurden (**Kreuzvalidierung**). Auf diese Weise wurden die **glaubhaften Aussagen mit einer deutlich über der Zufallswahrscheinlichkeit von 50% liegenden Häufigkeit zutreffend klassifiziert** – 14 der 17 *Zeuginnen* der Validierungsstichprobe wurden als glaubhaft eingestuft. Dagegen lag die **Trefferquote bei den unglaubhaften Aussagen auf dem Zufallsniveau**. Sowohl die *Täterinnen* als auch die *falschen Zeuginnen* wurden jeweils in etwa genauso oft falsch wie richtig klassifiziert. Die vorgefundene Gesamttrefferquote betrug 60.8%; nach Korrektur der ungleichen Basisraten von glaubhaften und unglaubhaften Aussagen lag die Gesamttrefferquote bei 66.2%. Verglichen mit dem Durchschnitt der in Abschnitt 2.1.6.2 zitierten diskriminanzanalytischen Trefferquoten (78% Treffer bei den erlebnisbezogenen und 77% Treffer bei den erfundenen Aussagen) waren die Hitraten der vorliegenden Studie in bezug auf die erlebnisbezogenen Aussagen etwas höher

(82.4%), hinsichtlich der erfundenen Aussagen jedoch deutlich niedriger (50%; Mittelwert der *Täterinnen* und *falschen Zeuginnen*). Die diskriminanzanalytischen Vorgehensweisen der meisten anderen Studien sind nicht direkt mit dem Prozedere in der vorliegenden Arbeit vergleichbar und. In Relation zur vorliegenden Studie dürften die diskriminanzanalytischen Trefferquoten der meisten unter Abschnitt 2.1.6.2 angeführten Untersuchungen die praktische Differenzierungsfähigkeit der *Kriterienorientierten Inhaltsanalyse* überschätzen. In der einzigen Studie, in welcher – wie hier – der Gesamtscore als Prädiktor der Glaubhaftigkeit eingesetzt wurde (Vrij et al., 2000), führte man keine Kreuzvalidierung der Klassifikationsregeln bzw. Trefferquoten durch. Dagegen ergaben sich die Hitraten von Ruby und Brigham (1998) sowie Köhnken et al. (1995) zwar im Rahmen einer Kreuzvalidierung, allerdings dienten jeweils alle berücksichtigten Einzelkriterien als eigenständige Prädiktoren der Glaubhaftigkeit. In den restlichen Studien wurde weder der Gesamtscore als einziger Prädiktor eingesetzt, noch eine Kreuzvalidierung durchgeführt. Insofern ist die vergleichsweise niedrige (basisratenkorrigierte) Gesamttrefferquote der vorliegenden Studie nicht überraschend. Auffallend ist die ausgeprägte Fehlertendenz in Richtung falsch positiver Klassifikationen; die unglaubhaften Aussagen wurden im Rahmen der Diskriminanzanalyse annähernd dreimal häufiger fehlklassifiziert als die erlebnisbezogenen Aussagen. Dieses Resultat steht im Einklang mit der allgemeinen empirischen Befundlage zur Treffsicherheit der inhaltsorientierten Glaubhaftigkeitsbeurteilung. Dabei ist allerdings anzumerken, daß diese Fehlertendenz gerade auf diskriminanzanalytischem Weg bislang nicht konsistent nachgewiesen werden konnte.

Mehr aus Gründen der Vergleichbarkeit mit anderen empirischen Befunden wurde in der vorliegenden Arbeit auch noch eine diskriminanzanalytische Klassifikationsprozedur durchgeführt, in welcher **alle 18 Kriterien als eigenständige Prädiktoren** der Glaubhaftigkeit eingesetzt wurden, wobei ebenfalls eine Kreuzvalidierung stattfand. Erstaunlicherweise war die hierbei resultierende Gesamttrefferquote (64.7%) kaum höher als in der diskriminanzanalytischen Klassifikation mit dem Gesamtscore als einzigem Prädiktor; die basisratenkorrigierten Gesamthitraten waren sogar identisch (66.2%). Die Trefferquote in bezug auf die erlebnisbezogenen Aussagen war deutlich überzufällig – zwölf der 17 *Zeuginnen* in der Validierungsstichprobe wurden zutreffend als glaubhaft kategorisiert. Die gleiche überzufällige Hitrate ergab sich auch hinsichtlich der konfabulierten Aussagen der *falschen Zeuginnen*. Dagegen war bei den erfundenen Aussagen der *Täterinnen* das Verhältnis von Richtig- und Falschklassifikationen in etwa ausgeglichen. In den Studien von Köhnken et al. (1995) sowie Ruby und Brigham (1998), in denen hinsichtlich der diskriminanzanalytischen Klassifikation analog verfahren wurde (alle berücksichtigten Kriterien als eigenständige Prädiktoren, Kreuzvalidierung der Klassifikation), ergaben sich etwas höhere Gesamttrefferquoten (vgl. Abschnitt

2.1.6.2). Bei Köhnken et al. (1995) wurden die glaubhaften Aussagen deutlich häufiger zutreffend klassifiziert (88%) als hier, während die Trefferquote bei den erfundenen Schilderungen (62%) der diesbezüglichen Hitrate in der vorliegenden Studie (61.8%; Mittelwert der *Täterinnen* und *falschen Zeuginnen*) entsprach. Umgekehrt wurden bei Ruby und Brigham (1998) in etwa genauso viele erlebnisbezogene Aussagen (70%, Mittelwert der Aussagen von farbigen und weißen Erzählern, vgl. Tabelle 4) zutreffend klassifiziert wie in der vorliegenden Arbeit; dafür war dort der Anteil korrekt kategorisierter Falschaussagen (69%, ebenfalls Mittelwert beider ethnischer Gruppen) höher als hier. Faßt man die beiden falschaussagenden Gruppen der vorliegenden Untersuchung zusammen, so zeigte sich auch bei dieser diskriminanzanalytischen Klassifikation eine Fehlerneigung zugunsten falsch positiver Klassifikationen (38.2% falsch positive vs. 29.4% falsch negative). Dieser Befund steht zudem im Einklang mit dem Ergebnis von Köhnken et al. (1995).

Zusammenfassend sprechen die Resultate der diskriminanzanalytischen Klassifikation dafür, daß die *Kriterienorientierte Inhaltsanalyse* nur zur Identifizierung glaubhafter, nicht jedoch zur Entlarvung unglaubhafter Aussagen beitragen kann. In der wohl aussagekräftigsten Klassifikationsprozedur, bei welcher der inhaltsanalytische Gesamtscore als einziger Prädiktor diente, resultierte mit 82.4% eine diagnostische Sensitivität deutlich über dem Zufallsniveau. Dagegen entsprach die Treffsicherheit bei den unglaubhaften Aussagen sowohl der *Täterinnen* als auch der *falschen Zeuginnen* (diagnostische Spezifität) der Zufallswahrscheinlichkeit. Zwar könnte man mit Steller (1989, S. 149) argumentieren: „This bias towards falsely classifying fictitious statements as truthful rather than committing the reverse error, is consistent with the theoretical basis of the method [...] CBCA (*Kriterienorientierte Inhaltsanalyse*, Anmerkung des Verfassers) is a truth verifying rather than a lie detection method.“ Allerdings gilt es zu beachten, daß die inhaltsorientierte Glaubhaftigkeitsbeurteilung in der Praxis ganz überwiegend zur Begutachtung von Anschuldigungen (sexueller Delikte) eingesetzt wird. Vor diesem Hintergrund ist die hohe Rate falsch positiver Klassifikationen (erfundene Aussagen, die als glaubhaft eingestuft werden) nicht mit dem Rechtsgrundsatz des Schutzes Unschuldiger vereinbar.

Zu der hier vorgenommenen diskriminanzanalytischen Klassifikationsprozedur sind – trotz der oben erläuterten Verbesserungen gegenüber früheren Studien – noch einige **kritische Anmerkungen** zu machen. Zwar wurde in der vorliegenden Studie eine Kreuzvalidierung der Klassifikation vorgenommen. Die dabei verwendete **„Hold-out sample“-Methode** (vgl. Abschnitt 5.2.3.1) ist allerdings prinzipiell für große Stichproben konzipiert. Die Stichprobenumfänge in der vorliegenden Untersuchung (jeweils 51 Aussagen [17 erlebnisbezogene und 34 erfundene] in der Konstruktions- bzw. Klassifi-

kationsstichprobe) waren jedoch eher gering, so daß die resultierenden Treffer- und Fehlerraten aus statistischer Perspektive grundsätzlich anfechtbar sind. Als Alternative zur „Hold-out sample“-Methode im Falle kleiner Stichproben schlägt Bortz (1999, S. 604) die **„Leave-one-out“-Methode** vor. Dabei wird jeder der N Fälle der Stichprobe anhand der Zuordnungsregeln klassifiziert, die jeweils an den restlichen N – 1 Fällen (Konstruktionsstichprobe) berechnet wurden. Übertragen auf die vorliegende Stichprobe hätte dies bedeutet, daß man für jede der N = 102 Aussagen anhand der jeweils restlichen 101 Aussagen Klassifikationsregeln hätte berechnen müssen, d.h. die Durchführung von 102 diskriminanzanalytischen Klassifikationsprozeduren wäre erforderlich gewesen. Da kein Statistik-Programmpaket mit der Möglichkeit der Automatisierung dieses äußerst aufwendigen Vorgangs zur Verfügung stand, wurde von der Durchführung der „Leave-one-out“-Methode abgesehen. Die Untersuchung von Köhnken et al. (1995) deutet zudem darauf hin, daß die „Hold-out sample“-Methode und die „Leave-one-out“-Methode alles in allem zu äquivalenten Einschätzungen der diagnostischen Differenzierungsfähigkeit führen. Die Autoren führten bei einer Stichprobe, die mit N = 59 (28 erlebnisbezogene und 31 erfundene Aussagen) noch geringer als die der vorliegenden Studie war, beide Methoden der Kreuzvalidierung durch. Bei der „Hold-out sample“-Methode ergab sich eine Gesamttrefferquote von 75%, mit der „Leave-one-out“-Methode resultierte eine Gesamttrefferquote von 73%. Allerdings unterschieden sich die beiden Kreuzvalidierungsverfahren in den resultierenden Werten für die Sensitivität und Spezifität. Während mit der „Hold-out sample“-Methode 88% der erlebnisbezogenen und 62% der konfabulierten Aussagen zutreffend kategorisiert wurden, führte die „Leave-one-out“-Methode zu 75% valide positiven und zu 71% valide negativen Zuordnungen. Die nahezu deckungsgleichen Gesamttrefferquoten, die sich mit beiden Verfahren ergaben, sprechen aber zumindest dafür, daß die „Hold-out sample“-Methode auch bei kleineren Stichproben zu validen Schlußfolgerungen hinsichtlich der allgemeinen diagnostischen Differenzierungsfähigkeit der *Kriterienorientierten Inhaltsanalyse* führt.

Auch die diagnostische Treffsicherheit des *GAT* wurde mit Hilfe der „Hold-out sample“-Methode analysiert, wobei zudem die Konstruktions- und die Klassifikationsstichprobe identisch waren mit denjenigen, die auch bei der Analyse der inhaltsorientierten Glaubhaftigkeitsbeurteilung herangezogen wurden. Insofern eignen sich die mit der „Hold-out sample“-Methode jeweils erzielten Gesamttrefferquoten sehr gut für einen direkten Vergleich der diagnostischen Leistungsfähigkeit der inhaltsorientierten Glaubhaftigkeitsbeurteilung einerseits und des *GAT* andererseits. Hierauf wird in Abschnitt 6.4 noch näher eingegangen.

Abschließend sei zum wiederholten Mal darauf hingewiesen, daß **für eine adäquate inhaltsorientierte Glaubhaftigkeitsbeurteilung neben den festgestellten Ausprägungsgraden der inhaltlichen Glaubhaftigkeitskriterien auch noch weitere Informationsquellen zur Verfügung stehen müssen** (in erster Linie die Ergebnisse gezielter Persönlichkeits- und Motivanalysen, ferner die genaue Rekonstruktion der Aussagegenese, gegebenenfalls eine Konstanzprüfung sowie ein Vergleich der in Frage stehenden Aussage mit gesichert erlebnisbezogenen bzw. erfundenen Aussagen derselben Person). Diese Daten sind dann im Rahmen der klinisch-intuitiven Urteilsbildung zu einem Gesamturteil hinsichtlich der Glaubhaftigkeit der Aussage zu integrieren. Insofern als die beschriebenen diskriminanzanalytischen Trefferquoten nur auf den festgestellten Ausprägungsgraden der Glaubhaftigkeitskriterien basieren, repräsentieren sie nicht die Treffsicherheit der inhaltsorientierten Glaubhaftigkeitsbeurteilung als diagnostischer Gesamtprozedur, sondern stellen eher einen unteren Schätzwert dar. Sofern die anderen genannten Informationsquellen für die diagnostische Urteilsbildung relevant sind, sollte sich erst bei deren Berücksichtigung die wahre Treffsicherheit der inhaltsorientierten Glaubhaftigkeitsbeurteilung zeigen (ein Punkt, der bislang jedoch noch nicht durch kontrollierte Studien angegangen wurde). Gleichwohl gilt, daß die Feststellung der inhaltlichen Aussagequalität anhand der Glaubhaftigkeitskriterien (*Kriterienorientierte Inhaltsanalyse*) das *zentrale* Element der diagnostischen Gesamtprozedur inhaltsorientierte Glaubhaftigkeitsbeurteilung ist. Angesichts der in numerischer Hinsicht eher mäßigen (wenngleich signifikanten) Differenzierungsfähigkeit der *Kriterienorientierten Inhaltsanalyse* in der vorliegenden Studie ist es unwahrscheinlich, daß sich bei einer Berücksichtigung zusätzlicher diagnostisch relevanter Daten wesentlich höhere Trefferquoten ergeben hätten als im Rahmen der hier vorgenommenen diskriminanzanalytischen Klassifikation.

6.2 Zum diagnostischen Potential des *Guilty Actions Tests* in der vorliegenden Studie

Auch in bezug auf den *GAT* soll der Diskussion der erzielten Trefferquoten eine genaue Erörterung der Befunde zur theoretischen Grundannahme des Verfahrens vorausgehen. Beginnen soll die Diskussion jedoch mit dem Vergleich der beiden SCR-Quantifizierungsmethoden, die in der vorliegenden Studie zur Anwendung kamen.

6.2.1 Vergleich der beiden SCR-Quantifizierungsmethoden

Als abhängiges physiologisches Maß im *GAT* wurden die Amplituden der phasischen Hautleitfähigkeitsreaktionen (SCRs) erfaßt, welche sich unmittelbar nach Darbietungsbeginn der jeweiligen Testitems manifestierten. Bei der Quantifizierung der SCR-Amplituden kamen zwei unterschiedliche Methoden zur Anwendung, welche sich hinsichtlich des zugrunde gelegten **Latenzzeitkriteriums** unterschieden. Gemäß **Quantifizierungsmethode A**, welche insbesondere in der experimentellen Grundlagenforschung gebräuchlich ist, wurden nur solche Hautleitfähigkeitsanstiege als reizbezogene SCRs gewertet, deren Anstiegspunkt mit einer Latenz von mindestens einer und maximal drei Sekunden nach Reizbeginn auftrat. Demgegenüber wurde bei **Quantifizierungsmethode B** ein breiteres Latenzzeitkriterium veranschlagt. Hier wurden jeweils die größten Hautleitfähigkeitsanstiege, die in dem Zeitintervall von einer bis zehn Sekunden nach Reizbeginn sowohl ihren Anstiegs- als auch ihren Gipfelpunkt hatten, als reizbezogene SCRs gewertet. Methode B entspricht eher den Gepflogenheiten in der forensischen Praxis; sie ist zudem mit der Auswertungsmethode vergleichbar, die die Bradley-Gruppe in den drei experimentellen Studien zum *GAT* anwandte.

In der vorliegenden Studie ergaben sich unter Verwendung der beiden unterschiedlichen SCR-Quantifizierungsmethoden weitestgehend **äquivalente Resultate**. So manifestierten sich die gleichen signifikanten Gruppenunterschiede in bezug auf die numerischen Scores. Auch die direkte statistische Analyse der SCR-Amplitudenwerte erbrachte für beide SCR-Quantifizierungsmethoden nahezu die gleichen statistisch bedeutsamen Haupt- und Wechselwirkungseffekte. Ferner zeigten sich auch nach der Herauspartialisierung der Kontrollvariablen im Rahmen kovarianzanalytischer Berechnungen für beide SCR-Quantifizierungsmethoden die gleichen signifikanten Effekte bezüglich der numerischen Scores und der SCR-Amplitudenwerte. Bei sämtlichen signifikanten Effekten unterschieden sich die jeweiligen statistischen Effektstärken kaum zwischen den beiden Quantifizierungsmethoden. Auch was die resultierenden Treffer- bzw. Fehlerquoten bei der Klassifizierung der Pbn angeht, zeigten sich keine wesentlichen Unterschiede zwischen den SCR-Quantifizierungsmethoden A und B.

In Anbetracht der weitestgehenden Deckungsgleichheit der mit beiden Auswertungsmethoden erzielten Resultate konzentriert sich die nachfolgende Diskussion nur noch auf die Ergebnisse gemäß SCR-Quantifizierungsmethode A, da letztere eher wissenschaftlichen Standards entspricht. Sofern die Ergebnisse gemäß Methode B doch einmal abwichen, wird dies an den entsprechenden Stellen erwähnt.

6.2.2 Gültigkeit der theoretischen Grundannahme in der vorliegenden Untersuchung

Gemäß der Grundannahme des *GAT* war zu erwarten, daß die drei experimentellen Gruppen sich in ihren physiologischen Reaktionsstärkemustern hinsichtlich der relevanten und irrelevanten Testitems unterschieden. Die *falschen Zeuginnen* (*Unschuldige ohne Tatwissen*) sollten auf die relevanten Items genauso stark reagieren wie auf die irrelevanten Items, da sie die zutreffenden ebenso wie die irrelevanten Alternativen weder wiedererkannten noch wahrheitswidrig verneinten. Demgegenüber sollten die *Zeuginnen* (*Unschuldige mit Tatwissen*) die relevanten Items als Tatdetails identifizieren und dementsprechend hier größere SCR-Amplituden zeigen als bei den irrelevanten Alternativen. Für die *Täterinnen* (*Schuldige*) wurde erwartet, daß der Reaktionsstärkeunterschied zwischen den relevanten und irrelevanten Items noch deutlicher ausgeprägt sein würde als bei den *Unschuldigen mit Tatwissen*. Im Gegensatz zur letztgenannten Gruppe erkannten die *Schuldigen* die zutreffenden Alternativen nämlich nicht nur wieder, sondern verneinten sie auch noch wahrheitswidrig.

Die postulierten Gruppenunterschiede wurden zunächst anhand der **numerischen Scores** überprüft, die nach dem Punktesystem von Lykken (1959) bestimmt wurden. Diesbezüglich zeigten sich aber nur signifikante erwartungskonforme Unterschiede zwischen den *Unschuldigen ohne Tatwissen* einerseits und den beiden übrigen experimentellen Gruppen andererseits. Dagegen **differenzierten** die numerischen Scores **nicht signifikant zwischen den *Schuldigen* und den *Unschuldigen mit Tatwissen***. Die fehlende Differenzierung zwischen *Schuldigen* und *Unschuldigen mit Tatwissen* könnte jedoch auch auf die **Inadäquanz des numerischen Auswertungsverfahrens** zurückzuführen sein. Wie in Abschnitt 2.2.3.3 eingehend erläutert wurde, ist dieses – für den konventionellen *TWT* konzipierte – Auswertungsverfahren grundsätzlich nicht dazu geeignet, das graduelle Ausmaß eines etwaigen Reaktionsstärkeunterschieds zwischen relevanten und irrelevanten Items zu quantifizieren, sondern spiegelt nur die Rangreihenposition wieder, welche die Reaktionsstärke auf das relevante Item innerhalb der Reaktionsstärken der gesamten Itemsequenz einer Frage einnimmt. Gleichwohl steht die mangelnde Differenzierung der vorliegenden Studie im **Kontrast zu den diesbezüglichen Befunden des Bradley-Arbeitskreises**. Bei Bradley und Warfield (1984) sowie Bradley und Rettinger (1992) differenzierten die numerischen Scores jeweils signifikant zwischen den *Schuldigen* und den *Unschuldigen mit Tatwissen*.

Die Problematik des numerischen Auswertungsverfahrens wurde umgangen, indem die **SCR-Amplitudenwerte** direkt statistisch analysiert wurden (wobei zur Annäherung an die Normalverteilung zunächst eine logarithmische Transformation erfolgte). Auf diese

Weise konnte die Grundannahme des *GAT* also unmittelbar überprüft werden. Es zeigte sich zwar eine signifikante Wechselwirkung der experimentellen Gruppenzugehörigkeit und des Itemtyps auf die SCR-Amplitude, die darauf zurückzuführen war, daß nur die SCR-Amplituden bei den relevanten Items, nicht jedoch diejenigen bei den irrelevanten Items zwischen den experimentellen Gruppen differenzierten. Anschlußtests ergaben jedoch, daß hinsichtlich der Reaktionsstärke auf die relevanten Items nur signifikante erwartungsgemäße Gruppenunterschiede zwischen den *Schuldigen* und *Unschuldigen ohne Tatwissen* bzw. zwischen den *Unschuldigen mit Tatwissen* und den *Unschuldigen ohne Tatwissen* bestanden. Dagegen war die **Differenzierung zwischen den *Schuldigen* und den *Unschuldigen mit Tatwissen* nicht statistisch bedeutsam**. Das heißt, daß die **theoretische Grundannahme des *GAT* in ihrem zentralen Aspekt nicht bestätigt** werden konnte. Die wahrheitswidrige Verneinung der relevanten Items durch die *Schuldigen* hatte offensichtlich keinen aktivierungssteigernden Effekt, der die SCR-Amplituden der *Schuldigen* bei den relevanten Items größer werden ließ als die diesbezüglichen SCR-Amplituden der *Unschuldigen mit Tatwissen*. Die psychophysiologische Reaktionsweise in dem hier durchgeführten *GAT* unterschied sich somit nicht von derjenigen im herkömmlichen *TWT*; der *GAT* differenzierte nur zwischen Personen mit und solchen ohne Tatwissen, innerhalb der Personen mit Tatwissen trennte er jedoch nicht zwischen Schuldigen und Unschuldigen.

Für die fehlende Differenzierung zwischen *Schuldigen* und *Unschuldigen mit Tatwissen* anhand des *GAT* in der vorliegenden Studie kommen neben der Invalidität der theoretischen Basisannahme auch noch einige **alternative Erklärungen** in Betracht. An erster Stelle muß das in der vorliegenden Untersuchung verwendete **Frageformat** problematisiert werden. Der postulierte Vorteil des *GAT* gegenüber dem herkömmlichen *TWT*, nämlich die Differenzierungsfähigkeit in bezug auf *Schuldige* vs. *Unschuldige mit Tatwissen*, wird dadurch ermöglicht, daß man die Fragen anders formuliert. Während die Fragen im *TWT* nur auf die Kenntnis tatbezogener Detailinformationen abstellen („The murder took place in a(n) ___? (a) service station (b) store (c) apartment ...“), werden im *GAT* die zur Auswahl stehenden Alternativen in Fragen eingebettet, welche die Täterschaft selbst thematisieren („You murdered the victim in a(n) ___? (a) service station (b) store (c) apartment ...“; Bradley et al., 1996, S. 156). Auf diese Weise wird im *GAT* gewährleistet, daß Täter bei der Verneinung der zutreffenden Alternative lügen, wohingegen *Unschuldige mit Tatwissen* die zutreffende Alternative zwar ebenfalls wiedererkennen, jedoch aufrichtig sind, indem sie darauf mit nein antworten. Die wahrheitswidrige Verneinung der relevanten Items soll mit einer Steigerung der autonomen Erregung einhergehen, anhand derer letztlich Täter von *Unschuldigen mit Tatwissen* zu unterscheiden sind. In der vorliegenden Studie war allerdings das Frageformat gegenüber dem *GAT*-Frageformat in den Untersuchungen der Bradley-Arbeitsgruppe abgewandelt.

Zwar richteten sich auch hier die Fragen sowohl auf das Vorhandensein von Tatkenntnissen als auch auf die Täterschaft, die Frageformulierung war jedoch wesentlich komplexer als bei Bradley und Kollegen. So bestanden die Fragen jeweils aus einem Haupt- und einem Nebensatz. Das Tatdetail wurde im Hauptsatz, die Täterschaft hingegen im Nebensatz der Frage angesprochen (z.B.: „Welche Farbe hatte das Fahrrad in dem Raum, in dem Sie die 100,- DM gestohlen haben? War es ... (a) rot? (b) weiß? (c) blau? ...“). Dies hat möglicherweise dazu geführt, daß der Aspekt der Tatbegehung für die Pbn zu wenig Salienz hatte bzw. daß der Aspekt des Tatwissens zu sehr ins Zentrum der Aufmerksamkeit gerückt wurde. In diesem Zusammenhang ist auch die Art und Weise zu bemängeln, in der die Pbn auf den *GAT* eingestimmt wurden. So hieß es zu Beginn der entsprechenden schriftlichen Instruktion: „... Bei dem Lügendetektortest werden Ihnen 10 Fragen gestellt. Diese Fragen beziehen sich auf die folgenden 10 Details des Tatorts bzw. Tathergangs: die Farbe des Fahrrads, das sich am Tatort befand; die Art des Getränks, das sich in dem Getränkekasten am Tatort befand ...“ (s. Anhang A, Instruktion G4). Diese Instruktion hat möglicherweise die Aufmerksamkeit der Pbn von vorneherein zu sehr auf den Aspekt des Tatwissens fokussiert. Das heißt, daß die Pbn die Verneinung der Testitems möglicherweise primär auf den Aspekt des eigenen Tatwissens und weniger auf den Aspekt der eigenen Täterschaft bezogen. Folglich bedeutete es für die *Unschuldigen mit Tatwissen* – ebenso wie für die *Schuldigen* – zu lügen, wenn sie die relevanten Items mit nein beantworteten. Dies könnte dann zu einer Angleichung der Reaktionsstärken von *Schuldigen* und *Unschuldigen mit Tatwissen* bei den relevanten Items geführt haben.

Die fehlende Differenzierung zwischen *Schuldigen* und *Unschuldigen mit Tatwissen* steht möglicherweise aber auch damit im Zusammenhang, daß die **elektrodermale Reagibilität** der Pbn in der vorliegenden Untersuchung insgesamt sehr gering war. Reanalysen ergaben, daß die Pbn durchschnittlich nur auf 17.3 der insgesamt 50 Testitems (34.6%; Pufferitems nicht berücksichtigt) überhaupt mit einer SCR reagierten. Im Durchschnitt lösten lediglich 12.1 der insgesamt 40 irrelevanten Items (30.3%) sowie 5.2 der insgesamt 10 relevanten Items (52%) eine SCR aus. Die Personen mit Tatwissen (*Schuldige* und *Unschuldige mit Tatwissen*) reagierten nur auf 64.1% bzw. 57.1% der Darbietungen relevanter Items mit einer SCR; bei den *Unschuldigen ohne Tatwissen* betrug die entsprechende Quote 34.7%.[32] Die niedrige Rate ausgelöster SCRs sagt natürlich noch nichts darüber aus, wie stark die ausgelösten SCRs waren. Hier wäre es insbesondere interessant, die SCR-Magnituden der vorliegenden Studie mit den entsprechenden Daten der Bradley-Gruppe zu vergleichen. Leider finden sich jedoch in

[32] Diese Angaben beziehen sich auf SCR-Quantifizierungsmethode A. Unter Anwendung von Methode B ergaben sich naturgemäß etwas höhere Reaktionsquoten (44.4% bei den Items insgesamt [ohne Pufferitems]; 40.9% bei den irrelevanten Items; 58.2% bei den relevanten Items [*Schuldige*: 68.8%; *Unschuldige mit Tatwissen*: 62.1%; *Unschuldige ohne Tatwissen*: 43.8%]).

deren Publikationen keine Angaben zur absoluten Höhe der SCR-Amplituden, sondern es werden nur die numerischen Scores mitgeteilt. Zur Relativierung der Reaktionsstärken in der vorliegenden Studie seien daher Daten aus einer eigenen Untersuchung (Gödert, Rill & Vossel, 2001) im Rahmen des sog. „Differentiation of deception"-Paradigmas herangezogen. In diesem experimentellen Paradigma geht es grundsätzlich um den Vergleich autonomer Erregung in zwei intraindividuell variierten Versuchsbedingungen, die sich lediglich darin unterscheiden, daß bestimmte Fragen in der einen Bedingung wahrheitsgemäß und in der anderen Bedingung wahrheitswidrig zu beantworten sind. Bei Gödert et al. (2001) wurde gezielt darauf geachtet, den emotionalen und motivationalen Anregungsgehalt der Untersuchungssituation sowie die mentale Beanspruchung der Pbn so gering wie möglich zu halten. So handelte es sich beim Reizmaterial um einfache Fragen zum Allgemeinwissen mit möglichst geringem autobiographischem Bezug. Die Fragen waren geschlossen, brauchten also nur mit ja bzw. nein beantwortet zu werden. Analog zur vorliegenden Studie, befanden sich die Pbn während der Befragung allein in einer abgeschirmten Kabine (kein Kontakt zum Versuchsleiter). Die Stimuli wurden allerdings nur optisch (Monitor) dargeboten. Sowohl bei Gödert et al. (2001) als auch in der vorliegenden Studie befanden sich die Pbn also in einer Befragungssituation. Der Anregungsgehalt der Befragungssituation bei Gödert et al. (2001) ist jedoch als weitaus niedriger einzustufen als der in der vorliegenden Studie. In diesem Zusammenhang ist besonders hervorzuheben, daß es im „Differentiation of deception"-Paradigma nicht um die Aufklärung von Täterschaft geht, so daß es für die Pbn dort grundsätzlich nichts zu gewinnen oder zu verlieren gibt. Insofern wäre zu erwarten gewesen, daß die elektrodermalen Reaktionen in der vorliegenden Studie alles in allem weitaus stärker ausfielen als bei Gödert et al. (2001). Dies war jedoch nicht der Fall. Bei Gödert et al. (2001) betrug die nach SCR-Quantifizierungsmethode A bestimmte SCR-Amplitude beim Stimulus-Onset im Durchschnitt aller Versuchsbedingungen 0.046 log µS. Im Vergleich dazu war die über alle Gruppen und Itemtypen gemittelte SCR-Magnitude der vorliegenden Untersuchung mit 0.055 log µS (SCR-Quantifizierungsmethode A) nur unwesentlich höher. Andererseits ist jedoch auch zu erwähnen, daß die in der vorliegenden Studie gemessenen SCRs der Pbn mit Tatwissen bei den relevanten Items (*Schuldige*: 0.133 log µS; *Unschuldige mit Tatwissen*: 0.108 log µS) sich deutlicher von den SCR-Amplituden bei Gödert et al. (2001) abhoben. Dort betrug etwa die SCR- Magnitude bei den wahrheitswidrig beantworteten Items lediglich 0.055 log µS. Die vergleichsweise geringe elektrodermale Reagibilität in der vorliegenden Studie wird jedoch auch noch in einem anderen Aspekt offensichtlich. So wurden in der vorliegenden Arbeit, ebenso wie bei Gödert et al. (2001), neben den SCRs beim Stimulus-Onset auch noch die SCRs quantifiziert, die unmittelbar nach der jeweiligen Itemausblendung auftraten, wobei zu berücksichtigen ist, daß in beiden Untersuchungen das Reizende als imperativer Stimulus für die Antwortgabe fungierte. Auch die SCRs nach Reizende

wurden jeweils unter Zugrundelegung des konservativen Latenzzeitkriteriums (1 bis 3 Sekunden, vgl. SCR-Quantifizierungsmethode A) ausgewertet. In den Versuchsbedingungen von Gödert et al. (2001), in denen – wie in der vorliegenden Arbeit – verbal geantwortet wurde (außerdem gab es noch Antwort per Tastendruck), betrug die durchschnittliche SCR-Amplitude 0.157 log µS. Dagegen belief sich die über alle Gruppen und Itemtypen gemittelte SCR-Amplitude in der vorliegenden Untersuchung nur auf 0.059 log µS; die Magnituden der *Schuldigen* und der *Unschuldigen mit Tatwissen* bei den relevanten Items betrugen lediglich 0.075 bzw. 0.079 log µS. (Auf die Befunde zu den SCRs nach Itemausblendung wird in der vorliegenden Arbeit nicht weiter eingegangen, da die Berücksichtigung dieser Reaktionen im Rahmen der psychophysiologischen Glaubhaftigkeitsbeurteilung unüblich ist. Es sei aber zumindest erwähnt, daß die Differenzierung der drei experimentellen Gruppen anhand dieser SCRs noch schwächer war als anhand der SCRs, die nach den Quantifizierungsmethoden A bzw. B bestimmt wurden.) Die Gegenüberstellung der vorliegenden Untersuchungsbefunde mit denjenigen von Gödert et al. (2001) spricht also dafür, daß die elektrodermale Reagibilität in der vorliegenden Studie recht gering war. Beim Vergleich der beiden Studien gilt es allerdings zu beachten, daß die Stichprobe bei Gödert et al. (2001) männlich, hier dagegen weiblich war. Es ist nicht abzuschätzen, inwiefern die Unterschiede in den elektrodermalen Reaktionsstärken auch geschlechtsbedingt sind.

Über die **Gründe für die geringe elektrodermale Reagibilität** der vorliegenden Stichprobe kann nur spekuliert werden. Als Erklärung bietet sich in erster Linie die Art und Weise der Befragung an. Die Präsentation der Fragen und Items erfolgte in vollständig automatisierter Form. Zudem befanden sich die Pbn während der Befragung allein in einer abgeschirmten Kabine, d.h. es bestand keinerlei Kontakt zur Versuchsleiterin. Auch waren die Interstimulus-Intervalle recht lang (20 bis 22 Sekunden von der Ausblendung eines Items bis zur Einblendung des nächsten); und die Befragung wurde noch zusätzlich dadurch in die Länge gezogen, daß den eigentlichen *GAT*-Fragen jeweils ein einleitender Satz vorangestellt wurde. Es ist denkbar, daß der Anregungsgehalt der Befragungssituation unter den genannten Faktoren litt.

Gemäß dieser Argumentation hätte man allerdings erwarten können, daß die SCRs zu Beginn der Befragung noch relativ stark waren und sich erst im Testverlauf zunehmend abschwächten. Zur Überprüfung dieser Hypothese wurde post hoc eine ergänzende dreifaktorielle ANOVA der SCR-Daten gerechnet, in welcher neben dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Itemtyp (Puffer, relevant, irrelevant) auch noch die Fragenposition im *GAT* (1 bis 10) als zusätzlicher Wiederholungsfaktor berücksichtigt wurde. Neben der Itemtyp × Gruppe-Wechselwirkung und dem Itemtyp-Haupteffekt, die oben bereits erläutert wurden, ergaben sich ein signi-

fikanter Haupteffekt der Fragenposition, $F(9,891) = 10.020$, $\varepsilon = .705$, $p < .01$, eine signifikante Fragenposition × Itemtyp-Interaktion, $F(18,1782) = 2.975$, $\varepsilon = .580$, $p < .01$, sowie eine signifikante dreifache Wechselwirkung der Faktoren Fragenposition, Itemtyp und Gruppe, $F(36,1782) = 1.593$, $\varepsilon = .580$, $p < .05$ (s. genauer Tabelle F.118 im Anhang F; diese Angaben beziehen sich auf SCR-Quantifizierungsmethode A; für Quantifizierungsmethode B ergaben sich äquivalente Resultate, s. Tabelle F.119 im Anhang F). Der Haupteffekt der Fragenposition ist in Abbildung F.2 im Anhang F illustriert (Quantifizierungsmethode A; bezüglich Methode B s. Abbildung F.3). Wie der Graphik zu entnehmen ist, kam es nur im Verlauf der ersten drei *GAT*-Fragen zu einer deutlichen Abschwächung der SCRs; von der vierten bis zur zehnten Frage blieb die Reaktionsstärke dann weitgehend konstant. Dies galt für alle drei experimentellen Gruppen gleichermaßen, da die Fragenposition × Gruppe-Interaktion nicht statistisch bedeutsam war (s. Tabelle F.118 bzw. F.119). Die signifikante Fragenposition × Itemtyp-Wechselwirkung, die in Abbildung F.4 (für Quantifizierungsmethode B s. Abbildung F.5) im Anhang F veranschaulicht ist, beruhte in erster Linie darauf, daß bei den Pufferitems von der ersten zur zweiten *GAT*-Frage eine deutlich stärkere Reduktion der SCR-Amplituden erfolgte als bei den beiden anderen Itemtypen. Dagegen waren die SCR-Amplitudenverläufe der relevanten und der irrelevanten Items über die zehn *GAT*-Fragen nahezu parallel, wobei sich allerdings eine leichte Konvergenz der Reaktionsstärken im Testverlauf andeutete. Zur genaueren Aufklärung der dreifachen Interaktion zwischen den Faktoren Fragenposition, Itemtyp und Gruppenzugehörigkeit sei auf die Ergebnisse der separaten zweifaktoriellen ANOVAs verwiesen, die jeweils für die einzelnen *GAT*-Fragen gerechnet wurden (s. Abschnitt 5.3.1.1.2 bzw. 5.3.2.1.2). Wie sich dort herausstellte, war die Gruppe × Itemtyp-Interaktion (bei den relevanten Items größere SCR-Magnituden der *Schuldigen* und *Unschuldigen mit Tatwissen* im Vergleich zu den *Unschuldigen ohne Tatwissen*, jedoch keine Gruppenunterschiede bei den irrelevanten Items) nur bei den ersten sieben *GAT*-Fragen (SCR-Quantifizierungsmethode A) bzw. bei den ersten sechs *GAT*-Fragen (Quantifizierungsmethode B) statistisch bedeutsam. Das heißt, **im Testverlauf nahm die differentielle Reaktionsweise der Pbn mit Tatwissen in bezug auf die relevanten vs. irrelevanten *GAT*-Items ab**. Um diesen Sachverhalt zu veranschaulichen, sind in Abbildung F.6 (für Quantifizierungsmethode B s. Abbildung F.7) im Anhang F die Verläufe der SCR-Magnituden über die zehn *GAT*-Fragen in Abhängigkeit von der experimentellen Gruppenzugehörigkeit und dem Itemtyp abgetragen. Aus der Abbildung geht hervor, daß die SCR-Amplituden aller drei Gruppen bei den irrelevanten und Pufferitems über die zehn *GAT*-Fragen weitgehend konstant blieben. Gleiches gilt für die SCR-Amplituden der *Unschuldigen ohne Tatwissen* (*falsche Zeuginnen*) bei den relevanten Items. Dagegen nahmen die Reaktionsstärken der *Unschuldigen mit Tatwissen* (*Zeuginnen*) bei den relevanten Items und noch deutlicher die der *Schuldigen* (*Täterinnen*) bei den relevanten Items im Testverlauf ab

bzw. glichen sich zunehmend an deren Reaktionsstärken bei den irrelevanten und Pufferitems an.

Es sei darauf hingewiesen, daß die vorgefundenen Verlaufsmuster der SCR-Magnituden **nicht mit der „Dichotomisierungstheorie“** (z.B. Ben-Shakhar, 1977) **vereinbar** sind. Diese Theorie soll die differentiellen Reaktionsstärken von Personen mit Tatwissen bei den relevanten vs. irrelevanten Items des herkömmlichen *TWT* erklären. Es wird postuliert, daß Personen mit Tatwissen eine dichotome Kategorisierung der Testitems in tatbezogene vs. nicht tatbezogene Reize (relevante vs. irrelevante Items) vornehmen. Habituation der durch die Testitems ausgelösten Orientierungsreaktionen soll nur innerhalb der beiden Reizkategorien stattfinden, jedoch nicht von einer Reizkategorie auf die andere generalisieren. Aus der größeren Anzahl irrelevanter im Vergleich zu relevanten Items ergibt sich, daß die Habituation innerhalb der Kategorie „irrelevante Items” deutlicher ausfallen sollte als innerhalb der Kategorie „relevante Items“. Die Vorhersagen der Dichotomisierungstheorie müßten prinzipiell auch für den *GAT* gelten. Wie jedoch Abbildung F.6 (bzw. Abbildung F.7) zu entnehmen ist, zeigte sich bei den SCRs auf die irrelevanten Items überhaupt keine Habituation, und zwar weder bei den Personen mit noch bei den Personen ohne Tatwissen. Somit konnte die im Testverlauf beobachtete Reduktion der Reaktionsstärken der Pbn mit Tatwissen auf die relevanten Items logischerweise nicht schwächer ausgeprägt sein als die vermeintliche Habituation der SCRs auf die irrelevanten Items. Zudem erstreckte sich die Senkung der Reaktionsstärken auf die relevanten Items nur bei den *Schuldigen* nahezu über den gesamten Testverlauf, während dieser Prozeß bei den *Unschuldigen mit Tatwissen* schon nach der dritten *GAT*-Frage abgeschlossen war.

Die Ergebnisse der nachträglichen dreifaktoriellen ANOVA sprechen also nicht zwingend dafür, daß die insgesamt geringe elektrodermale Reagibilität in der vorliegenden Studie auf den niedrigen Anregungsgehalt der Befragungssituation bzw. -prozedur zurückzuführen war. Wäre dies so gewesen, hätte im Verlauf des *GAT* eine deutlichere Absenkung der SCR-Amplituden (Habituation) erfolgen müssen, und zwar insbesondere bei den irrelevanten Items (alle 3 experimentellen Gruppen) und bei den relevanten Items, die den Pbn ohne Tatwissen präsentiert wurden. Allerdings läßt sich diesem Argument entgegenhalten, daß eine deutlichere Abschwächung der SCRs im Verlauf des *GAT* möglicherweise durch einen Bodeneffekt verhindert wurde. So gingen dem *GAT* ja eine sechsminütige Ruhemessung, ein dreieinhalbminütiger Zahlentest („Stimulationstest“), ein ca. dreieinhalbminütiger *GAT*-Probedurchgang und nicht zuletzt ein etwa fünfzehnminütiges Interview zur Erhebung der Zeugenaussage voraus (vgl. Abschnitt 4.3.3). Insofern ist nicht auszuschließen, daß zu dem Zeitpunkt, als die Pbn den *GAT* absolvierten, die Gewöhnung an die Untersuchungssituation und damit einhergehend

die Abschwächung der SCRs bereits sehr weit fortgeschritten und somit kaum noch fortsetzungsfähig war. In diesem Zusammenhang sollte auch darauf hingewiesen werden, daß in den Studien der Bradley-Gruppe dem *GAT* keine entsprechenden Untersuchungselemente vorausgingen.

In den Studien der Bradley-Gruppe bestand – ebenso wie in der vorliegenden Arbeit – der *GAT* jeweils aus 10 Fragen, so daß die numerischen Scores jeweils von null bis 20 variieren konnten. Insofern kann man zumindest die numerischen Scores der Bradley-Gruppe unmittelbar mit denjenigen der vorliegenden Studie vergleichen. Im Durchschnitt der Studien des Bradley-Arbeitskreises (Bradley & Warfield, 1984; Bradley & Rettinger, 1992; Bradley et al., 1996) betrugen die numerischen Scores 13.5 für die *Schuldigen*, 8.7 für die *Unschuldigen mit Tatwissen*[33] und 5.4 für die *Unschuldigen ohne Tatwissen*. In Relation dazu war der numerische Score der *Schuldigen* in der vorliegenden Studie (10.2 gemäß SCR-Quantifizierungsmethode A; 9.9 gemäß Methode B) deutlich niedriger, der Score der *Unschuldigen mit Tatwissen* (9.4 [Methode A]; 9.3 [Methode B]) hingegen geringfügig höher. Im Sinne der obigen Argumentation könnte man spekulieren, daß die alles in allem geringe elektrodermale Reagibilität der vorliegenden Stichprobe insgesamt dazu führte, daß die Reaktionsstärkeunterschiede zwischen relevanten vs. irrelevanten Items bei Pbn mit Tatwissen reduziert wurden. Bei den *Schuldigen* führte dies möglicherweise zu einer relativen Absenkung der numerischen Scores gegenüber den früheren Studien. Bei den *Unschuldigen mit Tatwissen* wurde diese relative Absenkung der numerischen Scores möglicherweise dadurch verhindert, daß diese Pbn im Gegensatz zu den früheren Studien die relevanten Items nicht nur wiedererkannten, sondern deren Verneinung vermeintlich als Lüge empfanden und dementsprechend stärker darauf reagierten. Der vergleichsweise niedrige numerische Score der *Unschuldigen ohne Tatwissen* in der vorliegenden Studie (3.7 [SCR-Quantifizierungsmethode A]; 3.9 [Methode B]) dürfte insbesondere darauf zurückzuführen sein, daß hier pro Frage ein irrelevantes Item mehr dargeboten wurde als in den Studien der Bradley-Gruppe (4 vs. 3 irrelevante Items). Folglich war die Zufallswahrscheinlichkeit dafür, daß bei den relevanten Items stärkere SCRs auftraten als bei den irrelevanten Items, in der vorliegenden Studie geringer als in den Untersuchungen von Bradley und Kollegen.

Es kann **ausgeschlossen** werden, **daß die fehlende Differenzierung zwischen *Schuldigen* und *Unschuldigen mit Tatwissen* darauf basierte, daß die letztgenannte Gruppe sich an mehr Tatdetails hätte erinnern können**. Vielmehr ergaben die nach dem *GAT* durchgeführten Gedächtnistests, daß sowohl das freie Erinnern als auch das

[33] Die Gruppe der Unschuldigen, die die relevanten Details in einem neutralen Kontext erlernten und sich des Tatbezugs dieser Informationen nicht bewußt waren („Innocent Associations Group“, Bradley & Warfield, 1984, S. 685f.), ist hier nicht berücksichtigt.

Wiedererkennen der kritischen Tatdetails bei den *Schuldigen* ebenso wie bei den *Unschuldigen mit Tatwissen* jeweils nahezu maximal waren. Dennoch könnte man spekulieren, daß die kritischen Tatdetails im Gedächtnis der *Zeuginnen* (*Unschuldige mit Tatwissen*) umfassender und komplexer abgespeichert waren als im Gedächtnis der *Schuldigen* (*Täterinnen*) und daß dies zu einer Nivellierung der intendierten Reaktionsstärkeunterschiede zwischen diesen beiden Gruppen bei den relevanten Items geführt hat. Dieser Gedanke sei nachfolgend etwas weiter ausgeführt. Wie schon mehrfach erwähnt, differenzierten bei Bradley und Rettinger (1992) die numerischen Scores signifikant zwischen den *Schuldigen* und den *Unschuldigen mit Tatwissen*. Der Grundannahme des *GAT* zufolge reflektiert dieser Gruppenunterschied einen spezifischen Täuschungseffekt. So soll die wahrheitswidrige Verneinung der relevanten Items seitens der *Schuldigen* mit einer Erhöhung der autonomen Erregung einhergehen, die bei der wahrheitsgemäßen Verneinung der relevanten Items (*Unschuldige mit Tatwissen*) nicht auftritt. Bradley und Rettinger (1992) ziehen in der Diskussion ihrer Befunde jedoch auch noch eine alternative, nämlich **gedächtnispsychologische Erklärung** in Betracht. Sie argumentieren, daß die *Schuldigen* die kritischen Tatdetails möglicherweise reichhaltiger und komplexer im Gedächtnis enkodiert hätten als die *Unschuldigen mit Tatwissen* und daß dies möglicherweise den Ausschlag gegeben habe für die stärkeren physiologischen Reaktionen der *Schuldigen* bei den relevanten Items („... the greater physiological responsiveness of the guilty subjects could be due to the richer and more complex memory codes associated with the key items“ [S. 58]). Während nämlich in der Studie von Bradley und Rettinger (1992) die *Unschuldigen mit Tatwissen* die kritischen Tatdetails nur aus der schriftlichen Instruktion kannten, wurden die *Schuldigen* der kritischen Details zusätzlich auch noch im Rahmen ihrer eigenen Handlungen gewahr. Hinsichtlich des postulierten Zusammenhangs zwischen der Reichhaltigkeit bzw. Komplexität der Gedächtnisenkodierung einerseits und der physiologischen Reaktivität andererseits berufen die Autoren sich auf die Theorie von Lang (1979). Was nun die vorliegende Studie betrifft, kann man genau umgekehrt argumentieren. Die *Täterinnen* kamen mit den kritischen Tatdetails vornehmlich im Zuge der eigenen Tathandlung in Berührung. Dagegen nahmen die *Zeuginnen* die kritischen Details sowohl als Bestandteil des beobachteten Diebstahls als auch als Bestandteil der eigenen Tätigkeit (Aufräumarbeit am Tatort) wahr. Beispielsweise sahen sie, wie die *Täterin* unter dem roten Teppich nach einem Notizzettel suchte; und sie mußten selber den roten Teppich saugen. Insofern ist es denkbar, daß die kritischen Details im Gedächtnis der *Unschuldigen mit Tatwissen* reichhaltiger und komplexer enkodiert wurden als im Gedächtnis der *Schuldigen* (wenngleich dieser Unterschied nicht in den durchgeführten Gedächtnistests zum Tragen kam). Im Sinne der Modellvorstellungen von Lang (1979) könnte dies zu einer relativen Erhöhung der autonomen Reaktionen der *Unschuldigen mit Tatwissen* bei den relevanten Items geführt haben bzw. dazu, daß der erwartete, durch die wahrheits-

widrige vs. wahrheitsgemäße Verneinung bedingte Reaktionsstärkeunterschied zwischen *Schuldigen* und *Unschuldigen mit Tatwissen* nicht zutage trat.

Im Rahmen der Versuchsplanung wurden neben der auf die kritischen Tatdetails bezogenen Gedächtnisleistung auch noch andere Faktoren als **potentielle Störvariablen** in Betracht gezogen, die die Befunde zur Validität des *GAT* möglicherweise hätten systematisch verfälschen können. So wurden die elektrodermale Labilität der Pbn, ihre Motivation, im *GAT* einen unschuldigen Eindruck zu hinterlassen, die subjektiv eingeschätzte persönliche Erfolgswahrscheinlichkeit im *GAT*, die Durchführung etwaiger Manipulationsmaßnahmen im *GAT*, etwaige Vorkenntnisse zur Glaubhaftigkeitsbeurteilung sowie etwaige frühere Teilnahmen an Untersuchungen zur Glaubhaftigkeitsbeurteilung im Versuchsablauf als Kontrollvariablen miterfaßt (zur Begründung s. Abschnitt 4.3.2) und anschließend im Rahmen ergänzender kovarianzanalytischer Berechnungen herauspartialisiert. Der oben diskutierte Haupteffekt der experimentellen Gruppenzugehörigkeit auf die Höhe der numerischen Scores und die auf den Gesamt-*GAT* bezogene Wechselwirkung der Gruppenzugehörigkeit und des Itemtyps auf die SCR-Amplituden erwiesen sich auch noch nach der Eliminierung der Kovariaten-Effekte als statistisch signifikant. Dabei blieben die Effektstärken gegenüber den ursprünglichen statistischen Analysen nahezu unverändert; und auch die kovarianzanalytisch geschätzten SCR-Magnitudenwerte unterschieden sich nur minimal von den empirisch vorgefundenen Reaktionsstärken. Das bedeutet, daß die **Befunde zur Validität des *GAT* nicht durch eine Konfundierung** der UV Status der aussagenden Person mit den genannten Kontrollvariablen **verfälscht** sind.

Zu den berücksichtigten Kontrollvariablen sind jedoch einige kritische Anmerkungen zu machen. So wurde die Variable elektrodermale Labilität erst im Anschluß an die experimentelle Verbrechenssimulation erfaßt. Die Häufigkeit elektrodermaler Spontanfluktuationen (NSRs), die als Kennwert der elektrodermalen Labilität dient, hängt jedoch teilweise auch von der tonischen Aktiviertheit ab (Vossel & Zimmer, 1998). Sofern also die drei Versuchsbedingungen (*Täterinnen, Zeuginnen, falsche Zeuginnen*) unterschiedlich aktivierend wirkten, könnte das in der vorliegenden Studie erfaßte Personenmerkmal „elektrodermale Labilität“ mit der differentiellen Aktiviertheit in den drei experimentellen Gruppen konfundiert sein. Die Gruppenunterschiede in der Anzahl der NSRs waren zwar nicht statistisch signifikant. Dennoch hob sich der Mittelwert der *Täterinnen* relativ deutlich von denen der beiden anderen Gruppen ab (s. Abschnitt 5.1). Es erscheint nicht unplausibel, daß der höhere Wert der *Täterinnen* u.a. auch einen erhöhten tonischen Aktiviertheitsgrad reflektiert, welcher von der experimentellen Aufgabe herrührt. An den durch einfache Befragung retrospektiv erfaßten Kontrollvariablen („Motivation, im *GAT* einen unschuldigen Eindruck zu hinterlassen“, „subjektiv einge-

schätzte persönliche Erfolgswahrscheinlichkeit im *GAT*“, „Durchführung von Manipulationsmaßnahmen im *GAT*“, „Vorkenntnisse zur Glaubhaftigkeitsbeurteilung“ und „frühere Teilnahmen an Untersuchungen zur Glaubhaftigkeitsbeurteilung“) kann man bemängeln, daß sie anfällig sind für irrtümliche und absichtliche Verfälschungen, deren Ausmaß im vorliegenden Fall nicht abzuschätzen ist. Insbesondere zur retrospektiv eingeschätzten persönlichen Erfolgswahrscheinlichkeit im *GAT* ist einschränkend anzumerken, daß über diese Variable indirekt die vor der Untersuchung bestehenden Annahmen der Pbn über die Zuverlässigkeit der psychophysiologischen Glaubhaftigkeitsbeurteilung erfaßt werden sollten. Es ist jedoch äußerst fragwürdig, ob die retrospektiven Angaben der Pbn in einem engen Zusammenhang mit ihren tatsächlichen prospektiven Erwartungen standen. Der Kritik an der Art und Weise der Erfassung der Kontrollvariablen ist jedoch entgegenzuhalten, daß aus versuchsökonomischen Gründen kein adäquateres Vorgehen möglich war. Eine sorgfältigere Erfassung der Kontrollvariablen hätte den ohnehin schon sehr komplexen und zeitintensiven Versuchsablauf noch mehr verlängert und war insofern den Pbn nicht zumutbar.

Abschließend sei noch eine weitere potentielle Erklärung für die fehlende Differenzierung zwischen den *Schuldigen* und den *Unschuldigen mit Tatwissen* in der vorliegenden Studie diskutiert. Hierfür wird zunächst auf einen speziellen Befund aus der Studie von Bradley und Warfield (1984) eingegangen. In dieser Untersuchung gab es u.a. auch eine Gruppe von *Unschuldigen mit Tatwissen*, die sich des Tatbezugs ihrer Kenntnisse nicht bewußt waren („Innocent Associations Group“, Bradley & Warfield, 1984, S. 685). Ähnlich wie die *Zeuginnen* der vorliegenden Studie, hatten die Pbn der „Innocent Associations“-Gruppe die Aufgabe, den Raum aufzuräumen, in welchem das simulierte Verbrechen stattfand. Dabei kamen sie mit denselben kritischen Details in Berührung wie die Pbn, die das Scheinverbrechen begingen. Im Gegensatz zur vorliegenden Studie fand das Scheinverbrechen jedoch zeitversetzt statt, d.h. für die Pbn der „Innocent Associations“-Gruppe bestand vor dem *GAT* keinerlei erkennbarer Zusammenhang zwischen den Einzelheiten, mit denen sie bei der Aufräumarbeit konfrontiert worden waren, und dem aufzuklärenden Verbrechen. Somit waren ihnen die relevanten Details des *GAT* gewissermaßen „zufällig“ aus einem deliktfremden Kontext bekannt. Während die anderen unschuldigen Pbn mit Tatwissen (eine Gruppe von Zeugen und eine Gruppe von Pbn, die schriftlich über den Tathergang informiert wurden) sich in den numerischen Scores nicht signifikant von den *Unschuldigen ohne Tatwissen* unterschieden, erzielte die „Innocent Associations“-Gruppe signifikant höhere Werte. Diesen überraschenden Befund diskutieren die Autoren dahingehend, daß „subjects in this group did not expect that some items on the GKT[34] would be the same as those involved in their

[34] „Guilty Knowledge Test“; zum Zeitpunkt dieser Publikation war der Begriff „Guilty Actions Test“ noch nicht als Bezeichnung für die modifizierte Version des Tatwissentests eingeführt.

innocent activities. Thus they may have become suspicious and very attentive to the crime-relevant items because of this unexpected coincidence.“ (Bradley & Warfield, 1984, S. 688) Die ***Unschuldigen mit Tatwissen*** (*Zeuginnen*) in der vorliegenden Studie waren sich zwar bewußt, daß sich Einzelheiten ihrer neutralen Aufräumarbeit mit Details des aufzuklärenden Diebstahls deckten. Dennoch – oder vielleicht sogar gerade deshalb – ist es **denkbar, daß sie das Erscheinen dieser Einzelheiten im Rahmen der *GAT*-Itemsequenzen besonders stark irritierte**, was möglicherweise eine Intensivierung der autonomen Reaktionen bei den relevanten Items bewirkte. Allerdings muß betont werden, daß sich bei Bradley und Warfield (1984) – im Gegensatz zur vorliegenden Studie – trotz allem eine signifikante Differenzierung zwischen den Tätern und der „Innocent Associations“-Gruppe ergab. Auch die Ergebnisse zweier Untersuchungen zum herkömmlichen *TWT* deuten darauf hin, daß die angesprochene Erklärungsmöglichkeit für die fehlende Differenzierung zwischen *Schuldigen* und *Unschuldigen mit Tatwissen* in der vorliegenden Studie unzutreffend ist. Giesen und Rollison (1980) sowie Stern, Breen, Watanabe und Perry (1981) unterzogen in ihren Simulationsstudien jeweils eine Gruppe von Tätern und eine Gruppe von *Unschuldigen mit Tatwissen* einem konventionellen *TWT*. Ähnlich wie bei Bradley und Warfield (1984) waren den Unschuldigen die relevanten Items des *TWT* aus einem neutralen, deliktfremden Kontext bekannt. Gemäß obiger Argumentation wäre also zu erwarten gewesen, daß die überraschende Konfrontation mit den bekannten (relevanten) Details die *Unschuldigen mit Tatwissen* besonders stark erregte und somit ein auf Schuld hindeutendes physiologisches Reaktionsmuster provozierte. Es ergab sich jedoch eine klare Differenzierung zwischen den *Schuldigen* und den *Unschuldigen mit Tatwissen*. Bei Giesen und Rollison (1980) wurden alle 20 *Unschuldigen mit Tatwissen* anhand der numerischen Scores zutreffend als unschuldig diagnostiziert, die entsprechende Trefferquote bei den *Schuldigen* betrug 19/20. In der Studie von Stern et al. (1981) klassifizierte man 23 der 26 *Unschuldigen mit Tatwissen* korrekt als unschuldig, gegenüber einer Hitrate von 25/26 bei den *Schuldigen*.

Zusammenfassend kann festgehalten werden, daß der zentrale Aspekt des diagnostischen Prinzips des *GAT*, nämlich die Differenzierung zwischen *Schuldigen* und *Unschuldigen mit Tatwissen*, in der vorliegenden Studie nicht bestätigt werden konnte. Auf den ersten Blick ist dies so zu interpretieren, daß die theoretische Grundannahme unzutreffend ist, derzufolge die wahrheitswidrige Verneinung der relevanten Items durch die *Schuldigen* mit einer Verstärkung der autonomen Reaktionen einhergehen soll, verglichen mit der wahrheitsgemäßen Verneinung der relevanten Items durch die *Unschuldigen mit Tatwissen*. Es wurden jedoch auch einige alternative Erklärungsansätze für den negativen Validitätsbefund der vorliegenden Studie erwogen. Hier ist an erster Stelle die potentielle Unangemessenheit des hier verwendeten Frageformats zu nennen, wel-

ches nicht vollständig mit demjenigen früherer *GAT*-Studien übereinstimmte und möglicherweise dazu führte, daß der nur für die Gruppe der *Schuldigen* (*Täterinnen*) anvisierte Effekt der wahrheitswidrigen Verneinung der relevanten Items auch bei den *Unschuldigen mit Tatwissen* (*Zeuginnen*) auftrat. Darüber hinaus ist nicht auszuschließen, daß die fehlende Differenzierung z.T. auch durch die insgesamt schwache elektrodermale Reagibilität in der vorliegenden Untersuchung bedingt ist, über deren Ursachen allerdings nur spekuliert werden kann. In diesem Zusammenhang verdient es zumindest Erwähnung, daß sich sowohl in den numerischen Scores als auch in den SCR-Magnituden eine hypothesenkonforme Differenzierung zwischen den *Schuldigen* und *Unschuldigen mit Tatwissen* vage andeutete, wobei freilich die beobachteten Mittelwertsunterschiede von einer etwaigen statistischen Bedeutsamkeit weit entfernt waren. Im Hinblick auf eine etwaige individualdiagnostische Anwendung sprechen die vorgefundenen Gruppenunterschiede dafür, daß sich der *GAT* nur zur Differenzierung zwischen Personen mit und solchen ohne Tatwissen eignet, jedoch nicht zur gezielten Identifizierung unschuldiger Personen, die tatspezifische Kenntnisse besitzen.

6.2.3 Treffsicherheit des *Guilty Actions Tests* in der vorliegenden Untersuchung

Angesichts der Tatsache, daß sich in den inferenzstatistischen Analysen keine signifikanten hypothesenkonformen Gruppenunterschiede zwischen *Schuldigen* und *Unschuldigen mit Tatwissen*, sondern nur zwischen Personen mit und solchen ohne Tatwissen ergaben, war nicht zu erwarten, daß sich für die Gruppe der *Unschuldigen mit Tatwissen* eine hohe diagnostische Treffsicherheit ergeben würde. Andererseits war es jedoch durchaus möglich, daß für die *Schuldigen* und die *Unschuldigen ohne Tatwissen* hohe Hitraten resultierten.

Die diagnostische Zuordnung der Pbn erfolgte zunächst anhand der **numerischen Scores**, wobei die – im herkömmlichen *TWT* übliche – **Entscheidungsregel** herangezogen wurde, derzufolge ein Testteilnehmer als schuldig zu klassifizieren ist, wenn er mehr als die Hälfte der möglichen Punkte erzielt. Auf diese Weise wurden sämtliche *Unschuldigen ohne Tatwissen* zutreffend als unschuldig diagnostiziert. Dagegen lag die Trefferrate bei den *Schuldigen* etwa auf dem Zufallsniveau von 50% (bei Zugrundelegung von SCR-Quantifizierungsmethode B wurden sogar deutlich weniger als die Hälfte der *Schuldigen* [13/34] zutreffend diagnostiziert). Die auf diese Weise erzielte Trefferquote bei den *Unschuldigen mit Tatwissen* ist aus zwei Gründen wenig aussagekräftig. Erstens bildet der numerische Score nicht das graduelle Ausmaß des Reaktionsstärkeunterschieds zwischen relevanten und irrelevanten Items ab und ist daher prinzi-

piell nicht geeignet, zwischen *Schuldigen* und *Unschuldigen mit Tatwissen* zu differenzieren. Zweitens sieht die verwendete Entscheidungsregel nur eine dichotome Kategorisierung vor, d.h. die im *GAT* anvisierte dritte Diagnosekategorie „unschuldig mit Tatwissen“ findet gar keine Berücksichtigung. Andererseits war im Sinne der Grundannahme des *GAT* jedoch zu erwarten, daß die *Unschuldigen mit Tatwissen* deutlich seltener als schuldig klassifiziert würden als die *Schuldigen*. Dies war jedoch nicht der Fall (bei Zugrundelegung von SCR-Quantifizierungsmethode B wurden sogar mehr *Unschuldige mit Tatwissen* als „schuldig“ klassifiziert [16 vs. 13]).

Das Problem, daß in der A-priori-Entscheidungsregel für die numerischen Scores nur eine dichotome Kategorisierung vorgesehen ist, wurde umgangen, indem die **numerischen Scores einer diskriminanzanalytischen Klassifikationsprozedur unterzogen** wurden, wobei das vorherzusagende Kriterium der experimentellen Gruppenzugehörigkeit entsprach, also die drei Ausprägungen „schuldig“, „unschuldig mit Tatwissen“ und „unschuldig ohne Tatwissen“ hatte. Analog zur diskriminanzanalytischen Vorgehensweise bei der inhaltsorientierten Glaubhaftigkeitsbeurteilung wurden die Klassifikationsregeln an einer Hälfte der Gesamtstichprobe (Konstruktionsstichprobe) berechnet und dann an der anderen Hälfte (Klassifikationsstichprobe) kreuzvalidiert („Hold-out sample“-Methode). Die Konstruktions- und die Klassifikationsstichprobe waren identisch mit denjenigen, die im Rahmen der Treffsicherheitsanalysen der inhaltsorientierten Glaubhaftigkeitsbeurteilung herangezogen wurden. Bei der Kreuzvalidierung wurden 14 von 17 *Unschuldigen ohne Tatwissen*, fünf von 17 *Unschuldigen mit Tatwissen* und vier von 17 *Schuldigen* zutreffend klassifiziert (diese Angaben beziehen sich auf SCR-Quantifizierungsmethode A; die Häufigkeiten bei Zugrundelegung von Methode B wichen davon nicht entscheidend ab). Das heißt, daß die Trefferquote nur bei den *Unschuldigen ohne Tatwissen* – hier allerdings deutlich – über der Zufallswahrscheinlichkeit von 33.3% lag. Die *Schuldigen* wurden v.a. als „unschuldig mit Tatwissen“ fehlklassifiziert. Die Falschzuordnungen bei den *Unschuldigen mit Tatwissen* verteilten sich in etwa gleichmäßig auf die Kategorien „schuldig“ und „unschuldig ohne Tatwissen“. Die beschriebenen Treffer- bzw. Fehlerraten sind insofern wenig aussagekräftig, als auch hier das grundsätzliche Problem besteht, daß die numerische Auswertungsmethode die im *GAT* intendierten Reaktionsstärkeunterschiede zwischen *Schuldigen* und *Unschuldigen mit Tatwissen* bei den relevanten Items nicht adäquat abbilden kann. Allerdings ist interessant, daß auch bei der diskriminanzanalytischen Klassifikation anhand der numerischen Scores die **Rate der als „schuldig“ klassifizierten *Unschuldigen mit Tatwissen* nicht geringer** war **als die Rate der als „schuldig“ klassifizierten *Schuldigen***. Noch auffälliger ist, daß **im Vergleich zu den *Unschuldigen mit Tatwissen* deutlich mehr *Schuldige* als „unschuldig mit Tatwissen“ eingestuft** wurden. Am Rande sei noch erwähnt, daß reanalysiert wurde, welche Wertebereiche der numerischen

Scores den drei möglichen diskriminanzanalytischen Zuordnungen entsprachen. Es zeigte sich, daß in der Diskriminanzanalyse solche Pbn als „unschuldig ohne Tatwissen“ klassifiziert wurden, deren numerische Scores im Bereich von null bis sechs lagen. Probandinnen mit numerischen Scores von sieben bis 13 wurden als „unschuldig mit Tatwissen“ klassifiziert, solche mit Scores von 14 und höher wurden der Gruppe der „Schuldigen“ zugeordnet. (Diese Intervalle gelten für SCR-Quantifizierungsmethode A. Bei Zugrundelegung von Methode B waren die entsprechenden Intervalle der numerischen Scores 0 – 6, 7 – 11 und 12 – 20.)

Um die Unzulänglichkeit der numerischen Auswertungsmethode im Hinblick auf eine etwaige Unterscheidung von *Schuldigen* und *Unschuldigen mit Tatwissen* vollständig zu umgehen, wurden **intraindividuelle Differenzmaße** gebildet, indem pro Pb die persönliche SCR-Magnitude bei den 40 irrelevanten *GAT*-Items von der persönlichen SCR-Magnitude bei den zehn relevanten Items subtrahiert wurde. Der Differenzwert diente dann im Rahmen einer **diskriminanzanalytischen Klassifikationsprozedur** („Hold-out sample“-Methode) als Prädiktor der Gruppenzugehörigkeit. Die so erhaltenen Klassifizierungsergebnisse belegen eindeutig, daß der ***GAT* in der vorliegenden Studie nicht zur Identifizierung von *Unschuldigen mit Tatwissen* geeignet** war. Sämtliche Pbn dieser Gruppe wurden fehlklassifiziert. Zudem fiel auf, daß auch aus den beiden anderen Gruppen niemand (bei SCR-Quantifizierungsmethode B lediglich eine *Unschuldige ohne Tatwissen*) als „unschuldig mit Tatwissen“ (fehl-) klassifiziert wurde. Die *Unschuldigen mit Tatwissen* wurden überwiegend als „unschuldig ohne Tatwissen“ (bei SCR-Quantifizierungsmethode B in etwa gleich häufig als „schuldig“ bzw. „unschuldig ohne Tatwissen“) fehlklassifiziert. Die höchste Trefferquote resultierte bei den *Unschuldigen ohne Tatwissen* (15/17 bei Quantifizierungsmethode A; 14/17 bei Methode B). Auch bei den *Schuldigen* lag die Hitrate mit 64.7% (SCR-Quantifizierungsmethoden A und B) deutlich über der Zufallswahrscheinlichkeit von 33.3%. Das einzig positive Resultat im Sinne der intendierten Differenzierung zwischen *Schuldigen* und *Unschuldigen mit Tatwissen* war, daß letztere deutlich seltener als „schuldig“ eingestuft wurden (6/17 [SCR-Quantifizierungsmethode A]; 8/17 [Methode B]), verglichen mit der Rate der als „schuldig“ klassifizierten *Schuldigen* (11/17).

Die Art und Weise der Kreuzvalidierung der diskriminanzanalytischen Klassifikation, also die „Hold-out sample“-Methode, wurde bereits im Zusammenhang mit der inhaltsorientierten Glaubhaftigkeitsbeurteilung kritisch gewürdigt. Daher sei an dieser Stelle nur auf die entsprechenden Ausführungen am Ende von Abschnitt 6.1.3 verwiesen.

Zusammenfassend bleibt festzuhalten, daß sich nur für die *Unschuldigen ohne Tatwissen* zufriedenstellende Trefferquoten ergaben. Diese Pbn wurden anhand der A-priori-

Entscheidungsregel ausnahmslos zutreffend für unschuldig befunden; und auch in den beiden diskriminanzanalytischen Klassifikationsprozeduren ergaben sich hohe Hitraten. Die Treffsicherheit in bezug auf die *Schuldigen* war eher mäßig. Sowohl bei Anwendung der A-priori-Entscheidungsregel für die numerischen Scores als auch in der diskriminanzanalytischen Klassifikation anhand der numerischen Scores wurden selbst die per Zufall zu erwartenden Hitraten von 50% bzw. 33.3% verfehlt. Allerdings lag die Trefferquote, welche in der diskriminanzanalytischen Klassifikation anhand der intraindividuellen Reaktionsstärkedifferenzen zwischen relevanten und irrelevanten Items erzielt wurde, deutlich über der Zufallswahrscheinlichkeit von 33.3%. Bezüglich der *Unschuldigen mit Tatwissen* sind aufgrund der Problematik der numerischen Auswertungsmethode nur die diskriminanzanalytischen Trefferquoten interpretierbar, bei denen die intraindividuelle Reaktionsstärkedifferenz zwischen relevanten und irrelevanten Items als Prädiktor diente. Die Tatsache, daß es hier nur Fehlklassifikationen gab, unterstreicht eindrucksvoll, daß der *GAT* zur gezielten Identifizierung von *Unschuldigen mit Tatwissen* völlig ungeeignet war. Während also innerhalb der Pbn mit Tatwissen keine Trennung zwischen Schuldigen und Unschuldigen gelang, wurden aber immerhin – im Sinne des herkömmlichen *TWT* – die Pbn mit Tatwissen einigermaßen zuverlässig von den Pbn ohne Tatwissen differenziert. Dies sei noch anhand einiger Zahlenwerte illustriert. Hätte man die Pbn mit Tatwissen (*Schuldige* und *Unschuldige mit Tatwissen*) zu einer Gruppe zusammengefaßt und den numerischen Score – so wie es von der Logik des *TWT* her eigentlich geboten ist – als Indikator von Tatwissen (nicht von Schuld) aufgefaßt, so wären gemäß der A-priori-Entscheidungsregel für die numerischen Scores 100% der Personen ohne Tatwissen und 45.6% der Personen mit Tatwissen (47.1% der *Schuldigen*; 44.1% der *Unschuldigen mit Tatwissen*) zutreffend diagnostiziert worden. Bei der diskriminanzanalytischen Klassifikation anhand der numerischen Scores hätten sich Hitraten von 82.4% für die Personen ohne Tatwissen und 70.6% für die Personen mit Tatwissen (82.4% der *Schuldigen*; 58.8% der *Unschuldigen mit Tatwissen*) ergeben (für genauere Angaben zur Diskriminanzanalyse s. Anhang F, Tabelle F.120). Unter Verwendung der intraindividuellen Reaktionsstärkedifferenz zwischen relevanten und irrelevanten Items als Prädiktor der dichotomen Gruppenzugehörigkeit hätten sich diskriminanzanalytische Trefferquoten von 88.2% für die Personen ohne Tatwissen und 50% für die Personen mit Tatwissen (64.7% der *Schuldigen*; 35.3% der *Unschuldigen mit Tatwissen*) ergeben (für genauere Angaben zur Diskriminanzanalyse s. Anhang F, Tabelle F.122).[35]

[35] Diese Angaben beziehen sich auf SCR-Quantifizierungsmethode A. Bei Zugrundelegung von Methode B beliefen sich die Trefferquoten gemäß der A-priori-Entscheidungsregel auf 100% für die Personen ohne Tatwissen und 42.7% für die Personen mit Tatwissen (38.2% der *Schuldigen*; 47.1% der *Unschuldigen mit Tatwissen*). In der diskriminanzanalytischen Klassifikation anhand der numerischen Scores betrugen die Hitraten 76.5% für die Personen ohne und 64.7% (70.6% bzw. 58.8%) für die Personen mit Tatwissen (für genauere Angaben zur Diskriminanzanalyse s. Anhang F, Tabelle F.121). Die diskriminanzanalytische Klassifikation anhand der intraindividuellen Differenzwerte erbrachte Treffer-

6.3 Zum diagnostischen Potential der naiven Glaubhaftigkeitsbeurteilung in der vorliegenden Studie

Aus dem Gesamtpool von 102 Aussagen wurden zwei Zufallsstichproben von je 15 Aussagen (5 *Täterinnen*, 5 *Zeuginnen*, 5 *falsche Zeuginnen*) gezogen. Die beiden Aussagenstichproben wurden jeweils 16 Personen dargeboten, die keine speziellen Vorkenntnisse auf dem Gebiet der Glaubhaftigkeitsbeurteilung besaßen. Diese naiven Beurteiler stuften auf einer standardisierten Skala die Glaubhaftigkeit der Aussagen ein. Die Präsentation der Aussagen erfolgte in Bild und Ton (Videoaufzeichnungen), so daß die naiven Beurteiler sich in ihren Einschätzungen neben dem Aussageinhalt auch noch vom aussagebegleitenden nonverbalen und paraverbalen Ausdrucksverhalten leiten lassen konnten**. Weder wenn man die beiden Aussagenstichproben separat analysierte, noch wenn man sie zu einer Stichprobe (von 30 Aussagen) zusammenfaßte, zeigten sich signifikante Gruppenunterschiede zwischen den *Täterinnen*, *Zeuginnen* und *falschen Zeuginnen* hinsichtlich der naiv eingeschätzten Glaubhaftigkeit**. Bei allen drei experimentellen Gruppen lag die durchschnittliche naive Beurteilung zwischen den beiden mittleren Skalenstufen 4 (*eher unglaubwürdig*) und 5 (*eher glaubwürdig*). Auch die Wechselwirkung zwischen den Faktoren Status der aussagenden Person und (naiver) Rater erwies sich in keiner der beiden Aussagenstichproben als statistisch bedeutsam. Lediglich der Rater-Haupteffekt war signifikant, allerdings nur in einer der beiden Aussagenstichproben. Diese Ergebnisse deuten darauf hin, daß die naive Glaubhaftigkeitsbeurteilung alles in allem nicht zwischen erlebnisbezogenen und erfundenen Aussagen zu differenzieren vermag.

Die individualdiagnostische Treffsicherheit der naiven Glaubhaftigkeitsbeurteilung wurde bestimmt, indem die Ratings auf der achtstufigen Skala zunächst dichotomisiert („unglaubhaft“ vs. „glaubhaft“) und dann zur tatsächlichen Glaubhaftigkeit der Aussagen in Beziehung gesetzt wurden. Über alle 32 naiven Beurteiler hinweg ergab sich eine **Gesamttrefferquote, die in etwa der Zufallswahrscheinlichkeit von 50% entsprach**. Dies galt **gleichermaßen für die erlebnisbezogenen** (*Zeuginnen*) **und konfabulierten Schilderungen** (*Täterinnen* und *falsche Zeuginnen*).

Zusammenfassend kann festgehalten werden, daß die Güte der naiven Glaubhaftigkeitsbeurteilung in der vorliegenden Studie insgesamt sehr gering war. Die wahr- und falschaussagenden Gruppen wurden im Durchschnitt als gleich glaubhaft bzw. unglaubhaft eingestuft. Dementsprechend lag auch die individualdiagnostische Treffsicherheit auf dem Zufallsniveau. Die Resultate der vorliegenden Untersuchung liegen somit **im**

---

raten von 82.4% für die Personen ohne und 55.9% (64.7% bzw. 47.1%) für die Personen mit Tatwissen (s. genauer Tabelle F.123).

Trend der sonstigen empirischen Befundlage zur naiven Glaubhaftigkeitsbeurteilung. Köhnken (1990, S. 62) faßt den Erkenntnisstand der bis dato erschienenen Literaturübersichten und Metaanalysen (z.B. DePaulo, Stone & Lassiter, 1985; Zuckerman et al., 1981) dahingehend zusammen, daß die naive Beurteilung in aller Regel Trefferquoten im Bereich von 45% bis 60% erziele. In den meisten Studien werde die per Zufall zu erwartende Hitrate von 50% nur unwesentlich (wenngleich statistisch signifikant) übertroffen. Auch in den Studien zur inhaltsorientierten Glaubhaftigkeitsbeurteilung, in welchen als Kontrollbedingung eine naive Beurteilung vorgenommen wurde, überstiegen die diesbezüglichen Trefferraten kaum das Zufallsniveau (Krahé & Kundrotas, 1992: 63%; Landry & Brigham, 1992: 47%; Steller et al., 1988, zit. nach Steller, 1989, S. 144ff.: 58%). Während sich in der vorliegenden Untersuchung die durchschnittlichen Glaubhaftigkeitsurteile (auf der achtstufigen Skala) bei den erlebnisbezogenen und erfundenen Aussagen nicht signifikant unterschieden, hielten die naiven Beurteiler bei Ruby und Brigham (1998) die Falschaussagen sogar für signifikant glaubhafter als die wahren Geschichten.

Da es denkbar ist, daß die Fähigkeit zur naiven Glaubhaftigkeitsbeurteilung **interindividuellen Schwankungen** unterliegt, wurden die Ratings auch noch getrennt nach den einzelnen Beurteilern analysiert. In den individuellen Reanalysen ergaben sich für zwölf der insgesamt 32 naiven Rater „erwartungskonforme“ Tendenzen in den Gruppenmittelwerten, d.h. diese zwölf Beurteiler hielten jeweils die erlebnisbezogenen Aussagen der *Zeuginnen* für glaubhafter als die erfundenen Schilderungen der *Täterinnen* und der *falschen Zeuginnen*. Der Effekt des Faktors Status der aussagenden Person erwies sich jedoch bei keinem der zwölf naiven Beurteiler als statistisch bedeutsam, was mit dem Befund korrespondiert, daß sich in den ursprünglichen statistischen Analysen für keine der beiden Aussagenstichproben eine signifikante Wechselwirkung zwischen den Faktoren Status der aussagenden Person und Rater ergab. Bei der differenzierten Betrachtung der Trefferraten einzelner Beurteiler stellte sich heraus, daß fünf der 32 naiven Rater jeweils sowohl die glaubhaften Aussagen der *Zeuginnen* als auch die unglaubhaften Aussagen der *Täterinnen* und der *falschen Zeuginnen* in jeweils mehr als der Hälfte der Fälle zutreffend diagnostizierten. Vier dieser fünf Rater gehörten auch zu der Gruppe der zwölf Beurteiler, deren Ratings auf der achtstufigen Skala bei den *Zeuginnen* tendenziell höher ausfielen als bei den *Täterinnen* und den *falschen Zeuginnen*. Somit betrug der Anteil der naiven Rater, deren Urteilsgüte tendenziell als zufriedenstellend eingestuft werden konnte, immerhin 12.5% (4/32). Zwar verfehlten bei diesen vier naiven Beurteilern die Gruppenunterschiede (höhere Glaubhaftigkeitsratings bei den *Zeuginnen* im Vergleich zu den *Täterinnen* und den *falschen Zeuginnen*) jeweils die Signifikanzgrenze; das Ausbleiben statistischer Signifikanz könnte jedoch auch durch die geringen Zellbesetzungen bedingt gewesen sein. So bearbeiteten die naiven Beur-

teiler ja jeweils nur fünf Aussagen von *Täterinnen*, fünf Schilderungen von *Zeuginnen* und fünf Bekundungen *falscher Zeuginnen*.

Um diese Hypothese zu untermauern, seien abschließend auch noch die Resultate einer **ergänzenden Datenerhebung** berichtet, bei welcher zwei naive Beurteiler (Rater X [31 Jahre, männlich, Sozialarbeiter], Rater Y [58 Jahre, weiblich, Küchengehilfin]) jeweils sämtliche 102 experimentellen Aussagen beurteilten, so daß im Hinblick auf die statistischen Analysen eine hohe Zellbesetzung gewährleistet war (jeweils n = 34 Aussagen von *Täterinnen*, *Zeuginnen* bzw. *falschen Zeuginnen*). Eine zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Rater (X und Y) ergab eine signifikante Wechselwirkung beider Faktoren (s. genauer Tabelle F.124 im Anhang F), die in Abbildung F.8 (Anhang F) graphisch illustriert ist. Rater X hielt die Aussagen der *Zeuginnen* für signifikant glaubhafter als die Schilderungen der *Täterinnen* und der *falschen Zeuginnen*, während die Beurteilungen von Rater Y nicht signifikant zwischen den Gruppen differenzierten (s. genauer Anhang F, Tabelle F.126 und F.127).[36] Die basisratenkorrigierte Gesamttrefferquote von Rater X war deutlich höher als die von Rater Y (s. Anhang F, Tabelle F.128). Rater X erzielte jedoch nur bei den glaubhaften Aussagen eine deutlich über der Zufallswahrscheinlichkeit liegende Trefferquote, schnitt dafür jedoch bei den unglaubhaften Aussagen der *Täterinnen* und der *falschen Zeuginnen* gleichermaßen schlecht ab. Zu der ergänzenden Datenerhebung ist einschränkend anzumerken, daß aus technischen Gründen keine Kontrolle von Sequenzeffekten gewährleistet werden konnte. Beide Rater bearbeiteten die Aussagen in der Reihenfolge 1, 2, ... 102. Zudem erfolgte die Präsentation der Aussagen unter unkontrollierten Rahmenbedingungen. Beide Rater schauten sich die Aussagevideos jeweils alleine zuhause an, wobei die Anzahl der am Stück bearbeiteten Aussagen variierte.

6.4 Vergleich der inhaltsorientierten, der psychophysiologischen und der naiven Glaubhaftigkeitsbeurteilung

Ein Vergleich der verschiedenen Methoden der Glaubhaftigkeitsbeurteilung läßt sich am einfachsten vornehmen, indem man die jeweils erzielten diagnostischen Trefferquoten gegenüberstellt. Bevor dies getan wird, sollen jedoch zunächst noch einmal die zentralen Aspekte des vorliegenden Versuchsplans sowie die vorliegenden Befunde zu

[36] In der zweifaktoriellen ANOVA erwiesen sich neben der Interaktion auch die beiden Haupteffekte als statistisch bedeutsam (s. Tabelle F.124 im Anhang F). Die Urteile von Rater X auf der achtstufigen Skala waren im Durchschnitt signifikant höher als die von Rater Y. Die Aussagen der *falschen Zeuginnen* wurden insgesamt für signifikant weniger glaubhaft gehalten als die Aussagen der *Zeuginnen*, während beide Gruppen sich nicht signifikant von den *Täterinnen* unterschieden (s. genauer Tabelle F.125 im Anhang F).

den Grundannahmen der inhaltsorientierten Glaubhaftigkeitsbeurteilung und des *GAT* kurz rekapituliert werden.

In der vorliegenden Untersuchung ging es um die Frage, ob die inhaltsorientierte und die psychophysiologische Glaubhaftigkeitsbeurteilung unterschiedlich gut zwischen glaubhaften und unglaubhaften Bekundungen differenzieren. Da die inhaltsorientierte Methode zur Beurteilung vermeintlicher Zeugenaussagen, der psychophysiologische Ansatz hingegen primär zur Beurteilung der Täterschaftsabstreitung durch Beschuldigte eingesetzt wird, mußte ein Paradigma entworfen werden, in welchem die Pbn gleichzeitig als mutmaßliche *Zeuginnen* und als Tatverdächtige in Frage kamen, so daß bei jeder Pb sowohl eine inhaltsorientierte Glaubhaftigkeitsbeurteilung (der vermeintlichen Zeugenaussage) als auch eine psychophysiologische Glaubhaftigkeitsbeurteilung (der Abstreitung des Tatvorwurfs) vorgenommen werden konnte. Daher wurde ein Teil der Pbn als *Täterinnen,* ein anderer Teil der Pbn als *Zeuginnen* in ein experimentelles Scheinverbrechen verwickelt. Die *Täterinnen* stritten anschließend den Tatvorwurf wahrheitswidrig ab und machten eine wahrheitswidrige „Zeugenaussage“. Die *Zeuginnen* stritten den Tatvorwurf wahrheitsgemäß ab und machten eine wahrheitsgemäße Zeugenaussage. Folglich war es im Rahmen dieses experimentellen Grunddesigns möglich, auf der Grundlage der gleichen Fälle sowohl die Treffsicherheit der inhaltsorientierten als auch die der psychophysiologischen Glaubhaftigkeitsbeurteilung zu bestimmen, und zwar jeweils getrennt für glaubhafte (*Zeuginnen*) und unglaubhafte Bekundungen (*Täterinnen*).

Die Erweiterung des experimentellen Grunddesigns um die Bedingung *falsche Zeuginnen* war aus zwei Gründen erforderlich. Zum einen sollte hierdurch kontrolliert werden, ob die *Täterinnen* aufgrund dessen, daß sie einen Erlebnisbezug zum Tathergang hatten, ihren erfundenen „Zeugenaussagen“ möglicherweise eine höhere inhaltliche Qualität verleihen konnten als Personen, die keinerlei Erlebnisbezug zum Tathergang hatten (*falsche Zeuginnen*). Wäre dies der Fall gewesen, so hätte man die inhaltsanalytische Trefferquote bei den *Täterinnen* nur noch eingeschränkt interpretieren dürfen. Zum anderen war die Einführung der Gruppe *falsche Zeuginnen* geboten, weil als psychophysiologische Beurteilungsmethode der *GAT* gewählt wurde. Dieses Verfahren soll grundsätzlich zwischen drei Personengruppen differenzieren, nämlich zwischen *Schuldigen, Unschuldigen mit Tatwissen* und *Unschuldigen ohne Tatwissen*. Die letztgenannte Personengruppe war erst durch Einführung der Bedingung *falsche Zeuginnen* im Versuchsplan repräsentiert, d.h. eine umfassende Validierung des *GAT* war erst in dem erweiterten Versuchsplan möglich.

Die Grundannahme der inhaltsorientierten Glaubhaftigkeitsbeurteilung („Undeutsch-Hypothese“) konnte in der vorliegenden Untersuchung prinzipiell bestätigt werden. Dies kam insbesondere darin zum Ausdruck, daß die inhaltsanalytischen Gesamtscores bei den glaubhaften Aussagen signifikant höher waren als bei den unglaubhaften. Dabei ist besonders hervorzuheben, daß die unglaubhaften Aussagen der *Täterinnen* sich im Gesamtscore nicht signifikant von den unglaubhaften Aussagen der *falschen Zeuginnen* unterschieden. Das bedeutet, daß die bei den *Täterinnen* erzielte Trefferquote durchaus interpretierbar war.

Die Grundannahme des *GAT* konnte dagegen nur teilweise bestätigt werden. Personen mit Tatwissen reagierten bei den relevanten Items signifikant stärker als Personen ohne Tatwissen. Der innovative Aspekt des *GAT* gegenüber dem herkömmlichen *TWT*, daß nämlich *Schuldige* aufgrund der wahrheitswidrigen Verneinung bei den relevanten Items stärker reagieren sollen als *Unschuldige mit Tatwissen*, fand jedoch keine Bestätigung. Dies bedeutet, daß der *GAT* grundsätzlich nicht zur direkten Überprüfung der Glaubhaftigkeit der Tatabstreitung geeignet ist, sondern nur zur Differenzierung zwischen Personen mit und solchen ohne tatspezifische Kenntnisse.

Tabelle 29 bietet nun eine Gegenüberstellung der prozentualen Trefferquoten der inhaltsorientierten Glaubhaftigkeitsbeurteilung, des *GAT* und der naiven Glaubhaftigkeitsbeurteilung. In Abschnitt 6.1.3 wurde erläutert, daß in bezug auf die inhaltsorientierte Glaubhaftigkeitsbeurteilung diejenigen Trefferquoten am aussagekräftigsten sind, die sich ergaben, wenn der inhaltsanalytische Gesamtscore im Rahmen einer Diskriminanzanalyse als Prädiktor der tatsächlichen Glaubhaftigkeit diente. Diese Trefferquoten sind auch in Tabelle 29 aufgeführt. In Abschnitt 6.2.3 wurde erläutert, daß im Hinblick auf den *GAT* diejenigen Trefferquoten am aussagekräftigsten sind, die resultierten, wenn man die intraindividuelle Reaktionsstärkedifferenz zwischen relevanten und irrelevanten Items als diskriminanzanalytischen Prädiktor der tatsächlichen Gruppenzugehörigkeit einsetzte. Diese Hitraten sind auch in Tabelle 29 abgetragen (wobei anzumerken ist, daß sie auf SCR-Quantifizierungsmethode A basieren). Als Kennwerte für die diagnostische Treffsicherheit der naiven Glaubhaftigkeitsbeurteilung sind in Tabelle 29 die durchschnittlichen Trefferquoten der 32 naiven Beurteiler abgetragen (vgl. Abschnitt 5.4.3).

Tabelle 29. Prozentuale Trefferquoten der inhaltsorientierten, der psychophysiologischen (*GAT*) und der naiven Glaubhaftigkeitsbeurteilung

	Status der aussagenden Person			
Beurteilungs-methode	*Täterinnen* (*Schuldige* [mit Tatwissen])	*Zeuginnen* (*Unschuldige mit Tatwissen*)	*falsche Zeuginnen* (*Unschuldige ohne Tatwissen*)	Gesamt-trefferquote
inhaltsorientiert	47.1	82.4	52.9	66.2 [a]
GAT	64.7	0	88.2	51.0
naiv	53.1	51.9	53.1	52.5 [a]

Anmerkung: [a] Gesamttrefferquote bei Korrektur der Basisraten von glaubhaften und unglaubhaften Aussagen.

Bei der Interpretation der Werte in Tabelle 29 ist zu berücksichtigen, daß die Trefferraten der naiven Glaubhaftigkeitsbeurteilung nicht wie diejenigen der inhaltsorientierten Beurteilung und des *GAT* auf diskriminanzanalytisch bestimmten Klassifikationsregeln beruhen. Dies schränkt die Vergleichbarkeit der naiven Trefferquoten mit denjenigen der beiden anderen Methoden ein. Die Vergleichbarkeit wäre besser, wenn auch die Hitraten der naiven Beurteilung auf einer diskriminanzanalytischen Klassifikationsprozedur basierten. Im vorliegenden Fall war die Aussagenstichprobe bei der naiven Glaubhaftigkeitsbeurteilung jedoch zu klein (jeweils 10 Aussagen von *Täterinnen*, *Zeuginnen* bzw. *falschen Zeuginnen*), um eine einigermaßen interpretierbare Diskriminanzanalyse durchführen zu können.[37] Ferner ist bei der Interpretation von Tabelle 29 zu beachten, daß es bei der inhaltsorientierten und der naiven Beurteilung jeweils nur zwei diagnostische Kategorien gibt („glaubhaft“ vs. „unglaubhaft“), wohingegen der *GAT* grundsätzlich zwischen drei diagnostischen Kategorien differenzieren soll („schuldig“, „unschuldig mit Tatwissen“, „unschuldig ohne Tatwissen“). Während also bei der inhaltsorientierten und der naiven Beurteilung die Wahrscheinlichkeit, allein durch Zufall eine richtige diagnostische Zuordnung zu treffen, jeweils 50% beträgt, ist die Zufallswahrscheinlichkeit beim *GAT* auf 33.3% zu beanschlagen.

[37] Der Vollständigkeit halber wurden dennoch diskriminanzanalytische Trefferquoten der naiven Beurteilung berechnet. Die Vorgehensweise und Resultate seien kurz beschrieben: Vorherzusagendes Kriterium war die tatsächliche Glaubhaftigkeit der Aussagen (glaubhaft vs. unglaubhaft), wobei die Aussagen der *Täterinnen* und der *falschen Zeuginnen* zur Kategorie „unglaubhaft“ zusammengefaßt wurden, während die Schilderungen der *Zeuginnen* die Kategorie „glaubhaft“ repräsentierten. Als Prädiktor der Kategorienzugehörigkeit diente das naive Glaubhaftigkeitsurteil auf der achtstufigen Ratingskala, wobei pro Aussage die Urteile der jeweils 16 Rater zu einem Wert (Prädiktor) gemittelt wurden. Die Klassifikationsregeln wurden anhand von Aussagenstichprobe II (vgl. Abschnitt 4.4.3 und 5.4.2) berechnet; die Kreuzvalidierung erfolgte an Aussagenstichprobe I. In der Kreuzvalidierung wurden die erfundenen Aussagen der *Täterinnen* und *falschen Zeuginnen* zu 40% (2 von 5) bzw. 80% (4 von 5) zutreffend als „unglaubhaft“ klassifiziert. Eine der fünf erlebnisbasierenden Aussagen der *Zeuginnen* (20%) wurde korrekt als „glaubhaft“ eingeordnet. Die basisratenkorrigierte Gesamttrefferquote (40%) war somit noch niedriger als diejenige in Tabelle 29. Die diskriminanzanalytische Trennung zwischen den beiden Aussagekategorien war nicht signifikant (s. Anhang F, Tabelle F.129).

Vor diesem Hintergrund fällt an Tabelle 29 zunächst auf, daß die Gesamttrefferquoten der inhaltsorientierten Glaubhaftigkeitsbeurteilung und des *GAT* jeweils über der Zufallswahrscheinlichkeit von 50% bzw. 33.3% lagen, wohingegen die diagnostische Treffsicherheit der naiven Beurteilung dem Zufallsniveau entsprach. Das bedeutet daß beide Methoden der forensischen Glaubhaftigkeitsbeurteilung der naiven Vorgehensweise überlegen sind. Letztere erzielte sowohl bei den glaubhaften Aussagen (*Zeuginnen*) als auch bei den unglaubhaften Aussagen (*Täterinnen* und *falsche Zeuginnen*) Trefferraten auf dem Zufallsniveau.

Die Überlegenheit der inhaltsorientierten gegenüber der naiven Glaubhaftigkeitsbeurteilung beschränkt sich auf die Identifizierung glaubhafter Aussagen. Während die inhaltsanalytische Trefferquote in bezug auf letztere (*Zeuginnen*) deutlich über dem Zufallsniveau lag, bewegten sich die Hitraten in bezug auf die unglaubhaften Aussagen (*Täterinnen* und *falsche Zeuginnen*) – ganz wie bei der naiven Beurteilung – auf dem Zufallsniveau.

Der *GAT* ist nur mittelbar mit der naiven Glaubhaftigkeitsbeurteilung vergleichbar, da es hier nicht um die dichotome Unterscheidung „glaubhaft vs. unglaubhaft“, sondern um die dreifache Unterscheidung „schuldig vs. unschuldig mit Tatwissen vs. unschuldig ohne Tatwissen“ geht. Dennoch fällt auf, daß bei den *Täterinnen* und den *falschen Zeuginnen* die mit dem *GAT* erzielten Hitraten nicht nur deutlich über der Zufallswahrscheinlichkeit von 33.3% lagen, sondern daß sie zudem auch höher waren als die Trefferquoten der naiven Beurteilung bei diesen beiden Gruppen. Dafür war die Null-Trefferrate des *GAT* bei den *Zeuginnen* deutlich geringer als die entsprechende Trefferrate der naiven Beurteilung.

Betrachtet man nur die Gesamttrefferquoten, so könnte man argumentieren, daß die Treffsicherheit des *GAT* deutlicher über der entsprechenden Zufallswahrscheinlichkeit (33.3%) lag als die der inhaltsorientierten Beurteilung (50%) und daß der *GAT* insofern die zuverlässigere Methode sei. Eine solche Argumentation wäre jedoch zu undifferenziert. Wie oben erläutert, sprechen die Befunde zu den Grundannahmen beider Ansätze dafür, daß prinzipiell nur die inhaltsorientierte Glaubhaftigkeitsbeurteilung zur direkten Überprüfung der Glaubhaftigkeit geeignet ist; hier ergaben sich signifikante Unterschiede zwischen glaubhaft und unglaubhaft aussagenden Personen. Dagegen zeigten Personen, die sich nur in bezug auf die Glaubhaftigkeit der Tatabstreitung, nicht jedoch hinsichtlich des vorhandenen Tatwissens unterschieden (*Täterinnen* vs. *Zeuginnen*), im *GAT* keine signifikant kontrastierenden elektrodermalen Reaktionsstärkemuster. Hier unterschieden sich lediglich Personen mit Tatwissen (*Täterinnen* und *Zeuginnen*) signifikant von Personen ohne Tatwissen (*falsche Zeuginnen*). Dies bedeutet, daß der *GAT*

prinzipiell nicht zur direkten Überprüfung der Glaubhaftigkeit der Täterschaftsabstreitung, sondern ausschließlich zur Diagnostik des tatbezogenen Kenntnisstands geeignet ist (woraus sich – wie beim herkömmlichen *TWT* – unter bestimmten Voraussetzungen indirekte Schlußfolgerungen bezüglich der Glaubhaftigkeit der Täterschaftsabstreitung ergeben können).

In Tabelle 30 sind die Trefferquoten aufgeführt, die sich ergaben, wenn man den *GAT* ausschließlich als diagnostisches Instrument zur Überprüfung von Tatwissen betrachtet. Rechnerisch wurden diese Werte ermittelt, indem man die Pbn im Rahmen einer Diskriminanzanalyse anhand der numerischen Scores (SCR-Quantifizierungsmethode A) als Personen mit Tatwissen vs. Personen ohne Tatwissen klassifizierte (vgl. Abschnitt 6.2.3). Es gilt zu beachten, daß in diesem Fall nur zwei diagnostische Kategorien existieren, so daß die Zufallswahrscheinlichkeit für korrekte Zuordnungen – ebenso wie bei der inhaltsorientierten und der naiven Beurteilung – 50% beträgt. In diesem Fall lag die Gesamttrefferquote des *GAT* ebenfalls deutlicher über der Zufallswahrscheinlichkeit als die Gesamttrefferquote der inhaltsorientierten Glaubhaftigkeitsbeurteilung (vgl. Tabelle 29). Zudem fällt auf, daß in diesem Fall die Treffsicherheit des *GAT* für beide Zielgruppen (Personen mit Tatwissen und Personen ohne Tatwissen) jeweils über der Zufallswahrscheinlichkeit von 50% lag, wohingegen die inhaltsorientierte Beurteilung nur bei den glaubhaften Aussagen, nicht jedoch bei den unglaubhaften Aussagen eine Hitrate über dem Zufallsniveau von 50% erzielte.

Tabelle 30. Prozentuale Trefferquoten des *GAT* hinsichtlich der Identifizierung von Personen mit und solchen ohne Tatwissen

Status der aussagenden Person		
mit Tatwissen	ohne Tatwissen	Gesamttrefferquote [a]
70.6	82.4	76.5

Anmerkung: [a] Korrektur der Basisraten von Personen mit und solchen ohne Tatwissen.

Auch dies darf jedoch nicht ohne weiteres so interpretiert werden, daß der *GAT* in seiner Eigenschaft als Instrument zur Identifizierung von Personen mit und solchen ohne Tatwissen zuverlässiger ist als die inhaltsorientierte Glaubhaftigkeitsbeurteilung. Wie schon mehrfach ausgeführt, ist die Feststellung der inhaltlichen Aussagequalität anhand der Glaubhaftigkeitskriterien zwar das zentrale, jedoch nicht das einzige Element der inhaltsorientierten Beurteilung. Normalerweise hat im Rahmen der diagnostischen Urteilsbildung eine Relativierung der festgestellten Inhaltsqualität an weiteren diagnostisch relevanten Informationen (Persönlichkeits- und Motivanalysen etc.) zu erfolgen. Den in Tabelle 29 dargestellten Trefferquoten liegen jedoch einzig und allein die inhaltsanalytischen Gesamtscores zugrunde. Insofern ist zumindest nicht auszuschließen,

daß sich bei einer alle diagnostischen Elemente umfassenden inhaltsorientierten Glaubhaftigkeitsbeurteilung höhere Hitraten ergeben hätten.

6.5 Resümee und Ausblick

Als Gesamtfazit der vorliegenden Untersuchung kann festgehalten werden, daß die diagnostische Differenzierungsfähigkeit der beiden forensischen Methoden höher war als die der naiven Glaubhaftigkeitsbeurteilung, welche nur der Zufallswahrscheinlichkeit entsprach. Entgegen der ursprünglichen diagnostischen Intention eignete sich der *GAT* jedoch nur zur Differenzierung von Personen mit und solchen ohne Tatwissen, nicht hingegen zur direkten Identifizierung solcher Personen mit Tatwissen, die den Tatvorwurf glaubhaft bzw. unglaubhaft abstritten. In der Eigenschaft als Instrument zur Überprüfung von Tatwissen erzielte der *GAT* eine Treffsicherheit, die deutlich über der Zufallswahrscheinlichkeit lag und zudem auch deutlich höher war als die Treffsicherheit der inhaltsorientierten Beurteilung bezüglich der Identifizierung glaubhafter vs. unglaubhafter Aussagen. Die Treffsicherheit der inhaltsorientierten Beurteilung lag zudem nur bei den glaubhaften, nicht jedoch bei den unglaubhaften Aussagen über dem Zufallsniveau. Da in der vorliegenden Studie die Vorgehensweise bei der inhaltsorientierten Glaubhaftigkeitsbeurteilung auf das zentrale Element, die Feststellung der inhaltlichen Aussagequalität, reduziert war, ist es denkbar, daß sich bei einer alle diagnostischen Elemente (insbesondere Relativierung der inhaltlichen Aussagequalität an aussagerelevanten Persönlichkeitseigenschaften und Vorkenntnissen) umfassenden Vorgehensweise höhere Trefferraten ergeben hätten. Allerdings ist dies eher unwahrscheinlich vor dem Hintergrund, daß die Befunde zur Gültigkeit der Grundannahme der inhaltsorientierten Glaubhaftigkeitsbeurteilung („Undeutsch-Hypothese“) in der vorliegenden Studie numerisch nur moderat ausgeprägt waren.

Es muß betont werden, daß die Resultate der vorliegenden Untersuchung keine endgültigen Schlußfolgerungen hinsichtlich des relativen diagnostischen Werts der inhaltsorientierten Glaubhaftigkeitsbeurteilung und des *GAT* zulassen. Zum einen wäre dazu zunächst eine Replikation der vorliegenden Ergebnisse im Rahmen einer Nachfolgeuntersuchung geboten. Zum anderen wäre es erforderlich, die Generalisierbarkeit der vorliegenden Befunde durch systematische Feldforschung zu überprüfen – ein Vorhaben, das jedoch angesichts des hierzulande geltenden strafrechtlichen Verwertungsverbots der psychophysiologischen Glaubhaftigkeitsbeurteilung bis auf weiteres wenig aussichtsreich erscheint.

Abgesehen von diesen grundsätzlichen Einwänden gegen eine Überbewertung der vorliegenden Befunde ist jedoch auch zu bedenken, daß die vorliegende Untersuchung in methodischer Hinsicht einige Angriffspunkte bietet. Im Hinblick auf den *GAT* wurde in erster Linie bemängelt, daß das hier verwendete Frageformat nicht vollständig mit dem Frageformat früherer Studien zum *GAT* übereinstimmte, was möglicherweise für die fehlende Differenzierung zwischen glaubhaft und unglaubhaft aussagenden Personen mit Tatwissen verantwortlich sein könnte. Bezüglich der inhaltsorientierten Glaubhaftigkeitsbeurteilung stehen insbesondere die beiden Argumente im Raum, daß das hier verwendete experimentelle Scheinverbrechen-Paradigma möglicherweise keine ausreichende externe Validität aufweist und daß die diagnostische Prozedur auf das zentrale Element, die Feststellung der inhaltlichen Aussagequalität mittels der *Kriterienorientierten Inhaltsanalyse*, reduziert war.

Im Hinblick auf die zukünftige Forschung erscheint es aussichtsreich, das hier verwendete Scheinverbrechen-Paradigma in einigen Aspekten zu optimieren. Die externe Validität der Befunde zur inhaltsorientierten Glaubhaftigkeitsbeurteilung könnte insbesondere erhöht werden, indem man das aussagerelevante Geschehen, also das Scheinverbrechenszenario, komplexer gestaltet. Ferner wäre es wichtig, neben den Aussagen selbst auch weitere diagnostisch relevanten Informationen ganz präzise zu erheben und in der diagnostischen Urteilsbildung mit zu berücksichtigen. Besonders interessant wäre es in diesem Zusammenhang, wenn der gesamte Begutachtungsprozeß von der Datenerhebung bis hin zur Urteilsbildung von professionellen Gutachtern bzw. Experten in der inhaltsorientierten Glaubhaftigkeitsbeurteilung vorgenommen würde, da man letztlich nur so der Forderung nach einer „angemessenen“ klinisch-intuitiven Urteilsbildung nachkommen kann.

Was die psychophysiologische Glaubhaftigkeitsbeurteilung mit dem *GAT* betrifft, so sollte in zukünftigen Untersuchungen darauf geachtet werden, daß das Frageformat exakt mit dem übereinstimmt, das in den Untersuchungen der Bradley-Arbeitsgruppe verwendet wurde. Darüber hinaus wäre es interessant, anstelle des *GAT* auch noch weitere neuere Verfahren direkt mit der inhaltsorientierten Methode zu vergleichen. Hier ist in erster Linie an den sog. „Directed Lie Control Question Test“ zu denken. Dieser ist eine Weiterentwicklung des Kontrollfragentests und soll einige von dessen zentralen Schwachstellen, insbesondere die mangelhafte Standardisierbarkeit und Objektivierbarkeit, verbessern.

7 Zusammenfassung

Die forensische Aussagepsychologie beschäftigt sich mit der Problematik der Aussage vor Gericht und ihrer Fehlerquellen. Die übergeordnete praktische Zielsetzung ist die Überprüfung des Realitätsgehalts von Aussagen, wobei es sich in erster Linie um Schilderungen vermeintlicher Zeugen sowie Einlassungen von Beschuldigten handelt. Grundsätzlich wird zwischen absichtlichen und irrtümlichen Aussageverfälschungen unterschieden. Der Begriff „Glaubwürdigkeit“ bezieht sich ausschließlich auf den intentionalen Aspekt der Aussage und ist gleichzusetzen mit Wahrheitsvorsatz bzw. Abwesenheit absichtlicher Täuschung. Glaubwürdigkeit ist gewährleistet, wenn ein Kommunikator an eine andere Person eine Information vermittelt, die nach Auffassung des Kommunikators zutreffend ist. Unglaubwürdigkeit bzw. Täuschung liegt dagegen vor, wenn bei einer anderen Person ein Eindruck zu erzeugen versucht wird, der nach Meinung des Kommunikators falsch ist.

Wichtig ist die Unterscheidung zwischen der „allgemeinen Glaubwürdigkeit“ einer Person im Sinne einer überdauernden Neigung, aufrichtig bzw. unehrlich zu sein, und der „speziellen Glaubhaftigkeit“ einer in Frage stehenden Aussage. Heute ist allgemein akzeptiert, daß Feststellungen über die allgemeine Glaubwürdigkeit einer Person keine zwingenden Schlußfolgerungen auf die Glaubhaftigkeit ihrer Aussage zu einem konkreten Sachverhalt zulassen. Zentraler Gegenstand forensisch-psychologischer Glaubwürdigkeitsbegutachtungen ist daher die Analyse der speziellen Glaubhaftigkeit der Aussage, d.h. die tatbestandsbezogenen Bekundungen und die damit einhergehenden Verhaltensmanifestationen stehen im Mittelpunkt.

In der forensischen Psychologie wurden verschiedene Möglichkeiten zur Glaubhaftigkeitsbeurteilung in Betracht gezogen. Während der Versuch, Merkmale der Gestik, der Mimik und des Sprechverhaltens als Glaubhaftigkeits- bzw. Täuschungsindikatoren nutzbar zu machen, bislang keine hinreichend praxisrelevanten Resultate erbrachte, gelangten insbesondere zwei methodische Ansätze zu praktischer Bedeutung: der inhaltsorientierte und der psychophysiologische. Bei der inhaltsorientierten Glaubhaftigkeitsbeurteilung wird in erster Linie analysiert, inwiefern eine Aussage bestimmte inhaltliche Merkmale aufweist, die für erlebnisbasierende Schilderungen typisch sein sollen. Der Hauptanwendungsbereich dieser Methode ist bislang die Begutachtung mutmaßlicher Opferzeugen von Sexualdelikten. Die psychophysiologische Glaubhaftigkeitsbeurteilung wird in erster Linie zur Begutachtung der Einlassungen von Beschuldigten eingesetzt. Anhand der während einer standardisierten Befragung oder Reizdarbietung gemessenen physiologischen Veränderungen beim Beschuldigten werden Rückschlüsse auf die Glaubhaftigkeit der Abstreitung des Tatvorwurfs gezogen.

Die rechtliche Situation in Deutschland unterscheidet sich für beide Methoden drastisch: Während auf dem inhaltsorientierten Ansatz basierende Glaubhaftigkeitsgutachten in deutschen Gerichtssälen schon seit mehreren Jahrzehnten als Beweismittel Anerkennung finden, wird der psychophysiologischen Glaubhaftigkeitsbeurteilung von der höchstrichterlichen Rechtsprechung kein ausreichender Beweiswert zugebilligt, woraus sich ein grundsätzliches strafrechtliches Verwertungsverbot ergibt. Dies wirft die Frage auf, ob bzw. inwiefern die psychophysiologischen und inhaltsorientierten Methoden sich in ihrem Beweiswert unterscheiden.

Ziel der vorliegenden Untersuchung war daher der direkte empirische Vergleich inhaltsorientierter und psychophysiologischer Methoden der forensischen Glaubhaftigkeitsbeurteilung. Die inhaltsorientierte Glaubhaftigkeitsbeurteilung erfolgte anhand der *Kriterienorientierten Inhaltsanalyse*, wobei 18 Glaubhaftigkeitskriterien zur Anwendung kamen. Die psychophysiologische Glaubhaftigkeitsbeurteilung erfolgte anhand des *Guilty Actions Tests* (*GAT*). Mit diesem Verfahren soll grundsätzlich differenzierbar sein, ob es sich bei einer verdächtigen Person um den Täter (Schuldigen), einen Unschuldigen mit Tatwissen oder einen Unschuldigen ohne Tatwissen handelt. Neben den beiden genannten Beurteilungsmethoden wurde als Kontrollbedingung auch noch eine naive, d.h. rein intuitive Einschätzung der Glaubhaftigkeit vorgenommen. Der Vergleich der drei diagnostischen Ansätze wurde im Rahmen einer experimentellen Simulationsstudie vorgenommen, an welcher drei Gruppen von Probandinnen (jeweils n = 34) teilnahmen. Die Probandinnen der ersten Gruppe begingen ein Scheinverbrechen (*Täterinnen* bzw. *Schuldige*). Anschließend stritten sie die Täterschaft wahrheitswidrig ab und bezichtigten in einer erfundenen „Zeugenaussage“ eine andere Person der Täterschaft. Die Probandinnen der zweiten Gruppe beobachteten das Delikt (*Zeuginnen* bzw. *Unschuldige mit Tatwissen*). Anschließend legten sie eine wahrheitsgemäße Zeugenaussage zum Tathergang ab und stritten die Täterschaft wahrheitsgemäß ab. Die Probandinnen der dritten Gruppe begingen weder das Scheinverbrechen noch beobachteten sie es. Statt dessen legten sie eine erfundene „Zeugenaussage“ zum Tathergang ab, stritten jedoch die Täterschaft wahrheitsgemäß ab (*falsche Zeuginnen* bzw. *Unschuldige ohne Tatwissen*). Die Glaubhaftigkeit der vermeintlichen Zeugenaussage wurde jeweils mit Hilfe der *Kriterienorientierten Inhaltsanalyse* beurteilt. Die Glaubhaftigkeit der Täterschaftsabstreitung beurteilte man jeweils psychophysiologisch, mit dem *GAT*. Die naive Glaubhaftigkeitsbeurteilung erfolgte, indem Personen, die keine theoretischen oder methodischen Vorkenntnisse auf dem Gebiet der Glaubhaftigkeitsbeurteilung hatten, Videos der vermeintlichen Zeugenaussagen (Bild und Ton) vorgespielt wurden, welche dann intuitiv hinsichtlich ihrer Glaubhaftigkeit einzuschätzen waren. Im Rahmen dieses Versuchsdesigns konnte somit überprüft werden, wie gut die verschiedenen Beurtei-

lungsmethoden im intendierten Sinne zwischen den experimentellen Gruppen differenzierten.

Die *Kriterienorientierte Inhaltsanalyse* differenzierte zwar erwartungskonform zwischen glaubhaften und unglaubhaften Aussagen, d.h. die Ausprägungsgrade einzelner Glaubhaftigkeitskriterien und der über alle Kriterien aufsummierte Gesamtscore waren in den erlebnisbasierenden Aussagen signifikant größer als in den erfundenen. Allerdings waren die vorgefundenen Gruppenunterschiede numerisch nur moderat ausgeprägt.

Das diagnostische Grundprinzip des *GAT* konnte nur teilweise bestätigt werden. Personen mit Tatwissen unterschieden sich in ihrem physiologischen Reaktionsstärkemuster auf die verschiedenen Testitems zwar signifikant und in der erwarteten Richtung von Personen ohne Tatwissen. Allerdings ergab sich innerhalb der Personen mit Tatwissen kein signifikanter Unterschied zwischen Schuldigen und Unschuldigen bzw. zwischen solchen Personen, die die Täterschaft unglaubhaft abstritten, und solchen, die den Tatvorwurf wahrheitsgemäß negierten.

Die Güte der naiven Glaubhaftigkeitsbeurteilung lag auf dem Zufallsniveau. Die erlebnisbasierenden und erfundenen Aussagen wurden durchschnittlich als gleich glaubhaft bzw. unglaubhaft eingestuft, so daß die individualdiagnostische Treffsicherheit sich um 50% bewegte. Allerdings deutete sich an, daß die Qualität der naiven Glaubhaftigkeitsbeurteilung auch erheblichen interindividuellen Schwankungen unterliegt.

Im Gegensatz zur naiven Glaubhaftigkeitsbeurteilung überstiegen die Gesamttrefferquoten der beiden forensisch-psychologischen Methoden jeweils die Zufallswahrscheinlichkeit. Mit der inhaltsorientierten Methode wurden allerdings nur die glaubhaften Aussagen überzufällig häufig korrekt klassifiziert, während die Treffsicherheit bezüglich der Identifizierung unglaubhafter Bekundungen dem Zufallsniveau entsprach. Mit dem *GAT* wurden sowohl die Personen mit Tatwissen als auch die Personen ohne Tatwissen überzufällig häufig als solche identifiziert, wobei die Treffsicherheit in bezug auf die Personen ohne Tatwissen höher war. Allerdings war der *GAT* völlig ungeeignet zur gezielten Identifizierung solcher Personen mit Tatwissen, die den Tatvorwurf wahrheitsgemäß abstritten. Bei Zugrundelegung von drei diagnostischen Urteilskategorien („schuldig“, „unschuldig mit Tatwissen“, „unschuldig ohne Tatwissen“) wurden die *Unschuldigen mit Tatwissen* nahezu ausschließlich fehlklassifiziert.

Die Treffsicherheit des *GAT* hinsichtlich der Unterscheidung von Personen mit und ohne Tatwissen war höher als die Treffsicherheit der inhaltsorientierten Glaubhaftig-

keitsbeurteilung in bezug auf die Identifizierung glaubhafter und unglaubhafter Schilderungen. Dies wurde jedoch nicht definitiv als Überlegenheit des *GAT* gegenüber der inhaltsorientierten Glaubhaftigkeitsbeurteilung bewertet, da bei der Interpretation der vorliegenden Untersuchungsergebnisse einige potentielle methodische Kritikpunkte in Rechnung zu stellen waren. So war, was die inhaltsorientierte Glaubhaftigkeitsbeurteilung angeht, zum einen diskussionswürdig, ob das verwendete experimentelle Scheinverbrechen-Szenario eine ausreichende externe Validität aufwies. Zum anderen wurde bemängelt, daß die diagnostische Prozedur auf das zentrale Element, die *Kriterienorientierte Inhaltsanalyse*, reduziert war. Aber auch die Resultate zum *GAT* unterschätzten möglicherweise das wirkliche diagnostische Potential dieser Methode. Diesbezüglich wurde insbesondere kritisiert, daß das in der vorliegenden Untersuchung verwendete Frageformat nicht exakt mit dem Frageformat früherer *GAT*-Studien übereinstimmte, was möglicherweise für die fehlende Differenzierung zwischen *Schuldigen* und *Unschuldigen mit Tatwissen* verantwortlich gewesen sein könnte.

8 Literaturverzeichnis

Amelang, M. & Zielinski, W. (1997). *Psychologische Diagnostik und Intervention*. Berlin: Springer.

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. (1970). *Psychologie der Zeugenaussage. Einführung in die forensische Aussagepsychologie*. Göttingen: Hogrefe.

Arntzen, F. (1983a). *Psychologie der Zeugenaussage. System der Glaubwürdigkeitsmerkmale* (2. Aufl.). München: Beck.

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

Arntzen, F. (1987). Forensische Aussagepsychologie. *Psychologische Rundschau, 38*, 165–166.

Arntzen, F. (1993). *Psychologie der Zeugenaussage. System der Glaubwürdigkeitsmerkmale* (3. Aufl.). München: Beck.

Asendorpf, J. & Wallbott, H. G. (1979). Maße der Beobachterübereinstimmung: Ein systematischer Vergleich. *Zeitschrift für Sozialpsychologie, 10*, 243–252.

Backster, C. (1962). Methods of strengthening our polygraph-technique. *Police, 6*, 61–68.

Backster, C. (1963). Polygraph professionalization through technique standardization. *Law and Order, 11*, 63–64.

Balloun, K. D. & Holmes, D. S. (1979). Effects of repeated examinations on the ability to detect guilt with a polygraphic examination: a laboratory experiment with a real crime. *Journal of Applied Psychology, 64*, 316–322.

Barland, G. H. & Raskin, D. C. (1975). An evaluation of field techniques in detection of deception. *Psychophysiology, 12*, 321–330.

Barland, G. H. & Raskin, D. C. (1976). *Validity and Reliability of Polygraph Examinations of Criminal Suspects*. Report No. 76-1, Contract 75-Ni-99-0001, National Institute of Law Enforcement and Criminal Justice, U.S. Department of Justice, Washington. (zitiert nach Ben-Shakhar & Furedy, 1990, S. 49)

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

Bender, R. (1982). Die Aussagepsychologie in der gerichtlichen Praxis. In A. Trankell (Ed.), *Reconstructing the past* (pp. 121–125). Stockholm: Norstedt and Soners.

Bender, R. & Nack, A. (1995). *Tatsachenfeststellung vor Gericht* (2. Aufl.). München: Beck.

Bender, R., Röder, S. & Nack, A. (1981). *Tatsachenfeststellung vor Gericht* (1. Aufl.). München: Beck.

Ben-Shakhar, G. (1977). A further study of the dichotomization theory in detection of information. *Psychophysiology*, *14*, 408–413.

Ben-Shakhar, G. & Furedy, J. J. (1990). *Theories and applications in the detection of deception: A psychophysiological and international perspective*. New York: Springer-Verlag.

Berlyne, D. E. (1960). *Conflict, arousal, and curiosity*. New York: McGraw-Hill.

Berning, B. R. (1992). *„Lügendetektion" aus interdisziplinärer Sicht: Eine psychologisch-juristische Abhandlung*. (Forschungsberichte aus dem Fachbereich Psychologie der Universität Osnabrück Nr. 81a, b; Band 1 und 2). Osnabrück: Selbstverlag der Universität Osnabrück.

Berning, B. R. (1993). „Lügendetektion": Eine interdisziplinäre Beurteilung. *Monatsschrift für Kriminologie, 76*, 242–255.

Bersh, P. J. (1969). A validation study of polygraph examiner judgements. *Journal of Applied Psychology*, *53*, 399–403.

Bortz, J. (1999). *Statistik für Sozialwissenschaftler*. Berlin: Springer.

Bortz, J. & Döring, N. (1995). *Forschungsmethoden und Evaluation*. Berlin: Springer.

Bortz, J., Lienert, G. A. & Boehnke, K. (1990). *Verteilungsfreie Methoden in der Biostatistik*. Berlin: Springer.

Boucsein, W. (1988). *Elektrodermale Aktivität: Grundlagen, Methoden und Anwendungen*. Berlin: Springer.

Boychuk, T. D. (1991). *Criteria-based content analysis of children´s statements about sexual abuse: A field-based validation study*. Unpublished doctoral dissertation, Arizona State University. (zitiert nach Greuel et al., 1998, S. 137)

Bradley, M. T. (1988). Choice and the detection of deception. *Perceptual and Motor Skills*, *66*, 43–48.

Bradley, M. T. & Janisse, M. P. (1981). Accuracy demonstrations, threat, and the detection of deception: Cardiovascular, electrodermal, and pupillary measures. *Psychophysiology*, *18*, 307–315.

Bradley, M. T., MacLaren, V. V. & Carle, S. B. (1996). Deception and nondeception in Guilty Knowledge and Guilty Actions Polygraph Tests. *Journal of Applied Psychology*, *81*, 153–160.

Bradley, M. T. & Rettinger, J. (1992). Awareness of crime-relevant information and the Guilty Knowledge Test. *Journal of Applied Psychology*, *77*, 55–59.

Bradley, M. T., & Warfield, J. F. (1984). Innocence, information, and the Guilty Knowledge Test in the Detection of Deception. *Psychophysiology*, *21*, 683–689.

Bundesgerichtshof. (1954). Urteil vom 16. Februar 1954 – 1 StR 578/53. *Entscheidungen des Bundesgerichtshofes in Strafsachen (BGHSt)*, *5*, 332–338.

Bundesgerichtshof. (1955). Urteil vom 14. Dezember 1954 – 5 StR 416/54. *Entscheidungen des Bundesgerichtshofes in Strafsachen (BGHSt)*, 7, 82–86;

Bundesgerichtshof. (2000a). Urteil vom 17. Dezember 1998 – 1 StR 156/98. *Entscheidungen des Bundesgerichtshofes in Strafsachen (BGHSt), 44*, 308–328.

Bundesgerichtshof. (2000b). Urteil vom 30. Juli 1999 – 1 StR 618/98. *Entscheidungen des Bundesgerichtshofes in Strafsachen (BGHSt), 45*, 164–182.

Bundesverfassungsgericht. (1981). Beschluß vom 18. August 1981 – 2 BvR 166/81. *Neue Zeitschrift für Strafrecht (NStZ), 1* (11), 446–447.

Cohen, J. (1968). Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213–220.

Cohen, J. (1972). Weighted Chi Square: An extension of the Kappa method. *Educational and Psychological Measurement, 32*, 61–74.

Coles, M. G. H., Gale, A. & Kline, P. (1971). Personality and habituation of the orienting reaction: Tonic and response measures of electrodermal activity. *Psychophysiology, 8*, 54–63.

Craig, R. A. (1995). *Effects of interviewer behaviour on children's statements of sexual abuse*. Unpublished manuscript. (zitiert nach Vrij & Akehurst, 1998, S. 18)

Craig, R. A., Scheibe, R., Raskin, D. C., Kircher, J. C. & Dodd, D. H. (1999). Interviewer questions and content analysis of children's statements of sexual abuse. *Applied Developmental Science, 3*, 77–85.

Crider, A. & Lunn, R. (1971). Electrodermal lability as a personality dimension. *Journal of Experimental Research in Personality, 5*, 145–150.

Crocker, L. M. & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando: Holt, Rinehart & Winston, Inc..

Davidson, P. O. (1968). Validity of the Guilty-Knowledge Technique: the effects of motivation. *Journal of Applied Psychology, 52*, 62–65.

Davis, R. C. (1961). Physiological responses as a means of evaluating information. In A. D. Biderman & H. Zimmer (Eds.), *The manipulation of human behavior* (pp. 142–168). New York: Wiley.

Dawson, M. E. (1980). Physiological detection of deception: Measurement of responses to questions and answers during countermeasure maneuvers. *Psychophysiology, 17*, 8–17.

DePaulo, B. M., Stone, J. I. & Lassiter, G. D. (1985). Deceiving and detecting deceit. In B. R. Schlenker (Ed.), *The self and social life* (pp. 323–370). New York: McGraw-Hill.

Dettenborn, H., Fröhlich, H.-H. & Szewczyk, H. (1984). *Forensische Psychologie. Lehrbuch der gerichtlichen Psychologie für Juristen, Kriminalisten, Psychologen, Pädagogen und Mediziner*. Berlin: VEB Deutscher Verlag der Wissenschaften.

Dickinson, J. R. & Smith, B. D. (1973). Nonspecific activity and habituation of tonic and phasic skin conductance in somatic complainers and controls as function of auditory stimulus intensity. *Journal of Abnormal Psychology, 82*, 404–413.

Döhring, E. (1964). *Die Erforschung des Sachverhalts im Prozeß*. Berlin: Duncker & Humblot.

Duprat, G. (1903). *Le mensonge: Étude de psychosociologie et pathologie*. Paris: Felix Alcan. (zitiert nach Furedy, 1986, S. 684)

Edelberg, R. (1967). Electrical properties of the skin. In C. C. Brown (Ed.), *Methods in psychophysiology* (pp. 1–53). Baltimore, MD: Williams and Wilkins.

Eisenberg, U. (1993). *Persönliche Beweismittel in der StPO*. München: Beck.

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

Elaad, E., Ginton, A. & Jungman, N. (1992). Detection measures in real-life criminal guilty knowledge tests. *Journal of Applied Psychology*, *77*, 757–767.

Esplin, P. W., Houed, T. & Raskin, D. C. (1988). *Applications of statement validity assessment*. Paper presented at the NATO Advanced Study Institute on Credibility Assessment, Maratea, Italy. (zitiert nach Raskin & Esplin, 1991a, S. 160ff.)

Fabian, T., Greuel, L. & Stadler, M. (1996). Möglichkeiten und Grenzen aussagepsychologischer Glaubwürdigkeitsbegutachtung. *Strafverteidiger, 16*, 347–351.

Fabian, T. & Wetzels, P. (1991). Zur gegenwärtigen Praxis von forensischen Psychologen und Psychologinnen. *Praxis der Forensischen Psychologie, 1*, 10–17.

Fahrenberg, J. (1969). Die Bedeutung individueller Unterschiede für die Methodik der Aktivierungsforschung. In W. Schönpflug (Hrsg.), *Methoden der Aktivierungsforschung* (S. 95–121). Bern: Huber.

Finn, R. H. (1970). A note on estimating the reliability of categorical data. *Educational and Psychological Measurement, 30*, 71–76.

Fischer, T. (1994). Glaubwürdigkeitsbeurteilung und Beweiswürdigung. *Neue Zeitschrift für Strafrecht, 14*, 1–5.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed). New York: Wiley.

Frick, T. & Semmel, M. I. (1978). Observer agreement and reliabilities of classroom observational measures. *Review of Educational Research, 48*, 157–184.

Fröhlich, W. D. (2000). *Wörterbuch zur Psychologie* (23. Aufl.). München: Deutscher Taschenbuch Verlag.

Furedy, J. J. (1986). Lie detection as psychophysiological differentiation: Some fine lines. In M. Coles, E. Donchin & S. W. Porges (Eds.), *Psychophysiology: Systems, processes, and applications – A handbook* (pp. 683–701). New York: Guilford.

Furedy, J. J. (1993). The ’control‘ question ’test‘ (CQT) polygrapher‘s dilemma: Logico-ethical considerations for psychophysiological practitioners and researchers. *International Journal of Psychophysiology*, *15*, 263–267.

Furedy, J. J. (1996). The North American polygraph and psychophysiology: Disinterested, uninterested, and interested perspectives. *International Journal of Psychophysiology, 21*, 97–105.

Furedy, J. J., Davis, C. & Gurevich, M. (1988). Differentiation of deception as a psychological process: A psychophysiological approach. *Psychophysiology, 25*, 683–688.

Furedy, J. J. & Heslegrave, R. J. (1991). The forensic use of the polygraph: A psychophysiological analysis of current trends and future prospects. In J. R. Jennings, P. K. Ackles & M. G. H. Coles (Eds.), *Advances in psychophysiology* (Vol. 4, pp. 157–189). London: Jessica Kingsley Publishers.

Giesen, M. & Rollison, M. A. (1980). Guilty knowledge versus innocent associations: effects of trait anxiety and stimulus context on skin conductance. *Journal of Research in Personality, 14*, 1–11.

Gödert, H. W., Rill, H.-G. & Vossel, G. (2001). Psychophysiological differentiation of deception: The effects of electrodermal lability and mode of responding on skin conductance and heart rate. *International Journal of Psychophysiology, 40*, 61–75.

Greuel, L., Offe, S., Fabian, A., Wetzels, P., Fabian, T., Offe, H. & Stadler, M. (1998). *Glaubhaftigkeit der Zeugenaussage*. Weinheim: Psychologie Verlags Union.

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

Hershkowitz, I., Lamb, M. E., Sternberg, K. J. & Esplin, P. W. (1997). The relationships among interviewer utterance type, CBCA scores and the richness of children´s responses. *Legal and Criminological Psychology, 2*, 169–176.

Höfer, E., Krause, S., Petersen, R., Sievers, K. & Köhnken, G. (1999). *Zur Reliabilität der Realkennzeichen in der Glaubwürdigkeitsbegutachtung*. 8. Arbeitstagung der Fachgruppe Rechtspsychologie der Deutschen Gesellschaft für Psychologie. Universität Erlangen-Nürnberg, Nürnberg. (Abstract)

Honts, C. R. (1994). Psychophysiological detection of deception. *Current Directions in Psychological Science, 3*, 77–82.

Honts, C. R. (1996). Criterion development and validity of the CQT in field application. *Journal of General Psychology, 123*, 309–324.

Honts, C. R., Hodes, R. L. & Raskin, D. C. (1985). Effects of physical countermeasures on the physiological detection of deception. *Journal of Applied Psychology, 70*, 177–187.

Honts, C. R. & Raskin, D. C. (1988). A field study of the validity of the directed lie control question. *Journal of Police Science and Administration, 16*, 56–61.

Honts, C. R., Raskin, D. C. & Kircher, J. C. (1987). Effects of physical countermeasures and their electromyographic detection during polygraph tests for deception. *Journal of Psychophysiology, 1*, 241–247.

Horneman, C. J. & O'Gorman, J. G. (1987). Individual differences in psychophysiological responsiveness in laboratory tests of deception. *Personality and Individual Differences, 8,* 321–330.

Horowitz, S. W., Kircher, J. C., Honts, C. R. & Raskin, D. C. (1997). The role of comparison questions in physiological detection of deception. *Psychophysiology, 34,* 108–115.

Horowitz, S. W., Lamb, M. E., Esplin, P. W., Boychuk, T. D., Krispin, O. & Reiter-Lavery, L. (1997). Reliability of criteria-based content analysis of child witness statements. *Legal and Criminological Psychology, 2,* 11–21.

Horvath, F. (1977). The effect of selected variables on interpretation of polygraph records. *Journal of Applied Psychology, 62,* 127–136.

Huffman, M. L. & Ceci, S. J. (1997). *Can criteria-based content analysis distinguish true and false beliefs of preschoolers? An exploratory analysis*. Unpublished manuscript, Cornell University, Ithaca, NY. (zitiert nach Ruby & Brigham, 1997, S. 718)

Iseler, A. (1967). Zur varianzanalytischen Schätzung der Auswertungsobjektivität von psychologischen Tests. *Diagnostica, 13,* 135–148.

Joffe, R. & Yuille, J. C. (1992). *Criteria-based content analysis: An experimental investigation*. Paper presented at the NATO Advanced Study Institute, „The Child Witness in Context: Cognitive, Social, and Legal Perspectives", Il Ciocco, Italy. (zitiert nach Greuel et al., 1998, S. 139f.)

Kainz, F. (1955). Gerichtliche Sprachpsychologie. *Sprachforum, 1,* 20–33.

Kasielke, E. (1965). Psychologische Begutachtung der Glaubwürdigkeit kindlicher und jugendlicher Zeugen. In H.-D. Schmidt & E. Kasielke (Hrsg.), *Psychologie und Rechtspraxis* (S. 87–105). Berlin: VEB Deutscher Verlag der Wissenschaften.

Katkin, E. S. & McCubbin, R. J. (1969). Habituation of the orienting response as a function of individual differences in anxiety and autonomic lability. *Journal of Abnormal Psychology, 74,* 54–60.

Kircher, J. C. & Raskin, D. C. (1988). Human versus computerized evaluation of polygraph data in a laboratory setting. *Journal of Applied Psychology, 73,* 291–302.

Köhnken, G. (1986). Verhaltenskorrelate von Täuschung und Wahrheit – Neue Perspektiven in der Glaubwürdigkeitsdiagnostik. *Psychologische Rundschau, 37,* 177–194.

Köhnken, G. (1987). Forschungsaufgaben in der Glaubwürdigkeitsdiagnostik. *Psychologische Rundschau, 38,* 167–168.

Köhnken, G. (1990). *Glaubwürdigkeit: Untersuchungen zu einem psychologischen Konstrukt*. München: Psychologie Verlags Union.

Köhnken, G., Schimossek, E., Aschermann, E. & Höfer, E. (1995). The cognitive interview and the assessment of the credibility of adults' statements. *Journal of Applied Psychology, 80,* 671–684.

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

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

Krahé, B. & Kundrotas, S. (1992). Glaubwürdigkeitsbeurteilung bei Vergewaltigungsanzeigen: Ein aussagenanalytisches Feldexperiment. *Zeitschrift für Experimentelle und Angewandte Psychologie, 39*, 598–620.

Krause, S. (1997). *Konzeption und Evaluation der Validität des Kieler Trainingsprogrammes zur Beurteilung der Glaubwürdigkeit von Zeugenaussagen (KTBG).* Unveröffentlichte Diplomarbeit im Fach Psychologie. Universität Kiel.

Krause, S. & Petersen, R. (keine Jahresangabe). *Konzeption und Evaluation der Reliabilität und Validität des Kieler Trainingsprogrammes zur Beurteilung der Glaubwürdigkeit von Zeugenaussagen (KTBG).* Unveröffentlichtes Manuskript, Institut für Psychologie der Universität Kiel.

Lamb, M. E. (1998). Assessment of children's credibility in forensic contexts. *Current Directions in Psychological Science, 7*, 43–46.

Lamb, M. E., Sternberg, K. J., Esplin, P. W. Hershkowitz, I. & Orbach, Y. (1997). Assessing the credibility of children´s allegations of sexual abuse: A survey of recent research. *Learning and Individual Differences, 9*, 175–194.

Lamb, M. E., Sternberg, K. J., Esplin, P. W., Hershkowitz, I., Orbach, Y. & Hovav, M. (1997). Criterion-based content analysis: A field validation study. *Child Abuse and Neglect, 21*, 255–264.

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Landry, K. L. & Brigham, J. C. (1992). The effect of training in criteria-based content analysis on the ability of detect deception in adults. *Law and Human Behavior, 16*, 663–676.

Lang, P. J., (1979). A bio-informational theory of emotional imagery. *Psychophysiology, 16*, 495–512.

Lehrl, S. (1977). *Mehrfachwahl-Wortschatz-Intelligenztest MWT-B.* Erlangen: Straube.

Leonhardt, C. (1930). Psychologische Beweisführung in Ansehung existenzstreitiger Vorgänge. *Archiv für die gesamte Psychologie, 75*, 545–558.

Leonhardt, C. (1934). Psychologische Beweisführung in Ansehung existenzstreitiger Vorgänge. *Zeitschrift für angewandte Psychologie, 46*, 229–245.

Lieblich, I., Kugelmass, S. & Ben-Shakhar, G. (1970). Efficiency of GSR detection of information as a function of stimulus set size. *Psychophysiology, 6*, 601–608.

Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Aufl.). Weinheim: Beltz.

Littmann, E. & Szewczyk, H. (1983). Zu einigen Kriterien und Ergebnissen forensisch-psychologischer Glaubwürdigkeitsbegutachtung von sexuell mißbrauchten Kindern und Jugendlichen. *Forensia, 4,* 55–72.

Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology, 43,* 385–388.

Lykken, D. T. (1960). The validity of the guilty knowledge technique: the effects of faking. *Journal of Applied Psychology, 43,* 385–388.

Lykken, D. T. (1974). Psychology and the lie detection industry. *American Psychologist, 29,* 725–739.

Lykken, D. T. (1979). The detection of deception. *Psychological Bulletin, 86,* 47–53.

Lykken, D. T. (1988). The case against polygraph testing. In a. Gale (Ed.), *The Polygraph Test: Lies, Truth and Science* (pp. 111–126). London: Sage.

Lykken, D. T. (1991). Why (some) Americans believe in the lie detector while others believe in the guilty knowledge test. *Integrative Physiological and Behavioral Science, 26,* 214–222.

Lykken, D. T. (1998). *A tremor in the blood: Uses and abuses of the lie detector* (2nd ed.). New York: Plenum Press.

Lykken, D. T. & Venables, P. H. (1971). Direct measurement of skin conductance: A proposal for standardization. *Psychophysiology, 8,* 656–672.

Matté, J. A. (1996). *Forensic psychophysiology using the polygraph*. Williamsville, N. Y.: J. A. M. Publications.

Michel, L. (1965). Die numerische Bestimmung der Auswertungsobjektivität von psychologischen Tests. *Diagnostica, 11,* 158–172.

Nack, A. (1982). Die Anwendung einer allgemeinen Beweislehre auf die Würdigung von Zeugenaussagen. In A. Trankell (Ed.), *Reconstructing the past* (pp. 127–130). Stockholm: Norstedt and Soners.

Offe, H. & Offe, S. (1994). Anforderungen an die Begutachtung der Glaubwürdigkeit von Zeugenaussagen bei Verdacht des sexuellen Mißbrauchs. *Praxis der Rechtspsychologie, 4,* 24–37.

O'Toole, D., Yuille, J. C., Patrick, C. J. & Iacono, W. G. (1994). Alcohol and the physiological detection of deception: Arousal and memory influences. *Psychophysiology, 31,* 253–263.

Patrick, C. J. & Iacono, W. G. (1989). Psychopathy, threat, and polygraph test accuracy. *Journal of Applied Psychology, 74,* 347–355.

Patrick, C. J. & Iacono, W. G. (1991). Validity of the Control Question Polygraph Test: The problem of sampling bias. *Journal of Applied Psychology, 76,* 229–238.

Petersen, R. (1997). *Konzeption und Evaluation der Reliabilität des Kieler Trainingsprogrammes zur Beurteilung der Glaubwürdigkeit von Zeugenaussagen (KTBG).* Unveröffentlichte Diplomarbeit im Fach Psychologie. Universität Kiel.

Podlesny, J. A. & Raskin, D. C. (1977). Physiological measures and the detection of deception. *Psychological Bulletin, 84*, 782–799.

Podlesny, J. A. & Raskin, D. C. (1978). Effectiveness of techniques and physiological measures in the detection of deception. *Psychophysiology, 15*, 344–359.

Porter, S. & Yuille, J. C. (1995). Credibility assessment of criminal suspects through statement analysis. *Psychology, Crime and Law, 1*, 319–331.

Porter, S. & Yuille, J. C. (1996). The language of deceit: An investigation of the verbal clues in the interrogation context. *Law and Human Behavior, 20*, 443–458.

Prüfer, H. (1986). *Aussagebewertung in Strafsachen*. Köln: Heymanns.

Rafky, D. M. & Sussman, R. C. (1985). Polygraphic reliability and validity: individual components and stress of issue in criminal tests. *Journal of Police Science and Administration, 13*, 283–294.

Raskin, D. C. (1979). Orienting and defense reflexes in the detection of deception. In H. D. Kimmel, E. H. Van Olst & J. F. Orlebeke (Eds.), *The Orienting Reflex in Humans* (pp. 587–605). Hillsdale, N. J.: Erlbaum.

Raskin, D. C. (1982). The scientific basis of polygraph techniques and their uses in the judicial process. In A. Trankell (Ed.), *Reconstructing the past: The role of psychologists in criminal trials* (pp. 317–371). Stockholm: Norstedt.

Raskin, D. C. (1986). The polygraph in 1986: Scientific, professional and legal issues surrounding application and acceptance of polygraph evidence. *Utah Law Review, 29*, 29–74.

Raskin, D. C. (1989). Polygraph techniques for the detection of deception. In D. C. Raskin (Ed.), *Psychological methods in criminal investigation and evidence* (pp. 247–295). New York: Springer.

Raskin, D. C. & Esplin, P. W. (1991a). 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). Wachington, DC: American Psychological Association.

Raskin, D. C. & Esplin, P. W. (1991b). Statement validity assessment: Interview procedures and content analysis of children's statements of sexual abuse. *Behavioral Assessment, 13*, 265–291.

Raskin, D. C. & Hare, R. D. (1978). Psychopathy and detection of deception in a prison population. *Psychophysiology, 15*, 126–139.

Raskin D. C. & Kircher, J. C. (1989). Comments on Furedy and Heslegrave: Misconceptions, misdescriptions, and misdirections. In J. R. Jennings, P. K. Ackles & M. G. H. Coles (Eds.), *Advances in psychophysiology* (Vol. 4, pp. 215–223). London: Jessica Kingsley Publishers.

Raskin, D. C. & Yuille, J. C. (1989). Problems in evaluating interviews of children in sexual abuse cases. In S. J. Ceci, M. P. Toglia & D. F. Ross (Eds.), *New perspectives on the child witness* (pp. 184–207). New York: Springer.

Reid, J. E. (1947). A revised questioning technique in lie-detection tests. *Journal of Criminal Law and Criminology, 37*, 542–547.

Reid, J. E. & Arther, R. O. (1953). Behavior symptoms of lie detector subjects. *Journal of Criminal Law and Criminology*, *44*, 104–108.

Reid, J. E. & Inbau, F. E. (1977). *Truth and deception. The polygraph ("lie-detector") technique* (2nd ed.). Baltimore, MD: Williams and Wilkins.

Rill, H.-G. & Vossel, G. (1998). Psychophysiologische Täterschaftsbeurteilung („Lügendetektion", „Polygraphie"): Eine kritische Analyse aus psychophysiologischer und psychodiagnostischer Sicht. *Neue Zeitschrift für Strafrecht,18*, 481–486.

Rovner, L. I., Raskin, D. C. & Kircher, J. C. (1979). Effects of information and practice on detection of deception. *Psychophysiology, 16*, 197–198.

Ruby, C. L. & Brigham, J. C. (1997). The usefulness of the criteria-based content analysis technique in distinguishing between truthful and fabricated allegations: A critical review. *Psychology, Public Policy, and Law*, *3*, 705–737.

Ruby, C. L. & Brigham, J. C. (1998). Can criteria-based content analysis distinguish between true and false statements of African-American speakers? *Law and Human Behavior, 22*, 369–388.

Rüßmann, H. (1997). Fragen der Umsetzung und Handhabung von Glaubwürdigkeitskriterien im Gerichtssaal. In L. Greuel, T. Fabian & M. Stadler (Hrsg.), *Psychologie der Zeugenaussage* (S. 151–160). Weinheim: Psychologie Verlags Union.

Rüth-Bemelmanns, E. (1984). *Experimentelle Erprobung der Kriterien zur Aussageanalyse*. Universität Köln: Unveröffentlichte Diplomarbeit.

Schandry, R. (1989). *Lehrbuch der Psychophysiologie: Körperliche Indikatoren psychischen Geschehens* (2. Aufl.). Weinheim: Psychologie Verlags Union.

Scheinberger, R. (1993). *Inhaltliche Realkennzeichen in Aussagen von Erwachsenen: Eine Simulationsstudie zur wissenschaftlichen Evaluation der Kriterienorientierten Aussageanalyse*. Unveröffentlichte Diplomarbeit, Freie Universität Berlin.

Schell, A. M., Dawson, M. E. & Filion, D. L. (1988). Psychophysiological correlates of electrodermal lability. *Psychophysiology*, *25*, 619–632.

Scherer, T. K. (1998). StPO §136a: Polygraphentests zukünftig verwertbar? (Kommentar zum Beschluß des BVerfG vom 15.10.1997 – 2 BVR 1211/97). *Strafverteidiger-Forum* (keine Bandangabe), 16–17.

Schlothauer, R. (1997). Strafprozessuale Fragen der Glaubwürdigkeitsbegutachtung von Zeugen durch psychologische Sachverständige. In L. Greuel, T. Fabian & M. Stadler (Hrsg.), *Psychologie der Zeugenaussage* (S. 143–149). Weinheim: Psychologie Verlags Union.

Schneider, E. (1987). *Beweis und Beweiswürdigung* (4. Aufl.). München: Franz Vahlen.

Siddle, D. A. T., O'Gorman, J. G. & Wood, L. (1979). Effects of electrodermal lability and stimulus significance on electrodermal response amplitude to stimulus change. *Psychophysiology*, *16*, 520–527.

Sokolov, E. N. (1963). *Perception and the conditioned reflex*. New York: MacMillan.

Stadler, M. (1997). Realitätskriterien und Wirklichkeitskriterien. In L. Greuel, T. Fabian & M. Stadler (Hrsg.), *Psychologie der Zeugenaussage* (S. 59–70). Weinheim: Psychologie Verlags Union.

Steller, M. (1984). *Extraversion, elektrodermale Labilität und psychophysiologische Täterschaftsermittlung*. Vortrag auf der 26. Tagung experimentell arbeitender Psychologen in Nürnberg 1984. (zitiert nach Steller, 1987, S. 33ff.)

Steller, M. (1987). *Psychophysiologische Aussagebeurteilung*. Göttingen: Hogrefe.

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

Steller, M. (1997). Psychophysiologische Täterschaftsermittlung („Lügendetektion“, „Polygraphie“). In M. Steller & R. Volbert (Hrsg.), *Psychologie im Strafverfahren* (S. 89–104). Bern: Huber.

Steller, M. & Boychuk, T. (1992). Children as witnesses in sexual abuse cases: Investigative interview and assessment techniques. In H. Dent & R. Flin (Eds.), *Children as witnesses* (pp. 47–71). Chichester: Wiley.

Steller, M. & Dahle, K.-P. (1997). Psychophysiologische Täterschaftsbeurteilung („Lügendetektion“): Unschuldsnachweis bei Verdacht auf sexuellen Mißbrauch? In L. Greuel, T. Fabian & M. Stadler (Hrsg.), *Psychologie der Zeugenaussage* (S. 309–323). Weinheim: Psychologie Verlags Union.

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 in criminal investigation and evidence* (pp. 217–245). New York: Springer.

Steller, M. & Volbert, R. (1997). Glaubwürdigkeitsbegutachtung. In Steller & Volbert (Hrsg.), *Psychologie im Strafverfahren* (S. 12–39). Bern: Huber.

Steller, M., Volbert, R. & Wellershaus, P. (1993). Zur Beurteilung von Zeugenaussagen. Aussagepsychologische Konstrukte und methodische Strategien. In L. Montada (Hrsg.), *Bericht über den 38. Kongreß der Deutschen Gesellschaft für Psychologie in Trier 1992, Band 2* (S. 367–376). Göttingen. Hogrefe.

Steller, M., Wellershaus, P. & Wolf, T. (1988). *Empirical validation of criteria-based content analysis*. Paper presented at the NATO Advanced Study Institute on Credibility Assessment, Maratea, Italy. (zitiert nach Steller, 1989, S. 144ff.)

Steller, M., Wellershaus, P. & Wolf, T. (1992). Realkennzeichen in Kinderaussagen: Empirische Grundlagen der kriterienorientierten Aussageanalyse. *Zeitschrift für experimentelle und angewandte Psychologie*, *39*, 151–179.

Stern, R. M., Breen, J. P., Watanabe, T. & Perry, B. S. (1981). Effect of feedback of physiological information on responses to innocent associations and guilty knowledge. *Journal of Applied Psychology*, *66*, 677–681.

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

Tedeschi, J. T. & Norman, N. (1985). Social power, self-presentation, and the self. In B. R. Schlenker (Ed.), *The self and social life* (pp. 293–322). New York: McGraw-Hill.

Tent, L. (1967). Psychologische Tatbestandsdiagnostik (Spurensymptomatologie, Lügendetektion). In U. Undeutsch (Hrsg.), *Handbuch der Psychologie: Band 11 Forensische Psychologie* (S. 185–259). Göttingen: Hogrefe.

Trankell, A. (1963). *Vittnespsykologins arbetsmetoder* (Aussagepsychologische Arbeitsmethoden). Stockholm: Bokförlaget Liber. (zitiert nach Undeutsch, 1967, S. 126)

Trankell, A. (1971). *Der Realitätsgehalt von Zeugenaussagen*. Göttingen: Vandenhoeck & Ruprecht.

Undeutsch, U. (1954). Die Entwicklung der gerichtspsychologischen Gutachtertätigkeit. In A. Wellek (Hrsg.), *Bericht über den 19. Kongreß der Deutschen Gesellschaft für Psychologie in Köln 1953* (S. 132–154). Göttingen: Hogrefe.

Undeutsch, U. (1956). Eine grundsätzliche Entscheidung des BGH über die Zuziehung von Sachverständigen zur Beurteilung von Aussagen Minderjähriger. *Praxis der Kinderpsychologie, 5,* 67–69.

Undeutsch, U. (1967). Beurteilung der Glaubhaftigkeit von Zeugenaussagen. In U. Undeutsch (Hrsg.), *Handbuch der Psychologie. Band 11: Forensische Psychologie* (S. 26–181). Göttingen: Hogrefe.

Undeutsch, U. (1982). Statement reality analysis. In A. Trankell (Ed.), *Reconstructing the past: the role of psychologists in criminal trials* (pp. 27–56). Stockholm: Norstedt and Soners.

Undeutsch, U. (1983a). Die psychophysiologische Täterschaftsermittlung. In F. Lösel (Hrsg.), *Kriminal-Psychologie: Grundlagen und Anwendungsbereiche* (S. 191–206). Weinheim: Beltz.

Undeutsch, U. (1983b). Vernehmung und non-verbale Information. In E. Kube, H. U. Störzer & S. Brugger (Hrsg.), *Wissenschaftliche Kriminalistik* (S. 389–418). Wiesbaden: Bundeskriminalamt.

Undeutsch, U. (1984). Courtroom evaluation of eyewitness testimony. *International Review of Applied Psychology, 33,* 51–67.

Undeutsch, U. (1989). The development of statement reality analysis. In J. Yuille (Ed.), *Credibility assessment* (pp. 101–120). Dordrecht: Kluwer.

Undeutsch, U. (1996). Die Untersuchung mit dem Polygraphen („Lügendetektor“) – eine wissenschaftliche Methode zum Nachweis der Unschuld. *Zeitschrift für Familienrecht, 6,* 329–331.

Undeutsch, U. (1997). Psychophysiologische Täterschaftsdiagnostik: Bedarf und Akzeptanz, insbesondere bei Verdacht des sexuellen Mißbrauchs. In L. Greuel, T. Fabian & M. Stadler (Hrsg.), *Psychologie der Zeugenaussage* (S. 303–308). Weinheim: Psychologie Verlags Union.

Venables, P. H. & Christie, M. J. (1980). Electrodermal activity. In I. Martin & P. H. Venables (Eds.), *Techniques in psychophysiology* (pp. 3–67). Chichester: Wiley.

Vossel, G. (1990). *Elektrodermale Labilität: Ein Beitrag zur differentiellen Psychophysiologie*. Göttingen: Hogrefe.

Vossel, G. & Roßmann, R. (1984). Electrodermal habituation speed and visual monitoring performance. *Psychophysiology, 21*, 97–100.

Vossel, G. & Zimmer, H. (1990). Psychometric properties of non-specific electrodermal response frequency for a sample of male students. *International Journal of Psychophysiology, 10*, 69–73.

Vossel, G. & Zimmer, H. (1998). *Psychophysiologie*. Stuttgart: Kohlhammer.

Vrij, A. (1998a). Nonverbal communication and credibility. In A. Memon, A. Vrij & R. Bull (Eds.), *Psychology and law: Truthfulness, accuracy, and credibility* (pp. 32–58). London: McGraw-Hill.

Vrij, A. (1998b). Physiological parameters and credibility: the polygraph. In A. Memon, A. Vrij & R. Bull (Eds.), *Psychology and law: Truthfulness, accuracy, and credibility* (pp. 77–101). London: McGraw-Hill.

Vrij, A. & Akehurst, L. (1998). Verbal communication and credibility: Statement validity assessment. In A. Memon, A. Vrij & R. Bull (Eds.), *Psychology and law: Truthfulness, accuracy, and credibility* (pp. 3–31). London: McGraw-Hill.

Vrij, A., Kneller, W. & Mann, S. (2000). The effect of informing liars about Criteria-Based Content Analysis on their ability to deceive CBCA-raters. *Legal and Criminological Psychology, 5*, 57–70.

Waid, W. M. & Orne M. T. (1980). Individual differences in electrodermal lability and the detection of information and deception. *Journal of Applied Psychology, 65*, 1–8.

Waid, W. M., Wilson, S. K. & Orne M. T. (1981). Cross-modal physiological effects of electrodermal lability in the detection of deception. *Journal of Personality and Social Psychology, 40*, 1118–1125.

Walschburger, P. (1975). Zur Standardisierung und Interpretation elektrodermaler Meßwerte in psychologischen Experimenten. *Zeitschrift für Experimentelle und Angewandte Psychologie, 22*, 514–533.

Walschburger, P. (1976). *Zur Beschreibung von Aktivierungsprozessen. Eine Methodenstudie zur psychophysiologischen Diagnostik*. Unveröffentlichte Dissertation, Universität Freiburg, Freiburg.

Wegener, H. (1992). *Einführung in die Forensische Psychologie* (2. Aufl.). Darmstadt: Wissenschaftliche Buchgesellschaft.

Wellershaus, P. (1992). Glaubhaftigkeit kindlicher Zeugenaussagen. *Psychomed, 4*, 20–24.

Wellershaus, P. & Wolf, T. (1989). *Kriterienorientierte Aussageanalyse. Empirische Überprüfung der diagnostischen Leistungsfähigkeit inhaltlicher Realkennzeichen zur Trennung wahrer von unwahren Zeugenaussagen*. Unveröffentlichte Diplomarbeit im Fach Psychologie. Universität Kiel. (zitiert nach Krause & Petersen, keine Jahresangabe, S. 22)

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). Wachington, DC: American Psychological Association.

Wetzels, P. (1990). Psychologische Glaubwürdigkeitsbegutachtung in Zivilverfahren: Ergebnisse einer Sachverständigenbefragung. *Rundbrief der Sektion Forensische und Kriminalpsychologie im BDP, 2*, 13–19.

Winkel, J. C. & Vrij, A. (1995). Verklaringen van kinderen in interviews: een experimenteel onderzoek naar de diagnostische waarde van Criteria Based Content Analysis. (Aussagen von Kindern in Vernehmungen: Eine experimentelle Untersuchung zum diagnostischen Wert der Kriterienorientierten Inhaltsanalyse). *Tijdschrift voor Ontwikkelingspsychologie, 22*, 61–74.

Wolf, P. & Steller, M. (1997). Realkennzeichen in Aussagen von Frauen. Zur Validierung der Kriterienorientierten Aussageanalyse für Zeugenaussagen von Vergewaltigungsopfern. In L. Greuel, T. Fabian & M. Stadler (Hrsg.), *Psychologie der Zeugenaussage* (S. 121–130). Weinheim: Psychologie Verlags Union.

Yuille, J. C. (1988). The systematic assessment of children´s testimony. *Canadian Psychology, 29*, 247–262.

Yuille, J. C. (1990). *Use of criteria-based content analysis*. Unpublished manuscript, University of British Columbia, Vancouver. (zitiert nach Zaparniuk, Yuille & Taylor, 1995, S. 345)

Yuille, J. C. & Cutshall, J. (1989). Analysis of the statements of victims, witnesses, and suspects. In J. Yuille (Ed.), *Credibility assessment* (pp. 175–191). Dordrecht: Kluwer.

Zaparniuk, J., Yuille, J. C. & Taylor, S. (1995). Assessing the credibility of true and false statements. *International Journal of Law and Psychiatry, 18*, 343–352.

Zimmer, H., Vossel, G. & Fröhlich, W. D. (1990). Individual differences in resting heart rate and spontaneous electrodermal activity as predictors of attentional processes: Effects on anticipatory heart rate deceleration and task performance. *International Journal of Psychophysiology, 8*, 249–259.

Zuckerman, M., DePaulo, B. M. & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 14, pp. 1–59). New York: Academic Press.

Anhang

Anhang A: Schriftliche Instruktionen und Utensilien für die Versuchspersonen[a]

Anhang A.1: Informationsblatt bezüglich Untersuchungsbedingungen und Einverständniserklärung

Anhang A.2: Instruktion 1 für die *Täterinnen*

Anhang A.3: Instruktion 2 für die *Täterinnen*

Anhang A.4: Instruktion 3 für die *Täterinnen*

Anhang A.5: Instruktion 4 für die *Täterinnen*

Anhang A.6: Instruktion 1 für die *Zeuginnen*

Anhang A.7: Instruktion 2 für die *Zeuginnen*

Anhang A.8: Instruktion 3 für die *Zeuginnen*

Anhang A.9: Instruktion 4 für die *Zeuginnen*

Anhang A.10: Instruktion 1 für die *falschen Zeuginnen*

Anhang A.11: Instruktion 2 für die *falschen Zeuginnen*

Anhang A.12: Wegskizze

Anhang A.13: Aufgabenzettel für die Reinigungskraft (Version für Verbrechenssimulation mit *Täterinnen* und *Zeuginnen*)

Anhang A.14: Aufgabenzettel für die Reinigungskraft (Version für Bedingung *falsche Zeuginnen*)

Anhang A.15: Versteckte Notizzettel mit den Ziffern des Zahlenschloßcodes

Anhang A.16: Notizzettel für die *Täterinnen*

Anhang A.17: *GAT*-Instruktion 1

Anhang A.18: *GAT*-Instruktion 2

Anhang A.19: *GAT*-Instruktion 3

Anhang A.20: *GAT*-Instruktion 4

Anhang A.21: Nachbefragungsbogen[b]

---

[a] Aus drucktechnischen Gründen wurde die Schriftgröße der Instruktionen etc. für den Anhang verringert.

[b] In der Bedingung *falsche Zeuginnen* war der Wiedererkennungstest im Nachbefragungsbogen leicht modifiziert (s. genauer Abschnitt 4.3.3).

Anhang A.1: Informationsblatt bezüglich Untersuchungsbedingungen und Einverständniserklärung

JOHANNES GUTENBERG-UNIVERSITÄT MAINZ
Psychologisches Institut
Abteilung Allgemeine Experimentelle Psychologie

Dipl.-Psych. Heinz Werner Gödert

Staudingerweg 9 . D-55099 Mainz . Telefon 06131 / 39-2422 . Fax 39-2480 . Email goedert@psych.uni-mainz.de

Experiment zur Glaubwürdigkeitsbegutachtung / Lügendetektion

Zu Ihrer Information

Alle Daten dieser Untersuchung werden anonym erhoben und gemäß den **Richtlinien des Datenschutzes** vertraulich behandelt. Ihr Name wird unabhängig von den restlichen Daten erfaßt, so daß diese keine Rückschlüsse auf Einzelpersonen zulassen.

- Für die Teilnahme am Experiment erhalten Sie am Ende eine **Grundvergütung von 10,- DM**.
- Wenn Sie bei der **Glaubwürdigkeitsbegutachtung** / dem **Lügendetektortest** als „**glaubwürdig**" bzw. „**unschuldig**" eingestuft werden, erhalten Sie eine **zusätzliche Belohnung von 15,- DM**.
- Außerdem haben Sie in einer abschließenden **Nachbefragung** zum Experiment die Möglichkeit, eine **weitere Zusatzbelohnung von bis zu 5,- DM** zu erhalten.
  (d.h. maximale Belohnung: 30,- DM)

Nach dem Experiment werden Sie **vollständig** über die Hintergründe der Untersuchung **aufgeklärt**.

Da wir für alle Teilnehmer vergleichbare Bedingungen schaffen müssen, dürfen zukünftige Versuchspersonen keine detaillierten Vorinformationen besitzen. Darum möchten wir Sie um **Verschwiegenheit** hinsichtlich der Untersuchung bitten. Falls Freunde oder Bekannte von Ihnen auch teilnehmen möchten, dann können Sie sie gerne an uns verweisen, ohne jedoch nähere Angaben zur Untersuchung zu machen.

Falls Sie Fragen haben, wenden Sie sich bitte an den Versuchsleiter.

Einverständniserklärung:

Hiermit bestätige ich, daß ich **freiwillig** an dieser Studie teilnehme und mir dessen bewußt bin, daß ich **jederzeit** - auch im Verlauf der Untersuchung - davon **zurücktreten kann**, ohne dadurch Nachteile zu erfahren. In diesem Fall erlischt der Anspruch auf die Grundvergütung von 10,- DM und die Zusatzbelohnungen von 15,- DM bzw. 5,- DM. Ferner erkläre ich mich mit den oben genannten Bedingungen **einverstanden** und verpflichte mich, bis zum Ende der gesamten Versuchsreihe (voraussichtlich Sommer 2000) keine Detailinformationen über das Experiment weiterzugeben.

Nachname: _____ Vorname: _____

Mainz, den _____ Unterschrift: _____

Im Anschluß an das Experiment erhalten Sie eine Kopie dieser Einverständniserklärung.

Anhang A.2: Instruktion 1 für die *Täterinnen*

T-1

Vielen Dank, daß Sie sich bereiterklärt haben, an einer Untersuchung zur Glaubwürdigkeitsbegutachtung / Lügendetektion teilzunehmen.

Ihre Aufgabe besteht zunächst darin, einen Diebstahl zu begehen. Anschließend sollen Sie die mit der Aufklärung des Delikts betraute gerichtspsychologische Expertin überlisten, indem Sie diese von Ihrer Unschuld überzeugen.

Wenn Sie den Diebstahl begehen, wird noch eine weitere weibliche Versuchsperson als **Zeugin anwesend** sein und Ihre Tat beobachten. Später werden sowohl Sie als auch die Zeugin verdächtigt werden, die Tat begangen zu haben.

Die wahre Täterin soll von einer gerichtspsychologischen Expertin entlarvt werden. Die gerichtspsychologische Expertin führt bei beiden Verdächtigen eine **mündliche Vernehmung** und einen **Lügendetektortest** durch und beurteilt dann anhand der Untersuchungsergebnisse, welche von beiden Verdächtigen den Diebstahl begangen hat. Die Untersuchung durch die gerichtspsychologische Expertin wird per Videokamera aufgezeichnet.

Die gerichtspsychologische Expertin ist nicht darüber informiert, ob Sie den Diebstahl tatsächlich begangen haben oder ob Sie den Diebstahl nur als Zeugin beobachtet haben.

Wenn es Ihnen gelingt, die gerichtspsychologische Expertin von Ihrer Unschuld bzw. von der Schuld der anderen tatverdächtigen Person zu überzeugen, erhalten Sie eine **zusätzliche Belohnung in Höhe von 15,- DM**. Diese Belohnung bleibt Ihnen jedoch versagt, wenn Sie von der gerichtspsychologischen Expertin als Täterin überführt werden.

Melden Sie sich bitte beim Versuchsleiter, sobald Sie diese Instruktion gelesen haben.

Wenn Sie bis hierher Fragen haben, wenden Sie sich bitte an den Versuchsleiter.

Anhang A.3: Instruktion 2 für die *Täterinnen*

T-2

Im folgenden erhalten Sie die Instruktion für die Durchführung des Diebstahls. **Bitte lesen Sie sich die Anweisungen zweimal genau durch:**

Ihre Aufgabe besteht darin, aus dem Büro von Prof. Kunze (Raum 02-527) **Geld zu entwenden**. Das Geld ist in einer **Geldkassette** verstaut, die irgendwo in Prof. Kunzes Büro versteckt ist. Das Schloß dieser Geldkassette läßt sich nur mit Hilfe einer **10-stelligen Zahlenkombination** öffnen.

Da Prof. Kunze sich Zahlen nicht gut merken kann, hat er die **einzelnen Ziffern der Zahlenkombination an unterschiedlichen Stellen seines Büros schriftlich niedergelegt**. Durch Zufall haben Sie erfahren, daß der Notizzettel mit der ersten Ziffer der Zahlenkombination sich in der **linken oberen Schublade des großen Schreibtischs** befindet. Außerdem enthält dieser Notizzettel auch Angaben über den Ort, an dem der Notizzettel mit der zweiten Ziffer der Zahlenkombination versteckt ist. Der Notizzettel mit der zweiten Ziffer der Zahlenkombination wiederum enthält außerdem auch Angaben über den Ort, an dem sich der Zettel mit der dritten Ziffer der Zahlenkombination befindet u.s.w. Der Zettel mit der letzten Ziffer der 10-stelligen Zahlenkombination enthält schließlich auch noch **Angaben über den Ort, an dem die Geldkassette verstaut ist**.

Gehen Sie also in das **Büro von Prof. Kunze (Raum 02-527)**.

Schauen Sie dann in der **linken oberen Schublade des großen Schreibtischs** (die mit dem **Deutschlandaufkleber**) nach dem Notizzettel mit der ersten Ziffer der 10-stelligen Zahlenkombination und notieren Sie sich diese. (Einen Zettel und einen Stift erhalten Sie vorher vom Versuchsleiter.)

Nachdem Sie alle 10 Ziffern der Zahlenkombination und das Versteck der Geldkassette herausgefunden haben, **öffnen Sie die Kassette und nehmen Sie das ganze Geld heraus**.

Anschließend verlassen Sie mit dem Geld den Raum und **kehren hierher zurück**.

Im Büro von Prof. Kunze wird sich **noch eine weitere weibliche Versuchsperson** befinden, die gerade dort arbeitet. Versuchen Sie, sich dieser Person gegenüber **möglichst unauffällig** zu **benehmen**.

Falls die andere Person Sie auf Ihr Tun hin anspricht, lassen Sie sich **möglichst plausible Ausreden** einfallen. Sie können z.B. vorgeben, daß Sie **Mitarbeiterin der zentralen Universitätsverwaltung sind und den Auftrag haben, im Büro von Prof. Kunze eine Inventarliste zu erstellen**.

Prägen Sie sich die obigen Anweisungen bitte gut ein und wiederholen Sie die wesentlichen Punkte noch einmal gegenüber dem Versuchsleiter.

Melden Sie sich bitte beim Versuchsleiter, sobald Sie diese Instruktion zweimal gelesen haben.

Wenn Sie noch Fragen haben, wenden Sie sich bitte an den Versuchsleiter.

Anhang A.4: Instruktion 3 für die *Täterinnen*

T-3

Sie befinden sich nun in folgender Situation:

Sie haben soeben einen **Gelddiebstahl begangen**. Dabei war eine andere weibliche Person anwesend, die Ihr Delikt vermutlich als **Zeugin** beobachtet hat.

Prof. Kunze will, daß der Diebstahl so schnell wie möglich aufgeklärt wird. Folgende Tatbestände sind den Ermittlungsbehörden bekannt: Der Diebstahl wurde heute im Büro von Prof. Kunze (Raum 02-527) verübt. Es wurden 100,- DM entwendet. Außerdem ist bekannt, daß während des Diebstahls neben dem Täter / der Täterin noch eine weitere Person am Tatort anwesend war, die möglicherweise den Diebstahl beobachtet hat.

Für sachdienliche Hinweise, die zur Ergreifung des Täters / der Täterin führen, ist eine Belohnung in Höhe von 15,- DM ausgesetzt. In diesem Zusammenhang bitten die Ermittlungsbehörden v.a. den mutmaßlichen Zeugen / die mutmaßliche Zeugin, sich zu melden und eine Zeugenaussage abzugeben.

Da Sie befürchten müssen, durch die Zeugenaussage der Person, die den Diebstahl beobachtet hat, überführt zu werden, beschließen Sie ihrerseits, ebendiese Person des Diebstahls zu bezichtigen. Auf diese Weise können Sie zudem auch noch die ausgesetzte Belohnung erhalten.

Daher wollen Sie die Untersuchungen zur Aufklärung des Diebstahls nutzen, um **sich selbst zu entlasten und die eigentliche Zeugin als Täterin zu belasten**. Die Untersuchungen zur Aufklärung des Diebstahls werden von einer **gerichtspsychologischen Expertin** vorgenommen und bestehen aus:

1) einer mündlichen Vernehmung und
2) einem Lügendetektortest.

Die gerichtspsychologische Expertin weiß nicht, ob Sie den Diebstahl tatsächlich begangen haben oder ob Sie den Diebstahl nur als Zeugin beobachtet haben.

Wenn es Ihnen gelingt, die gerichtspsychologische Expertin von Ihrer eigenen Unschuld bzw. von der Schuld der Zeugin zu überzeugen, erhalten Sie die ausgesetzte Belohnung in Höhe von 15,- DM. Mit anderen Worten: **Sie erhalten die ausgesetzte Belohnung in Höhe von 15,- DM nur, wenn es Ihnen gelingt, in der mündlichen Vernehmung glaubwürdig zu erscheinen und den Lügendetektortest zu bestehen.**

Melden Sie sich bitte beim Versuchsleiter, sobald Sie diese Instruktion gelesen haben.

Wenn Sie bis hierher Fragen haben, wenden Sie sich bitte an den Versuchsleiter.

Anhang A.5: Instruktion 4 für die *Täterinnen*

T-4

Es folgen nun die Untersuchungen zur Aufklärung des Gelddiebstahls.

Zunächst erfolgt die **mündliche Vernehmung durch die gerichtspsychologische Expertin**. Diese Expertin ist darin geschult, die Glaubwürdigkeit von Zeugenaussagen zu analysieren. Die Expertin ist nicht darüber informiert, ob Sie den Diebstahl tatsächlich begangen haben oder ob Sie den Diebstahl nur als Zeugin beobachtet haben.

Die einzige Möglichkeit, die gerichtspsychologische Expertin von Ihrer Unschuld zu überzeugen, besteht darin, daß Sie **die andere Person, die während des Diebstahls am Tatort anwesend war, beschuldigen, den Diebstahl begangen zu haben**.

Es reicht jedoch keineswegs aus, einfach nur zu behaupten, die andere Person habe das Geld gestohlen.

Bedenken Sie, daß die andere Person Sie ebenfalls des Diebstahls beschuldigt und dabei eine umfangreiche und plausible Zeugenaussage über den gesamten Tathergang abgibt. Sie können den Tatverdacht also nur dann überzeugend von sich auf die andere Person lenken, wenn auch Sie sich **eine umfangreiche und zugleich plausibel erscheinende Geschichte über den gesamten Tathergang ausdenken.** Es ist wichtig, daß Sie während Ihrer Aussage **so glaubwürdig wie möglich** erscheinen.

Zu Beginn der mündlichen Vernehmung wird die gerichtspsychologische Expertin Sie auffordern, einen **zusammenhängenden Bericht über den gesamten Tathergang** abzugeben. Dieser zusammenhängende Bericht ist für die Beurteilung der Glaubwürdigkeit Ihrer Aussage von besonderer Bedeutung.

Im folgenden haben Sie 15 Minuten Zeit, um sich auf Ihre Aussage, insbesondere auf den zusammenhängenden Bericht über den gesamten Tathergang vorzubereiten. Um die Vorbereitung auf Ihre Aussage zu erleichtern, können Sie sich in der Vorbereitungsphase erneut **am Tatort (Büro von Prof. Kunze) aufhalten**.

Gehen Sie also erneut in das Büro von Prof. Kunze (Raum 02-527). Setzen Sie sich dort an einen der Tische. Dort haben Sie 15 Minuten Zeit, um sich auf Ihre Aussage zum Hergang des Gelddiebstahls vorzubereiten.

Nach Ablauf der 15 Minuten wird der Versuchsleiter in den Raum treten und Ihnen das Ende der Vorbereitungszeit signalisieren. Anschließend wird er Sie in den Raum führen, in dem die mündliche Vernehmung durch die gerichtspsychologische Expertin erfolgt.

Prägen Sie sich die obigen Anweisungen bitte gut ein und wiederholen Sie die wesentlichen Punkte noch einmal gegenüber dem Versuchsleiter.

Melden Sie sich bitte beim Versuchsleiter, sobald Sie diese Instruktion gelesen haben.

Wenn Sie noch Fragen haben, wenden Sie sich bitte an den Versuchsleiter.

Anhang A.6: Instruktion 1 für die *Zeuginnen*

EZ-1

Vielen Dank, daß Sie sich bereiterklärt haben, an einer Untersuchung zur Glaubwürdigkeitsbegutachtung / Lügendetektion teilzunehmen.

Ihre Aufgabe besteht zunächst darin, als Zeugin einen Diebstahl zu beobachten, der von einer anderen weiblichen Versuchsperson begangen wird. Anschließend sollen Sie durch Ihre Zeugenaussage zur Ergreifung der Täterin beitragen.

Im Anschluß an die Tat wird die Täterin jedoch versuchen, den Verdacht von sich abzuwenden, indem sie **Sie** beschuldigt. Daher werden sowohl die wahre Täterin als auch Sie verdächtigt, die Tat begangen zu haben.

Die wahre Täterin soll von einer gerichtspsychologischen Expertin entlarvt werden. Die gerichtspsychologische Expertin führt bei beiden Verdächtigen eine **mündliche Vernehmung** und einen **Lügendetektortest** durch und beurteilt dann anhand der Untersuchungsergebnisse, welche von beiden Verdächtigen den Diebstahl begangen hat. Die Untersuchung durch die gerichtspsychologische Expertin wird per Videokamera aufgezeichnet.

Die gerichtspsychologische Expertin ist nicht darüber informiert, ob Sie den Diebstahl tatsächlich begangen haben oder ob Sie den Diebstahl nur als Zeugin beobachtet haben.

Wenn es Ihnen gelingt, die gerichtspsychologische Expertin von Ihrer Unschuld bzw. von der Schuld der anderen tatverdächtigen Person zu überzeugen, erhalten Sie eine **zusätzliche Belohnung in Höhe von 15,- DM**. Diese Belohnung bleibt Ihnen jedoch versagt, wenn Sie von der gerichtspsychologischen Expertin für schuldig befunden werden.

Melden Sie sich bitte beim Versuchsleiter, sobald Sie diese Instruktion gelesen haben.

Wenn Sie bis hierher Fragen haben, wenden Sie sich bitte an den Versuchsleiter.

Anhang A.7: Instruktion 2 für die *Zeuginnen*

EZ-2

Im folgenden erhalten Sie die Instruktion für den weiteren Verlauf der Untersuchung. **Bitte lesen Sie sich die Anweisungen zweimal genau durch:**

Versetzen Sie sich bitte in folgende Situation: Sie arbeiten als **Reinigungskraft** im Psychologischen Institut. Ihre Aufgabe besteht darin, das **Büro** von Prof. Kunze (Raum 02-527) **aufzuräumen bzw. zu reinigen**. Im Verlauf dieser Tätigkeit werden Sie **Zeugin eines Diebstahls** werden.

Gehen Sie also in das Büro von Prof. Kunze (Raum 02-527). Auf dem **großen Schreibtisch** finden Sie einen **Zettel** vor, den Prof. Kunze für Sie geschrieben hat. Auf dem Zettel hat Prof. Kunze **10 kleinere Aufräum- bzw. Reinigungsarbeiten aufgelistet**, die von Ihnen erledigt werden sollen. Beginnen Sie mit der Erledigung dieser Arbeiten.

Nach einer gewissen Zeit wird eine **weitere weibliche Person** das Büro betreten, um **etwas ganz Bestimmtes in Prof. Kunzes Büro zu suchen und zu entwenden**. Unterbrechen Sie dann Ihre Arbeit, um den Diebstahl beobachten zu können. Damit nicht zu sehr auffällt, daß Sie die Täterin beobachten, tun Sie am besten so, als würden Sie eine ganz normale Arbeitspause einlegen.

Beobachten Sie ganz genau das Verhalten der anderen Person.

Versuchen Sie, die andere Person in ein Gespräch zu verwickeln, um herauszufinden, was sie stehlen will.

Wenn die andere Person das Büro von Prof. Kunze wieder verlassen hat, erledigen Sie die restlichen Aufräum- bzw. Reinigungsarbeiten gemäß der Aufgabenliste von Prof. Kunze. Für den weiteren Verlauf der Untersuchung ist es **wichtig, daß Sie alle von Prof. Kunze aufgelisteten Aufgaben erledigt haben**. Verlassen Sie danach das Büro und kehren Sie hierher zurück.

Prägen Sie sich die obigen Anweisungen bitte gut ein und wiederholen Sie die wesentlichen Punkte noch einmal gegenüber dem Versuchsleiter.

Melden Sie sich bitte beim Versuchsleiter, sobald Sie diese Instruktion zweimal gelesen haben.

Wenn Sie noch Fragen haben, wenden Sie sich bitte an den Versuchsleiter.

Anhang A.8: Instruktion 3 für die *Zeuginnen*

EZ-3

Sie befinden sich nun in folgender Situation:

Sie sind soeben Zeugin eines Gelddiebstahls geworden.

Prof. Kunze will, daß der Diebstahl so schnell wie möglich aufgeklärt wird. Folgende Tatbestände sind den Ermittlungsbehörden bekannt: Der Diebstahl wurde heute im Büro von Prof. Kunze (Raum 02-527) verübt. Es wurden 100,- DM entwendet. Außerdem ist bekannt, daß während des Diebstahls neben dem Täter / der Täterin noch eine weitere Person am Tatort anwesend war, die möglicherweise den Diebstahl beobachtet hat.

Für sachdienliche Hinweise, die zur Ergreifung des Täters / der Täterin führen, ist eine Belohnung in Höhe von 15,- DM ausgesetzt. In diesem Zusammenhang bitten die Ermittlungsbehörden v.a. den mutmaßlichen Zeugen / die mutmaßliche Zeugin, sich zu melden und eine Zeugenaussage abzugeben.

Sie beschließen, zur Ergreifung der Täterin beizutragen und die ausgesetzte Belohnung zu erhalten, indem Sie **eine Zeugenaussage machen**.

Unglücklicherweise geraten neben der wahren Täterin aber auch **Sie** selber in Verdacht, da die Täterin zu ihrer eigenen Entlastung behauptet, sie habe **Sie** bei dem Diebstahl beobachtet. In den Untersuchungen zur Aufklärung des Diebstahls erhalten sie jedoch die Gelegenheit, **sich selbst zu entlasten und die wahre Täterin zu belasten**. Die Untersuchungen zur Aufklärung des Diebstahls werden von einer **gerichtspsychologischen Expertin** vorgenommen und bestehen aus:

1) einer mündlichen Vernehmung und
2) einem Lügendetektortest.

Die gerichtspsychologische Expertin weiß nicht, ob Sie den Diebstahl tatsächlich begangen haben oder ob Sie den Diebstahl nur als Zeugin beobachtet haben.

Wenn es Ihnen gelingt, die gerichtspsychologische Expertin von Ihrer eigenen Unschuld bzw. von der Schuld der wahren Täterin zu überzeugen, erhalten Sie die ausgesetzte Belohnung in Höhe von 15,- DM. Mit anderen Worten: **Sie erhalten die ausgesetzte Belohnung in Höhe von 15,- DM nur, wenn es Ihnen gelingt, in der mündlichen Vernehmung glaubwürdig zu erscheinen und den Lügendetektortest zu bestehen.**

Melden Sie sich bitte beim Versuchsleiter, sobald Sie diese Instruktion gelesen haben.

Wenn Sie bis hierher Fragen haben, wenden Sie sich bitte an den Versuchsleiter.

Anhang A.9: Instruktion 4 für die *Zeuginnen*

EZ-4

Es folgen nun die Untersuchungen zur Aufklärung des Gelddiebstahls.

Zunächst erfolgt die **mündliche Vernehmung durch die gerichtspsychologische Expertin**. Diese Expertin ist darin geschult, die Glaubwürdigkeit von Zeugenaussagen zu analysieren. Die Expertin ist nicht darüber informiert, ob Sie den Diebstahl tatsächlich begangen haben oder ob Sie den Diebstahl nur als Zeugin beobachtet haben.

Die einzige Möglichkeit, die gerichtspsychologische Expertin von Ihrer Unschuld zu überzeugen, besteht darin, daß Sie **die wahre Täterin beschuldigen, den Diebstahl begangen zu haben**.

Es reicht jedoch keineswegs aus, einfach nur zu behaupten, die Täterin habe das Geld gestohlen.

Bedenken Sie, daß die Täterin Sie ebenfalls des Diebstahls beschuldigt, so daß in bezug auf die Anschuldigungen Aussage gegen Aussage steht. Sie können den Tatverdacht gegen Sie also nur dann überzeugend entkräften und zur Überführung der wahren Täterin beitragen, wenn Sie **eine umfassende Zeugenaussage über den gesamten Tathergang ablegen.** Es ist wichtig, daß Sie während Ihrer Aussage **so glaubwürdig wie möglich** erscheinen.

Zu Beginn der mündlichen Vernehmung wird die gerichtspsychologische Expertin Sie auffordern, einen **zusammenhängenden Bericht über den gesamten Tathergang** abzugeben. Dieser zusammenhängende Bericht ist für die Beurteilung der Glaubwürdigkeit Ihrer Aussage von besonderer Bedeutung.

Im folgenden haben Sie 15 Minuten Zeit, um sich auf Ihre Aussage, insbesondere auf den zusammenhängenden Bericht über den gesamten Tathergang vorzubereiten. Um die Vorbereitung auf Ihre Aussage zu erleichtern, können Sie sich in der Vorbereitungsphase erneut **am Tatort (Büro von Prof. Kunze) aufhalten**.

Gehen Sie also erneut in das Büro von Prof. Kunze (Raum 02-527). Setzen Sie sich dort an einen der Tische. Dort haben Sie 15 Minuten Zeit, um sich auf Ihre Aussage zum Hergang des Gelddiebstahls vorzubereiten.

Nach Ablauf der 15 Minuten wird der Versuchsleiter in den Raum treten und Ihnen das Ende der Vorbereitungszeit signalisieren. Anschließend wird er Sie in den Raum führen, in dem die mündliche Vernehmung durch die gerichtspsychologische Expertin erfolgt.

Prägen Sie sich die obigen Anweisungen bitte gut ein und wiederholen Sie die wesentlichen Punkte noch einmal gegenüber dem Versuchsleiter.

Melden Sie sich bitte beim Versuchsleiter, sobald Sie diese Instruktion gelesen haben.

Wenn Sie noch Fragen haben, wenden Sie sich bitte an den Versuchsleiter.

Anhang A.10: Instruktion 1 für die *falschen Zeuginnen*

FZ-1

Vielen Dank, daß Sie sich bereiterklärt haben, an einer Untersuchung zur Glaubwürdigkeitsbegutachtung / Lügendetektion teilzunehmen.

Für den weiteren Verlauf der Untersuchung ist es notwendig, daß Sie sich **in folgende Situation hineinversetzen.** Lesen Sie sich daher bitte die folgenden Informationen zweimal genau durch.

Im Psychologischen Institut ist ein Gelddiebstahl begangen worden. Folgende Tatbestände sind den Ermittlungsbehörden bekannt: Der Diebstahl wurde **heute** im **Büro von Prof. Kunze (Raum 02-527)** verübt. Es wurden **100,- DM** entwendet. Außerdem ist bekannt, daß während des Diebstahls neben dem Täter / der Täterin noch eine **weitere Person am Tatort anwesend** war, die möglicherweise den Diebstahl beobachtet hat. Der Diebstahl soll so schnell wie möglich aufgeklärt werden.

Für sachdienliche Hinweise, die zur Ergreifung des Täters / der Täterin führen, ist eine Belohnung in Höhe von 15,- DM ausgesetzt. In diesem Zusammenhang bitten die Ermittlungsbehörden v.a. den mutmaßlichen Zeugen / die mutmaßliche Zeugin, sich zu melden und eine Zeugenaussage abzugeben.

Sie haben zwar die Tat nicht beobachtet. Sie benötigen jedoch gerade dringend Geld.
Daher **beschließen Sie, die ausgesetzte Belohnung zu erhalten, indem Sie eine Zeugenaussage erfinden, in welcher Sie eine fiktive Person des Diebstahls beschuldigen**.

Ihre Aussage wird jedoch nicht ohne weiteres als wahr / zutreffend akzeptiert werden, sondern Sie werden **von einer gerichtspsychologischen Expertin begutachtet**, die darin ausgebildet ist, die Glaubwürdigkeit von Zeugenaussagen zu analysieren. Die gerichtspsychologische Expertin führt mit Ihnen eine **mündliche Vernehmung** und einen **Lügendetektortest** durch und beurteilt dann anhand der Untersuchungsergebnisse, ob Sie den Diebstahl tatsächlich als Zeugin beobachtet haben oder nicht. Die Untersuchung durch die gerichtspsychologische Expertin wird per Videokamera aufgezeichnet.

Die gerichtspsychologische Expertin ist nicht darüber informiert, ob Sie den Diebstahl wirklich beobachtet haben oder nicht.

Sie erhalten die ausgesetzte Belohnung **in Höhe von 15,- DM** nur, wenn die gerichtspsychologische Expertin Ihre Zeugenaussage als glaubwürdig beurteilt. Es ist also wichtig, daß Sie während Ihrer Aussage **so glaubwürdig wie möglich erscheinen**. Besonders wichtig ist, daß Sie Ihre **Aussage gegenüber der gerichtspsychologischen Expertin so umfangreich und plausibel wie möglich** gestalten.

Melden Sie sich bitte beim Versuchsleiter, sobald Sie diese Instruktion zweimal gelesen haben.

Wenn Sie bis hierher Fragen haben, wenden Sie sich bitte an den Versuchsleiter.

Anhang A.11: Instruktion 2 für die *falschen Zeuginnen*

FZ-2

Sie haben sich entschlossen, die ausgesetzte Belohnung in Höhe von 15,- DM zu erhalten, indem Sie eine Zeugenaussage erfinden, in welcher Sie eine fiktive Person des Diebstahls beschuldigen. Wie bereits erwähnt wurde, werden Sie zunächst von einer **gerichtspsychologischen Expertin** vernommen, die darin ausgebildet ist, die Glaubwürdigkeit von Zeugenaussagen zu analysieren. Die gerichtspsychologische Expertin ist nicht darüber informiert, ob Sie den Diebstahl wirklich als Zeugin beobachtet haben oder nicht.

Am Anfang der Vernehmung wird die gerichtspsychologische Expertin Sie auffordern, einen **zusammenhängenden Bericht über den gesamten Tathergang** abzugeben. Dieser zusammenhängende Bericht ist für die Beurteilung der Glaubwürdigkeit Ihrer Aussage **von besonderer Bedeutung**. Er sollte **so umfangreich und zugleich so plausibel wie möglich** sein. Es ist wichtig, daß Sie während Ihrer Aussage **so glaubwürdig wie möglich** erscheinen.

Im folgenden haben Sie 15 Minuten Zeit, um sich auf Ihre Aussage, insbesondere auf den zusammenhängenden Bericht über den gesamten Tathergang vorzubereiten. Um die Vorbereitung auf Ihre Aussage zu erleichtern, können Sie sich in der Vorbereitungsphase **am wirklichen Tatort (Büro von Prof. Kunze) aufhalten**.

Gehen Sie also in das Büro von Prof. Kunze (Raum 02-527). Setzen Sie sich dort an einen der Tische. Dort haben Sie 15 Minuten Zeit, eine Zeugenaussage zum Hergang des Gelddiebstahls zu erfinden. Dabei können Sie sich an den räumlichen Details des Büros sowie an folgenden, bereits an die Öffentlichkeit gelangten Informationen orientieren:

- Der Gelddiebstahl wurde **heute** im **Büro von Prof. Kunze (Raum 02-527)** verübt.
- Es wurden **100,- DM** gestohlen.
- Während des Tathergangs war eine **weitere Person anwesend**, die den Diebstahl möglicherweise als Zeuge / Zeugin beobachtet hat.
- Möglicherweise kam es zum **Kontakt** zwischen dem Täter / der Täterin und dem Zeugen / der Zeugin.

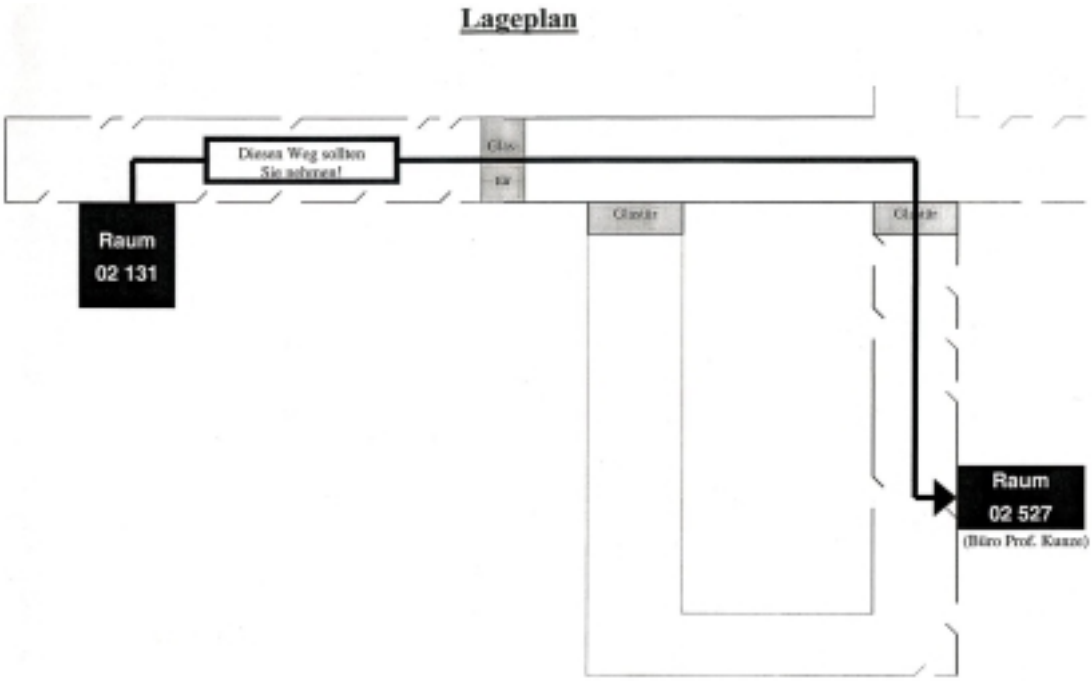
Nach Ablauf der 15 Minuten wird der Versuchsleiter in den Raum treten und Ihnen das Ende der Vorbereitungszeit signalisieren. Anschließend wird er Sie in den Raum führen, in dem die mündliche Vernehmung durch die gerichtspsychologische Expertin erfolgt.

Prägen Sie sich die obigen Anweisungen bitte gut ein und wiederholen Sie die wesentlichen Punkte noch einmal gegenüber dem Versuchsleiter.

Melden Sie sich bitte beim Versuchsleiter, sobald Sie diese Instruktion gelesen haben.

Wenn Sie noch Fragen haben, wenden Sie sich bitte an den Versuchsleiter.

Anhang A.12: Wegskizze



Anhang A.13: Aufgabenzettel für die Reinigungskraft (Version für Verbrechenssimulation mit *Täterinnen* und *Zeuginnen*)

Aufgabenzettel für Reinigungskraft

Bitte erledigen Sie folgende Arbeiten:

1. Die Schraube des Griffs der linken oberen Schreibtischschublade (die mit dem **Deutschland-Aufkleber**) hat sich gelockert und muß festgezogen werden. Genau in dieser Schublade befindet sich ein Schraubenzieher, den Sie dazu benutzen können.

2. Auf dem Tisch mit dem Tageslichtprojektor steht ein **Kaktus**, den Sie bitte gießen sollten.

3. An der Wand hängt ein **Bild mit Kühen**. Bitte entstauben Sie dessen Bilderrahmen. Hierzu finden Sie im Waschbecken einen Lappen.

4. Bitte wischen Sie auch den **Porzellanhund** ab, der in dem Regal an der linken Wand steht.

5. An der Wand beim Fenster hängt eine **Dart-Scheibe** mit Pfeilen. Nehmen Sie die Pfeile ab und legen Sie sie auf den Tisch, auf dem die Schreibmaschine steht.

6. In dem Holzschrank befindet sich sich eine schwarze **Jacke**, die auf den Bügel gehängt werden muß.

7. Saugen Sie bitte den **roten Teppich** auf dem Fußboden. Der Staubsauger befindet sich neben der Eingangstür.

8. Das **gelbe Fahrrad** ist an den Schrank angelehnt. Stellen Sie es bitte auf den Fahrradständer.

9. Beim Waschbecken steht ein **Kasten mit Wasserflaschen**. Räumen Sie die Flaschen, die auf dem Couchtisch stehen, in den Kasten ein.

10. Auf dem Couchtisch befindet sich eine **Obstschale mit Äpfeln**. Falls Äpfel aus der Obstschale gefallen sind, legen Sie sie zurück in die Obstschale.

Viele Grüße

W. Kunze

Anhang A.14: Aufgabenzettel für die Reinigungskraft (Version für Bedingung *falsche Zeuginnen*)

Aufgabenzettel für Reinigungskraft

Bitte erledigen Sie folgende Arbeiten:

1. Die Schraube des Griffs der linken oberen Schreibtischschublade (die mit dem **Schweiz-Aufkleber**) hat sich gelockert und muß festgezogen werden. Genau in dieser Schublade befindet sich ein Schraubenzieher, den Sie dazu benutzen können.

2. Auf dem Tisch mit dem Tageslichtprojektor steht eine **Rose**, die Sie bitte gießen sollten.

3. An der Wand hängt ein **Bild mit Fröschen**. Bitte entstauben Sie dessen Bilderrahmen. Hierzu finden Sie im Waschbecken einen Lappen.

4. Bitte wischen Sie auch den **Porzellanelefant** ab, der in dem Regal an der linken Wand steht.

5. An der Wand beim Fenster hängt ein **Gymnastikreifen**. Wischen Sie diesen ebenfalls mit dem Lappen ab.

6. In dem Holzschrank befindet sich sich ein weißer **Schal**, der auf den Bügel gehängt werden muß.

7. Saugen Sie bitte den **grauen Teppich** auf dem Fußboden. Der Staubsauger befindet sich neben der Eingangstür.

8. Das **silberne Fahrrad** hat platte Reifen. Bitte pumpen Sie beide Reifen auf.

9. Beim Waschbecken steht ein **Kasten Limonade**. Räumen Sie die Flaschen, die auf dem Couchtisch stehen, in den Kasten ein.

10. Auf dem Couchtisch befindet sich eine **Obstschale mit Nüssen**. Falls Nüsse aus der Obstschale gefallen sind, legen Sie sie zurück in die Obstschale.

Viele Grüße

W. Kunze

Anhang A.15: Versteckte Notizzettel mit den Ziffern des Zahlenschloßcodes

1. Ziffer: **5** nächste Ziffer: **unter Kaktus**

2. Ziffer: **2** nächste Ziffer: **unter Porzellanhund**

3. Ziffer: **8** nächste Ziffer: **gelbes Fahrrad (Satteltasche)**

4. Ziffer: **3** nächste Ziffer: **hinter dem Bild mit Kühen**

5. Ziffer: **6** nächste Ziffer: **Getränkekasten mit Wasserflaschen**

6. Ziffer: **1** nächste Ziffer: **rechte Tasche der Jacke im Holzschrank**

7. Ziffer: **4** nächste Ziffer: **unter der Obstschale mit Äpfeln**

8. Ziffer: **7** nächste Ziffer: **hinter der Dartscheibe an der Wand**

9. Ziffer: **0** nächste Ziffer: **unter dem roten Teppich**

10. Ziffer: **9** Ort der Geldkassette: **im untersten Regalfach im roten Karton**

Anhang A.16: Notizzettel für die *Täterinnen*

	Ort:	**Ziffer:**
1. Ziffer:	*Schreibtischschublade mit Deutschlandaufkleber*	
2. Ziffer:		
3. Ziffer:		
4. Ziffer:		
5. Ziffer:		
6. Ziffer:		
7. Ziffer:		
8. Ziffer:		
9. Ziffer:		
10. Ziffer:		

Anhang A.17: *GAT*-Instruktion 1

Sie haben soeben eine Zeugenaussage zu dem Gelddiebstahl im Psychologischen Institut abgegeben, in welcher Sie eine andere Person des Diebstahls beschuldigen.

Natürlich muß rein routinemäßig auch die Möglichkeit in Betracht gezogen werden, daß **Sie** in Wirklichkeit selber das Geld gestohlen haben und nun durch eine falsche Zeugenaussage die Ermittlungen in eine falsche Richtung lenken wollen, um den Verdacht von sich abzuwenden.

Im folgenden wird ein **Lügendetektortest** mit Ihnen durchgeführt, **um festzustellen, ob Sie den Gelddiebstahl begangen haben oder nicht**.

Bei dem Lügendetektortest werden Ihnen **verschiedene Fragen gestellt**. Während Sie diese Fragen beantworten, werden Ihre **körperlichen Reaktionen gemessen**. Anhand der Stärke der körperlichen Reaktionen wird der **Wahrheitsgehalt Ihrer Antworten ermittelt**.

Die Lügendetektion basiert darauf, daß **man beim Lügen aufgeregter ist als beim Wahrheitsagen**. Diese Aufregung zeigt sich auch in bestimmten körperlichen Veränderungen, die willentlich nur schwer kontrollierbar sind, die man aber mit Hilfe wissenschaftlicher Meßgeräte genau erfassen kann. Am deutlichsten zeigt sich die erhöhte Aufregung beim Lügen

- in der **Aktivität der Schweißdrüsen,**
- in der **Herzschlagfrequenz,**
- in der **peripheren Durchblutung** und
- in der **Atmung**.

Wissenschaftliche Studien haben gezeigt, daß die **Lügendetektion eine hohe Treffsicherheit bei der Identifizierung von Tätern und unschuldigen Personen aufweist**. Andererseits ist jedoch auch erwiesen, daß Personen mit hoher Intelligenz, die ihre Emotionen gut kontrollieren können, mitunter in der Lage sind, den Lügendetektor zu überlisten.

In dem folgenden Lügendetektortest werden Ihre Schweißdrüsenaktivität, Herzschlagfrequenz, periphere Durchblutung und Atmung gemessen. Dazu werden **jeweils 2 Meßfühler an Ihrer linken und rechten Hand** angelegt. **3 weitere Meßfühler und ein Gürtel** zur Erfassung von Bewegungen werden **an Ihrem Oberkörper** angebracht. Bitte legen Sie vor dem Anbringen der Meßfühler ggf. Ihren Schmuck (Uhren, Ringe, Ketten o.ä.) ab und entfernen Sie ggf. Kaugummis oder Bonbons aus Ihrem Mund. Nach dem Anbringen der Meßfühler nehmen Sie bitte in der Kabine Platz.

Vor dem eigentlichen Lügendetektortest werden zunächst eine **Ruhmessung** und ein **Vortest** durchgeführt.

Melden Sie sich bitte bei der Testleiterin, sobald Sie diese Instruktion gelesen haben.

Falls Sie bis hierher Fragen haben, wenden Sie sich bitte an die Testleiterin.

Anhang A.18: *GAT*-Instruktion 2

Es erfolgt nun eine **sechsminütige Ruhmessung**. Diese wird benötigt, um Ihre körperliche Aktivität während einer Ruhephase zu erfassen und um die Meßinstrumente genau einzustellen. Die Messung ist absolut ungefährlich.

Bitte setzen Sie sich während der Ruhemessung **bequem und entspannt** auf den Stuhl. Halten Sie die **Augen geöffnet**. Es ist wichtig, daß Sie sich **während der Ruhemessung möglichst wenig bewegen, nicht sprechen, nicht räuspern o.ä.** Die Kamera in der Kabine dient dazu, bewegungsbedingte Meßfehler zu erkennen.

Für die Ruhemessung - ebenso wie für die weiteren im Verlauf des Lügendetektortests erfolgenden Messungen - gilt folgendes: Versuche, durch bestimmte Bewegungen die Messungen zu beeinflussen bzw. zu stören, können leicht mit Hilfe der Videokamera entlarvt werden. Zudem sind bewegungsbedingte Meßfehler auch anhand ihrer charakteristischen Kurvenverläufe leicht identifizierbar. Da nur schuldige ProbandInnen (TäterInnen) ein Interesse daran haben können, die Messungen zu stören bzw. zu verfälschen, **werden körperliche Bewegungen und sonstige Störversuche seitens der Testperson grundsätzlich als Hinweis auf deren Täterschaft gewertet**.

Beginn und Ende der Ruhemessung werden Ihnen jeweils bekanntgegeben. Nach der Ruhemessung erhalten Sie die genauen Anweisungen für den weiteren Testablauf.

Melden Sie sich bitte bei der Testleiterin, sobald Sie diese Instruktion gelesen haben.

Falls Sie noch Fragen haben, wenden Sie sich bitte an die Testleiterin.

Anhang A.19: *GAT*-Instruktion 3

Im folgenden wird ein **Vortest** mit Ihnen durchgeführt.

Das Ziel dieses Vortests besteht darin, **Ihre typischen körperlichen Reaktionen beim Lügen und Wahrheitsagen zu erfassen**. Nur wenn dies gelingt, läßt sich auch der eigentliche Lügendetektortest mit Ihnen durchführen. Sollte sich dagegen zeigen, daß Ihre körperlichen Reaktionen beim Lügen und Wahrheitsagen nicht voneinander unterscheidbar sind, so würde dies bedeuten, daß Sie nicht als Probandin für die Lügendetektion geeignet sind. In diesem Fall erübrigt sich der eigentliche Lügendetektortest, und die Untersuchung wird nach dem Vortest abgebrochen.

Bitte wählen Sie jetzt eine Zahl zwischen 3 und 7 aus. Schreiben Sie die gewählte Zahl groß auf das leere Blatt, welches auf der Halterung vor Ihnen befestigt ist. Benutzen Sie dazu den Filzstift.

Im folgenden werden Ihnen 7 Fragen gestellt, die sich auf die von Ihnen gewählte bzw. aufgeschriebene Zahl beziehen. Die Fragen haben im einzelnen folgenden Wortlaut:

	Antwort
1) Haben Sie die Zahl 2 aufgeschrieben?	nein
2) Haben Sie die Zahl 3 aufgeschrieben?	nein
3) Haben Sie die Zahl 4 aufgeschrieben?	nein
4) Haben Sie die Zahl 5 aufgeschrieben?	nein
5) Haben Sie die Zahl 6 aufgeschrieben?	nein
6) Haben Sie die Zahl 7 aufgeschrieben?	nein
7) Haben Sie die Zahl 8 aufgeschrieben?	nein

Beantworten Sie alle 7 Fragen laut und deutlich mit „nein". Dadurch ist gewährleistet, daß Sie bei 6 Fragen wahrheitsgemäß antworten. Bei der einen Frage hingegen, welche die tatsächlich von Ihnen aufgeschriebene Zahl beinhaltet, ist die Antwort „nein" eine Lüge.

Die Fragen werden Ihnen jeweils sowohl optisch (auf dem Bildschirm) als auch akustisch (über Lautsprecher) dargeboten. **Antworten Sie jeweils erst, nachdem die Frage auf dem Bildschirm ausgeblendet wurde.** Der zeitliche Abstand zwischen den Fragen beträgt ca. 20 Sekunden.

Während der Befragung werden Ihre körperlichen Reaktionen gemessen. Es ist wichtig, daß Sie sich während der Messung **möglichst wenig bewegen, nicht räuspern o.ä.** Mit der Kamera in der Kabine werden bewegungsbedingte Meßfehler kontrolliert.

Antworten Sie stets nur mit „nein", sagen Sie sonst nichts.

Antworten Sie erst, nachdem die Frage auf dem Bildschirm ausgeblendet wurde.

Melden Sie sich bitte bei der Testleiterin, sobald Sie diese Instruktion gelesen haben.

Falls Sie noch Fragen haben, wenden Sie sich bitte an die Testleiterin.

Anhang A.20: *GAT*-Instruktion 4

Es folgt nun der **eigentliche Lügendetektortest, in dem festgestellt werden soll, ob Sie bezüglich des Diebstahls der 100,- DM schuldig oder unschuldig sind**.

Bei dem Lügendetektortest werden Ihnen **10 Fragen** gestellt. Diese Fragen beziehen sich auf die folgenden **10 Details des Tatorts bzw. Tathergangs:**

- die **Farbe des Fahrrads**, das sich am Tatort befand
- die **Art des Getränks**, das sich **in dem Getränkekasten** am Tatort befand
- die **Tierfigur aus Porzellan**, die sich am Tatort befand
- die **Art der Pflanze**, die sich am Tatort befand
- die **Obstsorte**, die sich **in der Obstschale** am Tatort befand
- das **Motiv des Aufklebers**, der sich **auf einer der Schreibtischschubladen** am Tatort befand
- die **Art des Sportgeräts**, das am Tatort **an der Wand** hing
- das **Tierbild**, das am Tatort **an der Wand** hing
- die **Art des Kleidungsstücks**, das sich **in dem Holzschrank** am Tatort befand
- die **Farbe des Teppichs**, der sich am Tatort befand

Zu den 10 Fragen des Lügendetektortests werden Ihnen **jeweils 6 Antwortalternativen** vorgegeben. Die Antwortalternativen werden Ihnen nacheinander dargeboten. Der zeitliche Abstand zwischen den einzelnen Fragen bzw. Antwortalternativen beträgt ca. 20 Sekunden.

In dem folgenden Kasten finden Sie ein **Beispiel** für eine Testfrage mit 6 Antwortalternativen. Das Beispiel dient nur der Veranschaulichung und ist nicht Bestandteil des nachfolgenden Lügendetektortests.

Bei dem Gelddiebstahl wurde ein ganz bestimmter Betrag entwendet. Wie hoch war der Geldbetrag, den Sie gestohlen haben? Waren es ...		Antwort
a)	20,- DM?	nein
b)	50,- DM?	nein
c)	10,- DM?	nein
d)	100,- DM?	nein
e)	500,- DM?	nein
f)	200,- DM?	nein

Beantworten Sie bei dem Lügendetektortest jede Antwortalternative mit **„nein“**. Sagen Sie sonst nichts.

Wenn Sie **unschuldig** sind, haben Sie bei dem Test **nichts zu befürchten**. Als Unschuldige sagen Sie nämlich bei allen Antwortalternativen die **Wahrheit**, sofern Sie instruktionsgemäß immer mit „nein“ antworten. Die Aufrichtigkeit Ihrer Antworten ist - genauso wie bei den 8 aufrichtigen Antworten in dem vorhin durchgeführten Zahlentest - daran erkennbar, daß Sie keine oder nur sehr schwache körperliche Reaktionen zeigen.

Falls Sie aber den **Gelddiebstahl begangen** haben, so reagieren Sie bei allen 10 Fragen des Lügendetektortests auf jeweils eine der 6 Antwortalternativen mit einer **Lüge**, sofern Sie instruktionsgemäß mit „nein“ antworten (im Beispiel: Antwortalternative d). Diese Lügen sind - genauso wie die Lüge in dem vorhin durchgeführten Zahlentest - anhand der starken körperlichen Reaktionen erkennbar. **Falls Sie entgegen der Instruktion mit „ja“ antworten, so verraten Sie sich dadurch selbst**. Wenn Sie also schuldig sind, besteht für Sie nur eine äußerst **geringe Chance, den Lügendetektor zu überlisten**.

Es ist natürlich auch möglich, daß Sie die zutreffenden Antwortalternativen gar **nicht wiedererkennen**. In diesem Fall haben Sie bei dem Lügendetektortest ebenfalls nichts zu befürchten, da Sie ja nicht bewußt lügen können, wenn Sie die zutreffenden Antwortalternativen mit „nein“ beantworten.

Die Fragen und Antwortalternativen werden Ihnen jeweils sowohl optisch (auf dem Bildschirm) als auch akustisch (über Lautsprecher) dargeboten.

Antworten Sie jeweils erst, nachdem die entsprechende Antwortalternative auf dem Bildschirm aus<u></u>geblendet wurde.

Antworten Sie stets laut und deutlich mit „nein“. Sagen Sie sonst nichts.

Es folgt nun noch ein **kurzer Probedurchgang**, um zu überprüfen, ob Sie die Instruktionen für den Lügendetektortest verstanden haben. Der Probedurchgang ist für das Testergebnis völlig irrelevant. Nach dem Probedurchgang beginnt dann endgültig der Lügendetektortest.

Melden Sie sich bitte bei der Testleiterin, sobald Sie diese Instruktion gelesen haben.

Falls Sie noch Fragen haben, wenden Sie sich bitte an die Testleiterin.

Anhang A.21: Nachbefragungsbogen

Vp_____

Auf den folgenden Fragebogen werden Ihnen noch einige Fragen gestellt, die sich auf den Gelddiebstahl und die Untersuchung durch die gerichtspsychologische Expertin (mündliche Vernehmung und Lügendetektortest) beziehen. Die gerichtspsychologische Expertin erfährt **nicht**, welche Antworten Sie auf die folgenden Fragen geben, so daß Ihre Antworten **keinerlei Einfluß** darauf haben, ob die gerichtspsychologische Expertin Sie bzgl. des Gelddiebstahls als schuldig oder unschuldig beurteilt. **Sie können die folgenden Fragen also völlig aufrichtig beantworten!**

Bitte umblättern.

Im folgenden wird überprüft, wie gut Sie sich an bestimmte Einzelheiten des Tatorts erinnern können. Beantworten Sie dazu die folgenden 10 Fragen. **Für jede richtige Antwort erhalten Sie eine zusätzliche Belohnung von 0,25 DM**.

1. Welche Farbe hatte das Fahrrad in dem Raum, in dem die 100,- DM gestohlen wurden?

2. Welches Getränk war in dem Getränkekasten in dem Raum, in dem die 100,- DM gestohlen wurden?

3. Welches Tier stellte die Porzellanfigur in dem Raum dar, in dem die 100,- DM gestohlen wurden?

4. Welche Pflanze befand sich in dem Raum, in dem die 100,- DM gestohlen wurden?

5. Welche Obstsorte befand sich in der Obstschale in dem Raum, in dem die 100,- DM gestohlen wurden?

6. Welcher Landesaufkleber befand sich auf einer der Schubladen in dem Raum, in dem die 100,- DM gestohlen wurden?

7. Welches Sportgerät hing an der Wand des Raumes, in dem die 100,- DM gestohlen wurden?

8. Welche Tiere waren auf dem Bild in dem Raum, in dem die 100,- DM gestohlen wurden?

9. Welches Kleidungsstück befand sich in dem Holzschrank des Raumes, in dem die 100,- DM gestohlen wurden?

10. Welche Farbe hatte der Teppich in dem Raum, in dem die 100,- DM gestohlen wurden?

Bitte umblättern.

Wie sehr waren Sie während der **mündlichen Vernehmung** motiviert, die gerichtspsychologische Expertin von der Glaubwürdigkeit Ihrer Aussage zu überzeugen? Bitte ankreuzen:

Während der mündlichen Vernehmung war ich ...

gar nicht motiviert	kaum motiviert	etwas motiviert	ziemlich motiviert	stark motiviert	äußerst motiviert
1	**2**	**3**	**4**	**5**	**6**

... die gerichtspsychologische Expertin von der Glaubwürdigkeit meiner Aussage zu überzeugen.

Wie sehr waren Sie bei dem **Lügendetektortest** motiviert, einen unschuldigen Eindruck zu hinterlassen? Bitte ankreuzen:

Bei dem Lügendetektortest war ich ...

gar nicht motiviert	kaum motiviert	etwas motiviert	ziemlich motiviert	stark motiviert	äußerst motiviert
1	**2**	**3**	**4**	**5**	**6**

... einen unschuldigen Eindruck zu hinterlassen.

Wie hoch schätzen Sie die Wahrscheinlichkeit ein, daß es Ihnen während der **mündlichen Vernehmung** gelungen ist, die gerichtspsychologische Expertin von der Glaubwürdigkeit Ihrer Zeugenaussage zu überzeugen? Bitte ankreuzen:

0%	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**

Wie hoch schätzen Sie die Wahrscheinlichkeit ein, daß es Ihnen bei dem **Lügendetektortest** gelungen ist, einen unschuldigen Eindruck zu hinterlassen bzw. die Begehung des Diebstahls glaubwürdig abzustreiten? Bitte ankreuzen:

0%	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**

Bitte umblättern.

Im folgenden werden Ihnen nochmals die 10 Fragen zu den Einzelheiten des Tatorts gestellt. Allerdings werden diesmal bei jeder Frage 6 Antwortalternativen vorgegeben. **Kreuzen Sie jeweils die zutreffende Antwortalternative an**. Sie erhalten erneut **für jede richtige Antwort eine zusätzliche Belohnung von 0,25 DM**.

1. Welche Farbe hatte das Fahrrad in dem Raum, in dem die 100,- DM gestohlen wurden? War es ...

❍ rot?
❍ weiß?
❍ blau?
❍ gelb?
❍ schwarz?
❍ grün?

2. Welches Getränk war in dem Getränkekasten in dem Raum, in dem die 100,- DM gestohlen wurden? War es ...

❍ Milch?
❍ Bier?
❍ Cola?
❍ Wein?
❍ Fruchtsaft?
❍ Wasser?

3. Welches Tier stellte die Porzellanfigur in dem Raum dar, in dem die 100,- DM gestohlen wurden? War es ...

❍ eine Giraffe?
❍ ein Hund?
❍ ein Schwan?
❍ ein Hase?
❍ eine Katze?
❍ ein Pferd?

4. Welche Pflanze befand sich in dem Raum, in dem die 100,- DM gestohlen wurden? War es ...

❍ eine Primel?
❍ eine Palme?
❍ ein Kaktus?
❍ ein Efeu?
❍ ein Farn?
❍ eine Geranie?

5. Welche Obstsorte befand sich in der Obstschale in dem Raum, in dem die 100,- DM gestohlen wurden? Waren es ...

❍ Birnen?
❍ Trauben?
❍ Bananen?
❍ Orangen?
❍ Pflaumen?
❍ Äpfel?

Bitte umblättern.

6. Welcher Landesaufkleber befand sich auf einer der Schubladen in dem Raum, in dem die 100,- DM gestohlen wurden? War es ...

❍ ein Großbritannien-Aufkleber?
❍ ein Portugal-Aufkleber?
❍ ein Deutschland-Aufkleber?
❍ ein Frankreich-Aufkleber?
❍ ein Italien-Aufkleber?
❍ ein Spanien-Aufkleber?

7. Welches Sportgerät hing an der Wand des Raumes, in dem die 100,- DM gestohlen wurden? War es ...

❍ ein Tennisschläger?
❍ ein Schlittschuh?
❍ eine Taucherbrille?
❍ ein Golfschläger?
❍ eine Dartscheibe?
❍ ein Springseil?

8. Welche Tiere waren auf dem Bild in dem Raum, in dem die 100,- DM gestohlen wurden? Waren es ...

❍ Delphine?
❍ Schafe?
❍ Hirsche?
❍ Kühe?
❍ Ziegen?
❍ Hühner?

9. Welches Kleidungsstück befand sich in dem Holzschrank des Raumes, in dem die 100,- DM gestohlen wurden? War es ...

❍ ein Hemd?
❍ eine Jacke?
❍ ein Kleid?
❍ ein Pullover?
❍ eine Hose?
❍ eine Krawatte?

10. Welche Farbe hatte der Teppich in dem Raum, in dem die 100,- DM gestohlen wurden? War er ...

❍ blau?
❍ weiß?
❍ gelb?
❍ grün?
❍ rot?
❍ schwarz?

Bitte umblättern.

Haben Sie bei der **mündlichen Vernehmung** durch die gerichtspsychologische Expertin irgendeine **Strategie, Taktik oder Technik** angewendet, um Ihre Aussage glaubwürdiger erscheinen zu lassen? Bitte ankreuzen:

nein	**ja**
0	1

Wenn **ja**: welche?

Haben Sie bei bei dem **Lügendetektortest** irgendeine **Strategie, Taktik oder Technik** angewendet, um einen unschuldigen Eindruck zu hinterlassen? Bitte ankreuzen:

nein	**ja**
0	1

Wenn **ja**: welche?

Bitte umblättern.

Anhang B: Abbildungen vom Tatort

Anhang B.1: Tatort in den Bedingungen *Täterinnen* und *Zeuginnen*
Anhang B.2: Tatort in der Bedingung *falsche Zeuginnen*

Anhang B.1: Tatort in den Bedingungen *Täterinnen* und *Zeuginnen*

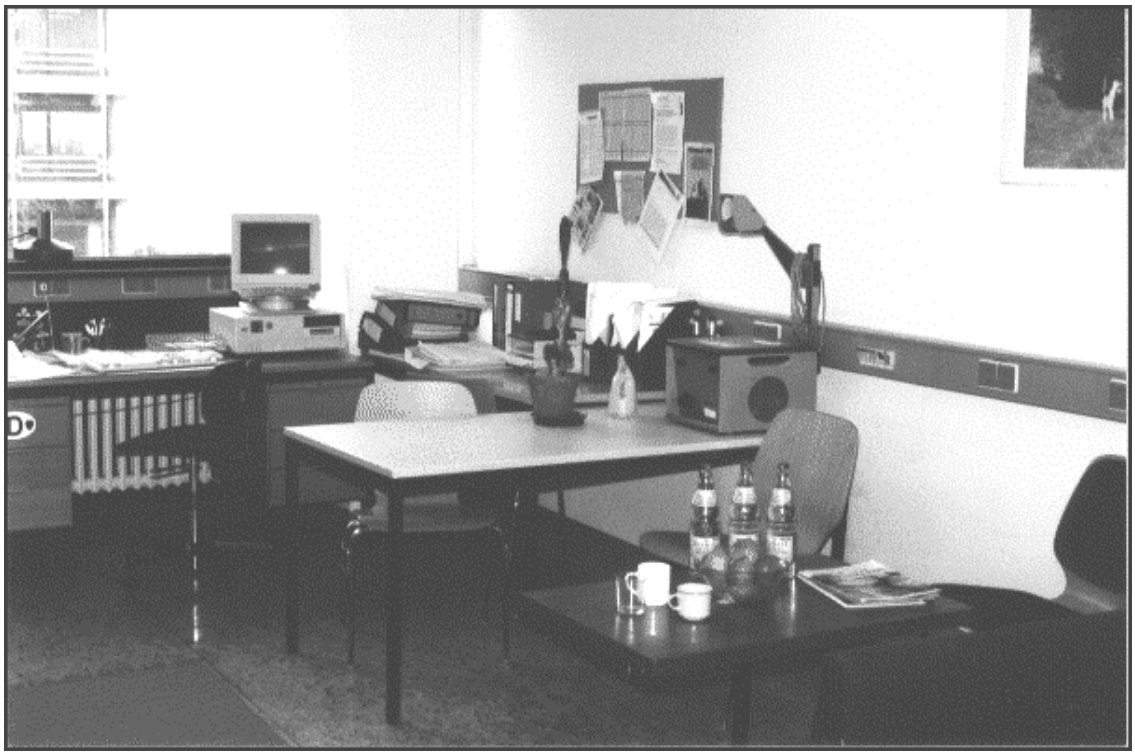


Abbildung B.1.1. Ansicht des Tatorts für die *Täterinnen* und *Zeuginnen*.

Abbildung B.1.2. Tatortdetail „gelbes Fahrrad“.

Abbildung B.1.3. Tatortdetail „Wasserkasten“.

Abbildung B.1.4. Tatortdetail „Porzellanhund“.

Abbildung B.1.5. Tatortdetail „Kaktus“.

Abbildung B.1.6. Tatortdetail „Obstschale mit Äpfeln“.

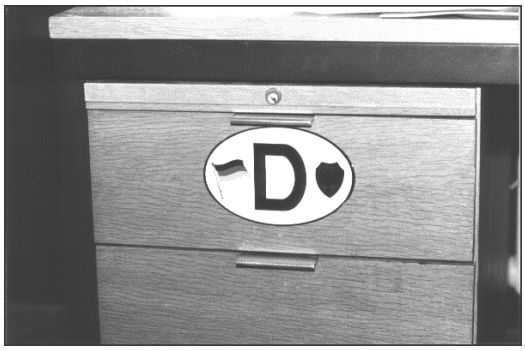


Abbildung B.1.7. Tatortdetail „Deutschland-Aufkleber“.

Abbildung B.1.8. Tatortdetail „Dartscheibe“.

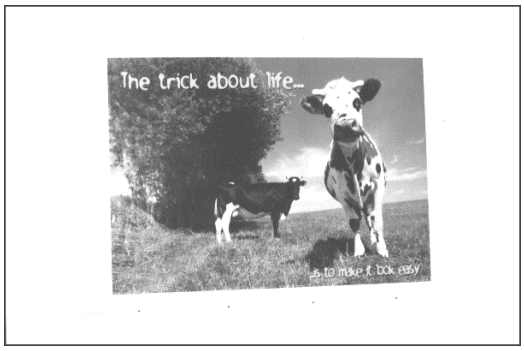


Abbildung B.1.9. Tatortdetail „Bild mit Kühen“.

Abbildung B.1.10. Tatortdetail „Jacke“.

Abbildung B.1.11. Tatortdetail „roter Teppich“.

Anhang B.2: Tatort in der Bedingung *falsche Zeuginnen*

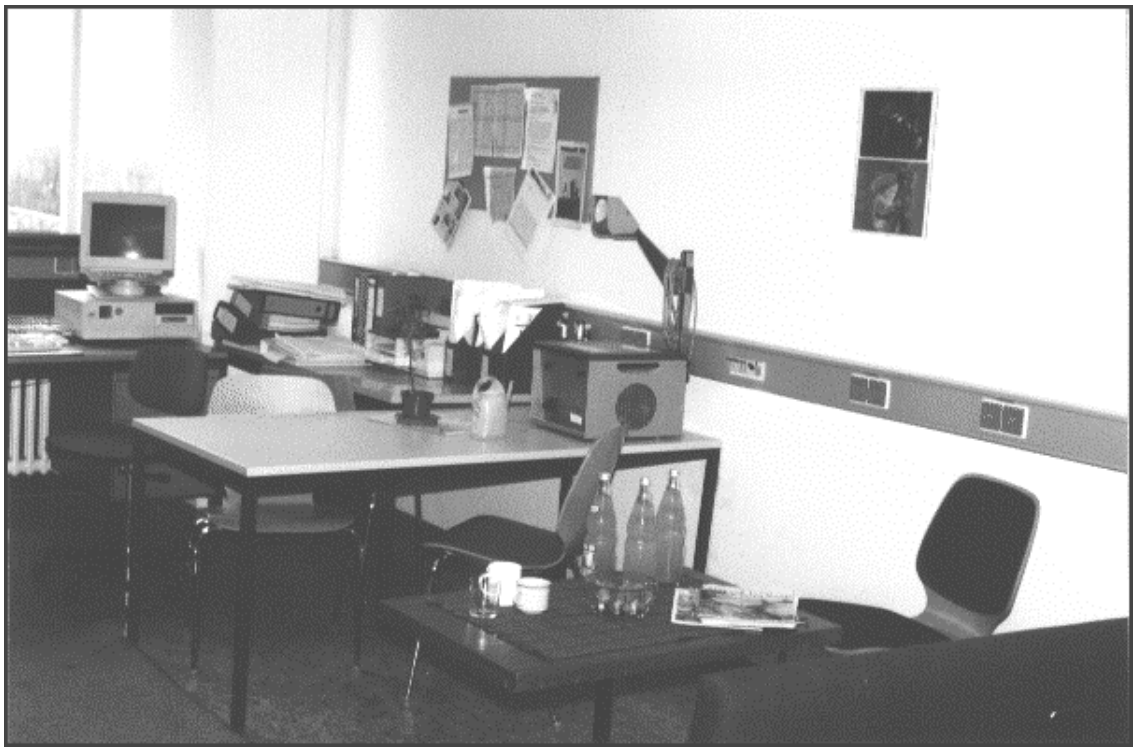


Abbildung B.2.1. Ansicht des Tatorts für die *falschen Zeuginnen*.

Abbildung B.2.2. Tatortdetail „silbernes Fahrrad“.

Abbildung B.2.3. Tatortdetail „Limonadenkasten“.

Abbildung B.2.4. Tatortdetail „Porzellanelefant“.

Abbildung B.2.5. Tatortdetail „Rose“.

Abbildung B.2.6. Tatortdetail „Obstschale mit Nüssen“.

Abbildung B.2.7. Tatortdetail „Schweiz-Aufkleber“.

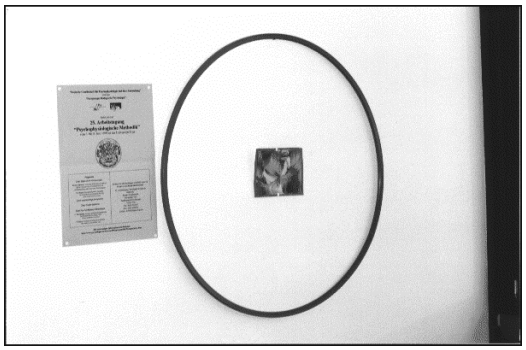


Abbildung B.2.8. Tatortdetail „Gymnastikreifen“.

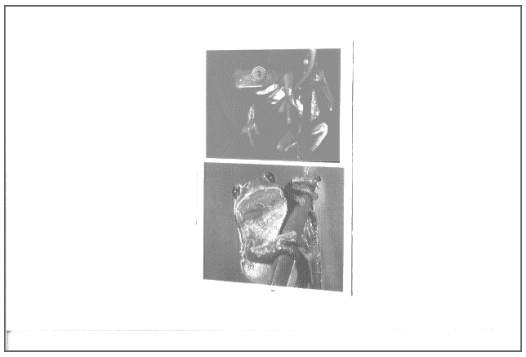


Abbildung B.2.9. Tatortdetail „Bild mit Fröschen“.

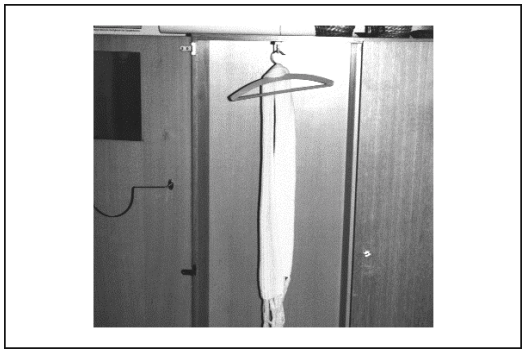


Abbildung B.2.10. Tatortdetail „Schal“.

Abbildung B.2.11. Tatortdetail „grauer Teppich“.

Anhang C: Protokoll- und Auswertungsbogen[a]

Anhang C.1: Protokollbogen für Versuchsleiter 2

Anhang C.2: Ratingbogen für die inhaltsorientierte Glaubhaftigkeitsbeurteilung

Anhang C.3: Ratingbogen für die naive Glaubhaftigkeitsbeurteilung

[a] Aus drucktechnischen Gründen wurde die Schriftgröße für den Anhang geändert.

Anhang C.1: Protokollbogen für Versuchsleiter 2 (Auszug)

VI:_____ **vp**____**g**

- **Aussageanalytische Exploration - Bemerkungen:**

- **Soziodemographische Daten:**

Alter:	**Muttersprache:**	**Dominante Hand:** ❍ Rechtshänder ❍ Linkshänder
Studienfach / Beruf:		**Semesterzahl:**
Hobbys:		
Sehschwächen (farbenblind?):	**Korrektur** (Brille o.ä.?):	
Waren Sie vor dieser Untersuchung auch schon einmal Zeugin eines Diebstahls? ❍ nein ❍ ja		
Erfahrung mit psychol. Exp.? ❍ nein ❍ ja	ggf. **Häufigkeit der Teilnahme**:	
Frühere Teilnahme an einer Untersuchung zur Lügendetektion bzw. zur Glaubwürdigkeitsbegutachtung ? ❍ nein ❍ ja Wenn ja: Worum ging es in dieser Untersuchung genau?		
Vorkenntnisse zur Lügendetektion bzw. Glaubwürdigkeitsbegutachtung? ❍ nein ❍ ja Wenn ja: Welche?		
Art der Anwerbung (Aushang, Ansprechen etc.):		
Befinden in der Kabine (Platzangst o.ä.):		
sonstige Anmerkungen:		

- **Ruhemessung - Bemerkungen:**

Temp. / Luftfeuchtigkeit zu Beginn: _____ / _____ Anzahl NSRs: _____ Sonstiges:

VI:_____ **vp**____**g**

- **Zahlentest: (A U S G E W Ä H L T E Z A H L A N K R E U Z E N ! ! !)**

Trial	Wortlaut	Antwort	Bemerkungen
1	Bezüglich der Zahl, die Sie aufgeschrieben haben: Haben Sie die Zahl 2 aufgeschrieben?	nein	
2	Haben Sie die Zahl 3 aufgeschrieben?	nein	
3	Haben Sie die Zahl 4 aufgeschrieben?	nein	
4	Haben Sie die Zahl 5 aufgeschrieben?	nein	
5	Haben Sie die Zahl 6 aufgeschrieben?	nein	
6	Haben Sie die Zahl 7 aufgeschrieben?	nein	
7	Haben Sie die Zahl 8 aufgeschrieben?	nein	

- **Probedurchgang:**

Trial	Wortlaut	Antwort	Bemerkungen
1	Die 100,- DM wurden aus dem Büro einer ganz bestimmten Person entwendet.		
2	Welche Person war Inhaber des Büros, in dem Sie die 100,- DM gestohlen haben? War es ...		
3	Prof. Schuster?	nein	
4	Prof. Hinze?	nein	
5	Prof. Meyer?	nein	
6	Prof. Müller ?	nein	
7	Prof. Kunze?	nein	
8	Prof. Bauer?	nein	

VI:_____ **vp**____**g**

V I D E O A U F N A H M E ! ! !

- **GAT:**

Trial	Wortlaut	Antwort	Bemerkungen
1	In dem Raum, in dem die 100,- DM gestohlen wurden, befand sich ein Fahrrad mit einer ganz bestimmten Farbe.		
2	Welche Farbe hatte das Fahrrad in dem Raum, in dem Sie die 100,- DM gestohlen haben? War es ...		
3	rot?	nein	
4	weiß?	nein	
5	blau?	nein	

- **Sonstige Bemerkungen:**

Temp. / Luftfeuchtigkeit am Ende der Untersuchung: _____ / _____

Anhang C.2: Ratingbogen für die inhaltsorientierte Glaubhaftigkeitsbeurteilung

Aussage-Nr.: **Rater:**

Häufigkeitstabelle der Signierungen im Aussagetext

	schwach vorhanden 1	mittel vorhanden 2	stark vorhanden 3
4. Kontextuelle Einbettung			
5. Beschreibung nonverbaler Interaktionen			
6. Wiedergabe von Gesprächen			
7. Komplikationen im Handlungsverlauf			
8. Ausgefallene Einzelheiten			
9. Nebensächliche (überflüssige) Einzelheiten			
10. Phänomengemäße Schilderung unverstand. Handl.-elemente			
11. Inhaltliche Verschachtelungen			
12. Schilderung eigenpsychischer Vorgänge			
13. Schilderung psychischer Vorgänge anderer Beteiligter			
14. Spontane Verbesserungen der eigenen Aussage			
15. Zugeben von Erinnerungslücken			
16. Einwände gegen die Richtigkeit der eigenen Aussage			
17. Selbstbelastungen der aussagenden Person			
18. Inschutznahme des Täters			

Anzahl der Details	

Gesamtskalierung

1.	**Logische Konsistenz & Widerspruchsfreiheit**	nicht vorhanden **0**	schwach vorhanden **1**	mittel vorhanden **2**	stark vorhanden **3**
2.	**Inkontinenz der Aussage (Unstrukturierte Darstellung)**	nicht vorhanden **0**	schwach vorhanden **1**	mittel vorhanden **2**	stark vorhanden **3**
3.	**Quantitativer Detailreichtum**	nicht vorhanden **0**	schwach vorhanden **1**	mittel vorhanden **2**	stark vorhanden **3**

4.	**Kontextuelle Einbettung**	nicht vorhanden **0**	schwach vorhanden **1**	mittel vorhanden **2**	stark vorhanden **3**
5.	**Beschreibung nonverbaler Interaktionen**	nicht vorhanden **0**	schwach vorhanden **1**	mittel vorhanden **2**	stark vorhanden **3**
6.	**Wiedergabe von Gesprächen**	nicht vorhanden **0**	schwach vorhanden **1**	mittel vorhanden **2**	stark vorhanden **3**
7.	**Komplikationen im Handlungsverlauf**	nicht vorhanden **0**	schwach vorhanden **1**	mittel vorhanden **2**	stark vorhanden **3**
8.	**Ausgefallene Einzelheiten**	nicht vorhanden **0**	schwach vorhanden **1**	mittel vorhanden **2**	stark vorhanden **3**
9.	**Nebensächliche (überflüssige) Einzelheiten**	nicht vorhanden **0**	schwach vorhanden **1**	mittel vorhanden **2**	stark vorhanden **3**
10.	**Phänomengemäße Schild. unverstand. Handl.-elemente**	nicht vorhanden **0**	schwach vorhanden **1**	mittel vorhanden **2**	stark vorhanden **3**
11.	**Inhaltliche Verschachtelungen**	nicht vorhanden **0**	schwach vorhanden **1**	mittel vorhanden **2**	stark vorhanden **3**
12.	**Schilderung eigenpsychischer Vorgänge**	nicht vorhanden **0**	schwach vorhanden **1**	mittel vorhanden **2**	stark vorhanden **3**
13.	**Schilderung psychischer Vorgänge anderer Beteiligter**	nicht vorhanden **0**	schwach vorhanden **1**	mittel vorhanden **2**	stark vorhanden **3**
14.	**Spontane Verbesserungen der eigenen Aussage**	nicht vorhanden **0**	schwach vorhanden **1**	mittel vorhanden **2**	stark vorhanden **3**
15.	**Zugeben von Erinnerungslücken**	nicht vorhanden **0**	schwach vorhanden **1**	mittel vorhanden **2**	stark vorhanden **3**
16.	**Einwände gegen die Richtigkeit der eigenen Aussage**	nicht vorhanden **0**	schwach vorhanden **1**	mittel vorhanden **2**	stark vorhanden **3**
17.	**Selbstbelastungen der aussagenden Person**	nicht vorhanden **0**	schwach vorhanden **1**	mittel vorhanden **2**	stark vorhanden **3**
18.	**Inschutznahme des Täters**	nicht vorhanden **0**	schwach vorhanden **1**	mittel vorhanden **2**	stark vorhanden **3**

- Ich halte die Aussage für:

1 2 3 4 5 6 7 8 9 10

Unglaubwürdig **glaubwürdig**

- In meinem Urteil bin ich mir:

1 2 3 4 5 6 7 8 9 10

sehr unsicher **sehr sicher**

Anhang C.3: Ratingbogen für die naive Glaubhaftigkeitsbeurteilung

Liebe Versuchsteilnehmerin / Lieber Versuchsteilnehmer,

Im folgenden sehen Sie 15 Videoaufzeichnungen von **Zeugenaussagen**. Die Zeugenaussagen beziehen sich allesamt auf einen Diebstahl, der im Rahmen eines Experiments simuliert wurde.

Bei den aussagenden Personen handelt es sich teilweise um **echte Zeuginnen**, d.h. die Personen haben das, was sie erzählen, tatsächlich erlebt.

Teilweise handelt es sich bei den aussagenden Personen aber auch um **falsche „Zeuginnen“**, d.h. die Zeugenaussagen sind erfunden.

Ihre Aufgabe besteht darin, die **Glaubwürdigkeit** der Zeugenaussagen zu **beurteilen**.

Verwenden Sie dazu die vorgegebene **8-stufige Beurteilungsskala** (s. Muster).
Kreuzen Sie jeweils an, für wie glaubwürdig bzw. unglaubwürdig Sie die Aussagen halten.

Muster der Beurteilungsskala:

Ich halte die Aussage für:

1	2	3	4	5	6	7	8
äußerst unglaubwürdig	sehr unglaubwürdig	ziemlich unglaubwürdig	eher unglaubwürdig	eher glaubwürdig	ziemlich glaubwürdig	sehr glaubwürdig	äußerst glaubwürdig

Nachdem Sie Ihr Urteil auf der Skala abgegeben haben, **beschreiben Sie bitte auch noch grob in Stichworten, worauf sich Ihre Einschätzung stützt**.

An dieser Untersuchung nehmen insgesamt 20 Personen teil.
Derjenige Beurteiler, der die Aussagen insgesamt am zutreffendsten beurteilt, erhält am Ende der Untersuchungsreihe eine Belohnung von 20,- DM!

Beurteiler Nr.: _____

Zeugenaussage Nr.: _____

Ich halte die Aussage für:

1	2	3	4	5	6	7	8
äußerst unglaubwürdig	sehr unglaubwürdig	ziemlich unglaubwürdig	eher unglaubwürdig	eher glaubwürdig	ziemlich glaubwürdig	sehr glaubwürdig	äußerst glaubwürdig

Meine Einschätzung stützt sich auf:

Zeugenaussage Nr.: _____

Ich halte die Aussage für:

1	2	3	4	5	6	7	8
äußerst unglaubwürdig	sehr unglaubwürdig	ziemlich unglaubwürdig	eher unglaubwürdig	eher glaubwürdig	ziemlich glaubwürdig	sehr glaubwürdig	äußerst glaubwürdig

Meine Einschätzung stützt sich auf:

Zeugenaussage Nr.: _____

Ich halte die Aussage für:

1	2	3	4	5	6	7	8
äußerst unglaubwürdig	sehr unglaubwürdig	ziemlich unglaubwürdig	eher unglaubwürdig	eher glaubwürdig	ziemlich glaubwürdig	sehr glaubwürdig	äußerst glaubwürdig

Meine Einschätzung stützt sich auf:

Anhang D: Abbildungen zu den psychophysiologischen Untersuchungen

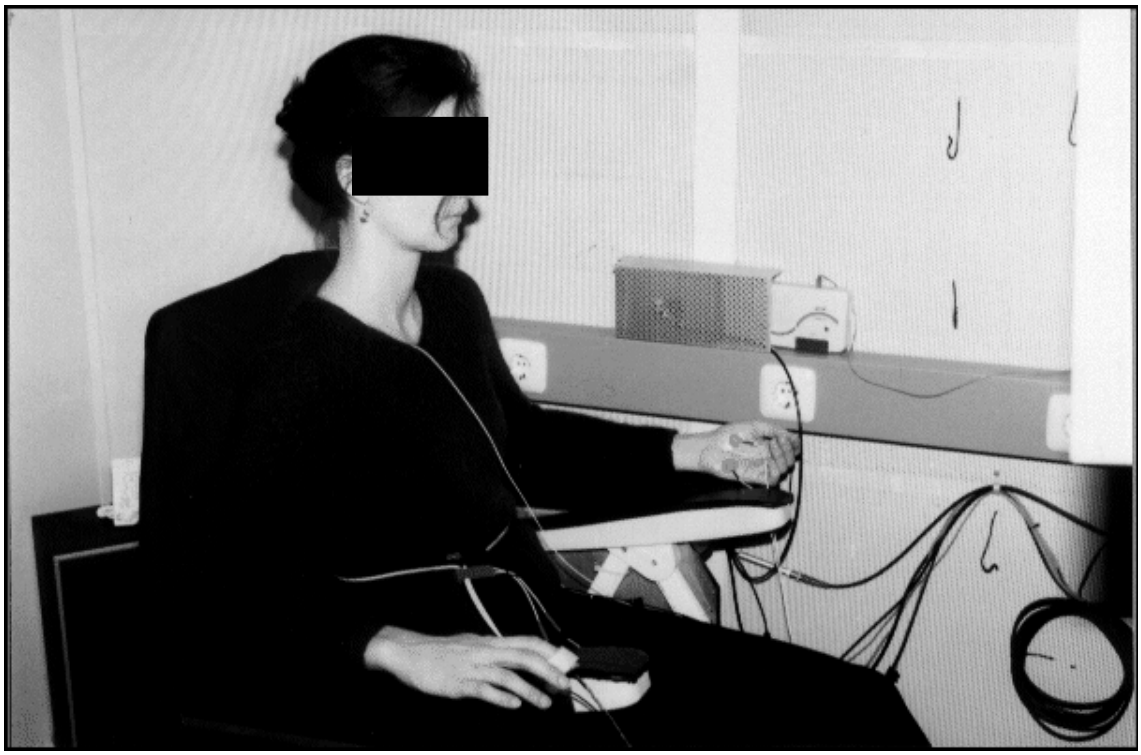


Abbildung D.1. Versuchsperson in Meßkabine.

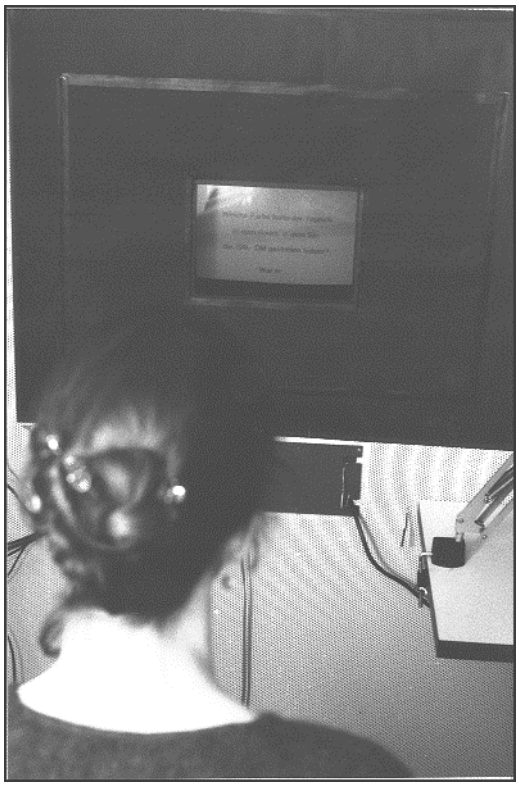


Abbildung D.2. Reizpräsentations-monitor aus Sicht der Versuchsperson

Anhang E: Beschreibung des Auswertertrainings zur *Kriterienorientierten Inhaltsanalyse*

Das Ratertraining wurde in der Zeit vom 03.04. bis 17.04.2000 an sieben jeweils zwei- bis dreistündigen Sitzungsterminen im Psychologischen Institut der Universität Mainz durchgeführt. Es wurde vom Verfasser dieser Arbeit geleitet und war nach dem Vorbild des ***Kieler Trainingsprogramms zur Beurteilung der Glaubwürdigkeit von Zeugenaussagen*** (***KTBG***; Krause, 1997; Petersen, 1997; vgl. auch Höfer et al., 1999) konzipiert. Allerdings war es aus organisatorischen Gründen gegenüber dem KTBG zeitlich stark gestrafft. An der Schulung nahmen drei Rater teil.

Dem eigentlichen praktischen Anwendungstraining war ein Literaturstudium vorgeschaltet. Hierfür wurde den Trainingsteilnehmern drei Wochen vor Beginn der praktischen Übungsphase folgende Literatur ausgehändigt, die in Einzelarbeit zu lesen war und einen umfassenden Eindruck von der inhaltsorientierten Glaubhaftigkeitsbeurteilung vermitteln sollte:

- Köhnken, G. (1990). *Glaubwürdigkeit: Untersuchungen zu einem psychologischen Konstrukt*. München: Psychologie Verlags Union. (S. 82–117)
- Krahé, B. & Kundrotas, S. (1992). Glaubwürdigkeitsbeurteilung bei Vergewaltigungsanzeigen: Ein aussagenanalytisches Feldexperiment. *Zeitschrift für Experimentelle und Angewandte Psychologie, 39*, 598–620.
- Littmann, E. & Szewczyk, H. (1983). Zu einigen Kriterien und Ergebnissen forensisch-psychologischer Glaubwürdigkeitsbegutachtung von sexuell mißbrauchten Kindern und Jugendlichen. *Forensia, 4*, 55–72.
- Wellershaus, P. (1992). Glaubhaftigkeit kindlicher Zeugenaussagen. *Psychomed, 4*, 20–24.
- Steller, M. & Köhnken, G. (1989). Criteria-based statement analysis. Credibility assessment of children´s statements in sexual abuse cases. In D. C. Raskin (Hrsg.), *Psychological methods in criminal investigation and evidence* (S. 217–245). New York: Springer.
- Steller, M., Wellershaus, P. & Wolf, T. (1992). Realkennzeichen in Kinderaussagen: Empirische Grundlagen der kriterienorientierten Aussageanalyse. *Zeitschrift für experimentelle und angewandte Psychologie, 39*, 151–179.
- Wolf, P. & Steller, M. (1997). Realkennzeichen in Aussagen von Frauen. Zur Validierung der Kriterienorientierten Aussageanalyse für Zeugenaussagen von Vergewaltigungsopfern. In L. Greuel, T. Fabian & M. Stadler (Hrsg.), *Psychologie der Zeugenaussage* (S. 121–130). Weinheim: Psychologie Verlags Union.

Das Training selbst war in einzelne, aufeinander aufbauende und zunehmend schwieriger werdende Übungsschritte aufgegliedert. Die Grundlage des Trainings bildete eine ausführliche Operationalisierung der Realkennzeichen. Die Operationalisierungsvorschriften zu den 19 Glaubhaftigkeitskriterien wurden wörtlich aus dem *KTBG* übernommen. Darüber hinaus beinhaltete die Raterschulung **fünf** wesentliche praktische **Übungselemente**, die sich mit Petersen (1997) und Krause (1997) als Wiedererkennungs-, Simultan-, Skalierungs-, Anker- sowie Herstellungstraining charakterisieren lassen.

Beim **Wiedererkennungstraining** geht es darum, einzelne Glaubhaftigkeitskriterien im Übungsmaterial zu identifizieren, wobei der Ausprägungsgrad der Kriterien irrelevant ist. Entsprechend sollen beim **Simultantraining** mehrere Kriterien gleichzeitig und ohne Bezugnahme zum Ausprägungsgrad im Übungsmaterial wiedererkannt werden. Ziel des **Skalierungstrainings** ist es, unterschiedliche Ausprägungsgrade der Glaubhaftigkeitskriterien zu erkennen bzw. durch eine korrekte Verknüpfung von Intensität und Häufigkeit kriterienerfüllender Textstellen übereinstimmende Einschätzungen der Ausprägungsgrade zu erzielen. Beim **Ankertraining** gilt es, mehrere Aussagen zunächst zu lesen und anschließend in jeder Aussage die Glaubhaftigkeitskriterien zu skalieren. **Herstellungstraining** bedeutet, daß die Trainingsteilnehmer selbst Aussagen produzieren, die vorgegebene Kriterien in vorgegebenen Ausprägungsgraden aufweisen.

Bei der Durchführung des Ratertrainings wurde darauf geachtet, eine durch zu frühzeitige Beendigung eines Trainingselements bzw. durch zu schnelle Steigerung des Schwierigkeitsgrades bedingte Überforderung und damit verbundene Frustrationserlebnisse und Motivationsverluste seitens der Trainingsteilnehmer zu vermeiden. Während des gesamten Trainings erhielten die Teilnehmer Rückmeldungen über ihren Leistungsstand. Diskrepanzen zwischen den Einschätzungen der einzelnen Rater wurden ausführlich diskutiert. Die Aussagen, die im Training als Anschauungs- bzw. Übungsmaterial dienten, waren dem *KTBG* (Krause, 1997; Petersen, 1997) und der Diplomarbeit von Scheinberger (1993) entnommen.

Im folgenden ist der **Ablauf** des Ratertrainings stichwortartig beschrieben:

1. Sitzung (03.04.2000)**:**

- Beschreibung des Trainingsablaufs
- Besprechung der vorgegebenen Literatur; Erfragung des Wissens über die Kriterien und Vorstellung einzelner Kategorien der *Kriterienorientierten Inhaltsanalyse*
- Definition der Begriffe allgemeine und spezielle Glaubwürdigkeit, Täuschung und Lüge
- Besprechung der grundlegenden Hypothesen und Elemente der forensischen Glaubhaftigkeitsbeurteilung
- Allgemeine Informationen über Beurteilungsfehler
- Grobe Definition der Glaubhaftigkeitskriterien
- Besprechung ersten Übungsmaterials (Wiedererkennungstraining in der Gruppe)
- Vorstellung der Operationalisierungsvorschriften aus dem *KTBG* und Veranschaulichung durch Beispiele

2. Sitzung (04.04.2000)**:**

- Wiederholung der Operationalisierungsvorschriften
- Bearbeitung von Kurzbeispielen (Wiedererkennungstraining in Einzelarbeit)
- Besprechung der Kurzbeispiele in der Gruppe
- Hausaufgabe: Identifizierung der Kriterien in Übungstexten (Simultantraining)

3. Sitzung (05.04.2000)**:**

- Abfrage der Kategorien und Kriterien (Welche Kategorien gibt es? Welche Kriterien gehören hinein?); Erläuterung der Kriterien durch die Trainingsteilnehmer
- gemeinsame Besprechung der Hausaufgabe
- Herstellungstraining: Trainingsteilnehmer denken sich in Einzelarbeit Beispiele zu den einzelnen Kriterien aus und stellen diese anschließend in der Gruppe vor.
- Wiedererkennungstraining: Die jeweils anderen Trainingsteilnehmer sollen herausfinden, um welche Kriterien es sich handelt.
- Hausaufgabe: Identifizierung der Kriterien in Übungstexten (Simultantraining)

4. Sitzung (07.04.2000)**:**

- Abfrage der Kategorien und Kriterien; Diskussion von Unklarheiten bezüglich der Operationalisierungsvorschriften
- gemeinsame Besprechung der Hausaufgabe
- Festlegung der Vorgehensweise bei der Signierung der Aussagen (Wie sollen kriterienerfüllende Textstellen markiert werden, um eine optimale Nachvollziehbarkeit der Ratings zu gewährleisten?)
- Vorstellung der Skalierungsregeln (aus dem *KTBG* übernommen)
- Hausaufgabe: Identifizierung und Skalierung der Kriterien in einem Übungstext (Skalierungstraining)

5. Sitzung (10.04.2000)**:**

- Abfrage der Kategorien und Kriterien; Diskussion von Unklarheiten bezüglich der Operationalisierungsvorschriften
- Abfrage der Vorgehensweise beim Signieren
- Abfrage der Skalierungsregeln
- Skalierungstraining anhand von Kurzbeispielen (Einzelarbeit)
- Besprechung der Kurzbeispiele in der Gruppe
- gemeinsame Besprechung der Hausaufgabe
- Herstellungstraining: Jeder Trainingsteilnehmer greift sich 5 Kriterien (aus den Kriterien 4 – 18) heraus und erfindet dazu in Einzelarbeit jeweils 3 Beispiele unterschiedlicher Ausprägung.

- Besprechung und Diskussion der entwickelten Beispiele: Ein Teilnehmer stellt seine Beispiele vor. Die anderen Teilnehmer müssen das Kriterium erkennen und die entsprechenden Beispiele bezüglich der Kriterienausprägung in eine Rangreihe bringen.
- Hausaufgabe: Identifizierung und Skalierung der Kriterien in einem Übungstext (Skalierungstraining)

6. Sitzung (11.04.2000)**:**

- Durchführung eines Wissenstests zur *Kriterienorientierten Inhaltsanalyse* (entnommen aus dem *KTBG*)
- Abfrage der Kategorien und Kriterien; Diskussion von Unklarheiten bezüglich der Operationalisierungsvorschriften
- gemeinsame Besprechung der Hausaufgabe
- Herstellungstraining: Jeder Trainingsteilnehmer greift sich 5 Kriterien (aus den Kriterien 4 – 18) heraus und erfindet dazu in Einzelarbeit jeweils 3 Beispiele unterschiedlicher Ausprägung.
- Besprechung und Diskussion der entwickelten Beispiele: Ein Teilnehmer stellt seine Beispiele vor. Die anderen Teilnehmer müssen das Kriterium erkennen und den Ausprägungsgrad auf der vierstufigen Skala einordnen.
- Hausaufgabe: Identifizierung und Skalierung der Kriterien in mehreren Übungstexten (Ankertraining)

7. Sitzung (17.04.2000)**:**

- Abfrage der Kategorien und Kriterien; Diskussion von Unklarheiten bezüglich der Operationalisierungsvorschriften
- Rückmeldung der Ergebnisse aus dem Wissenstest zur *Kriterienorientierten Inhaltsanalyse*; Besprechung der Fehler
- gemeinsame Besprechung der Hausaufgabe
- abschließende Besprechung noch bestehender Unklareiten seitens der Trainingsteilnehmer

Anhang F: Statistischer Anhang

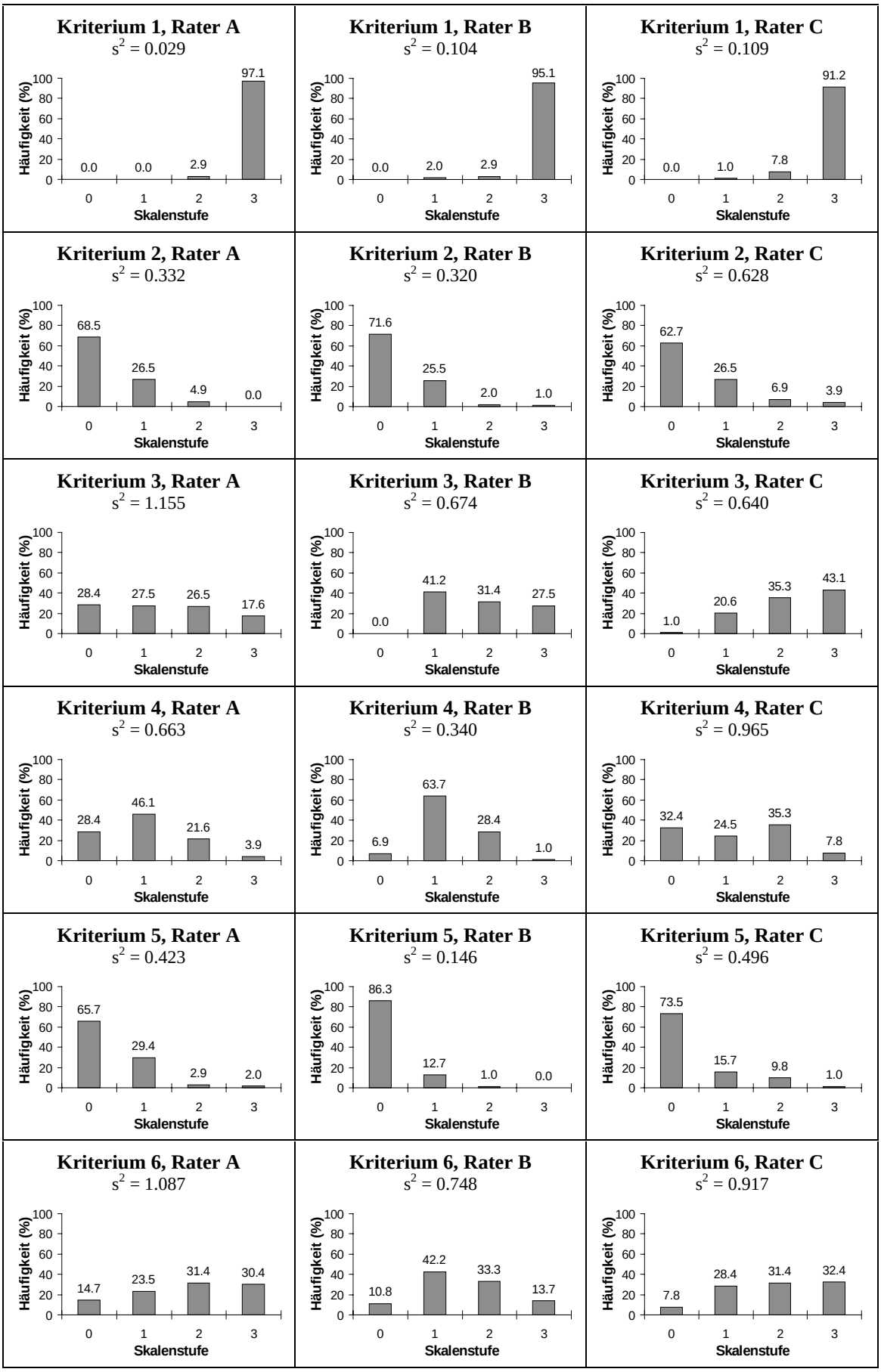


Abbildung F.1. Häufigkeitsverteilungen der Ratings.

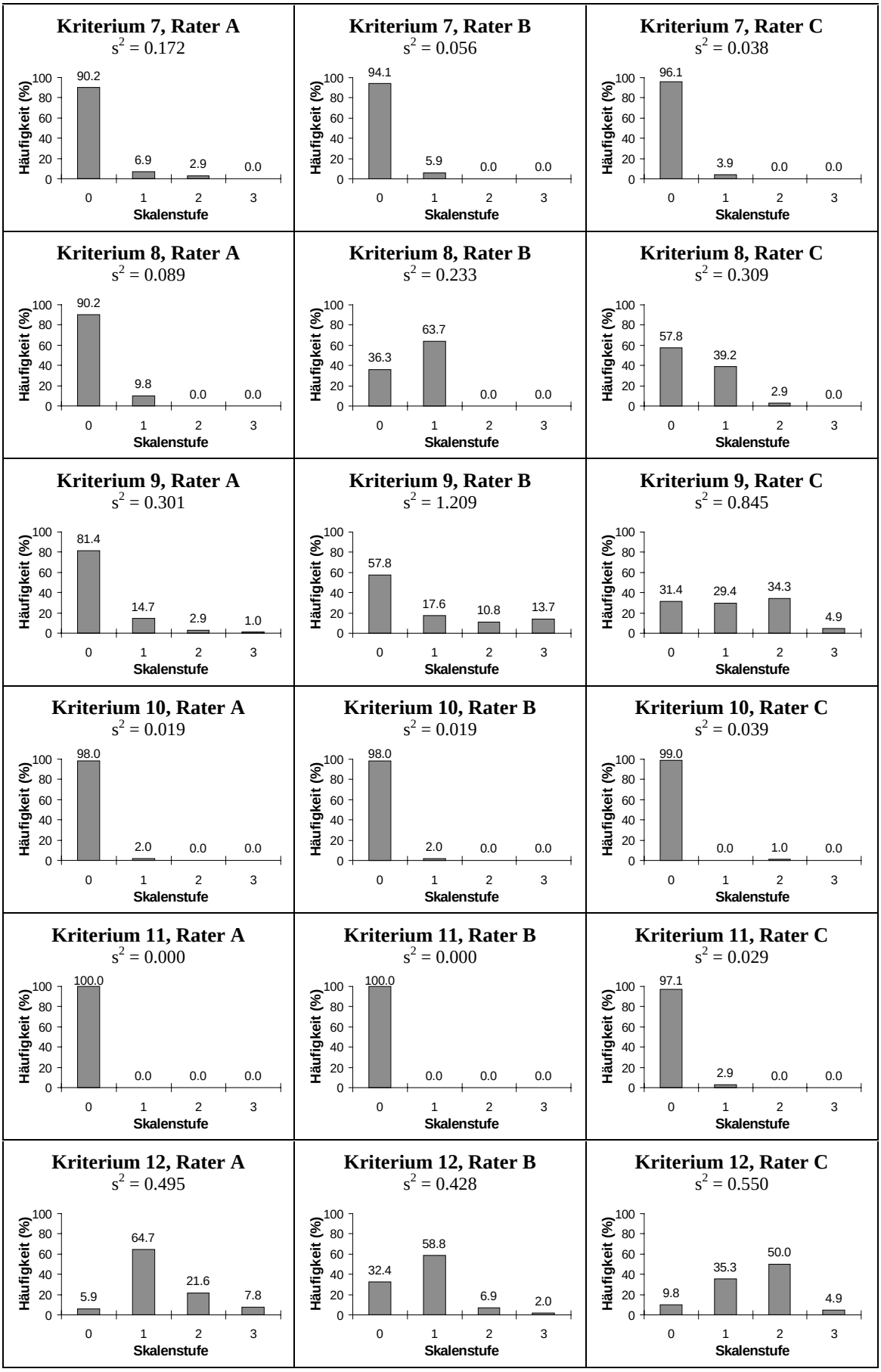


Abbildung F.1 (Fortsetzung). Häufigkeitsverteilungen der Ratings.

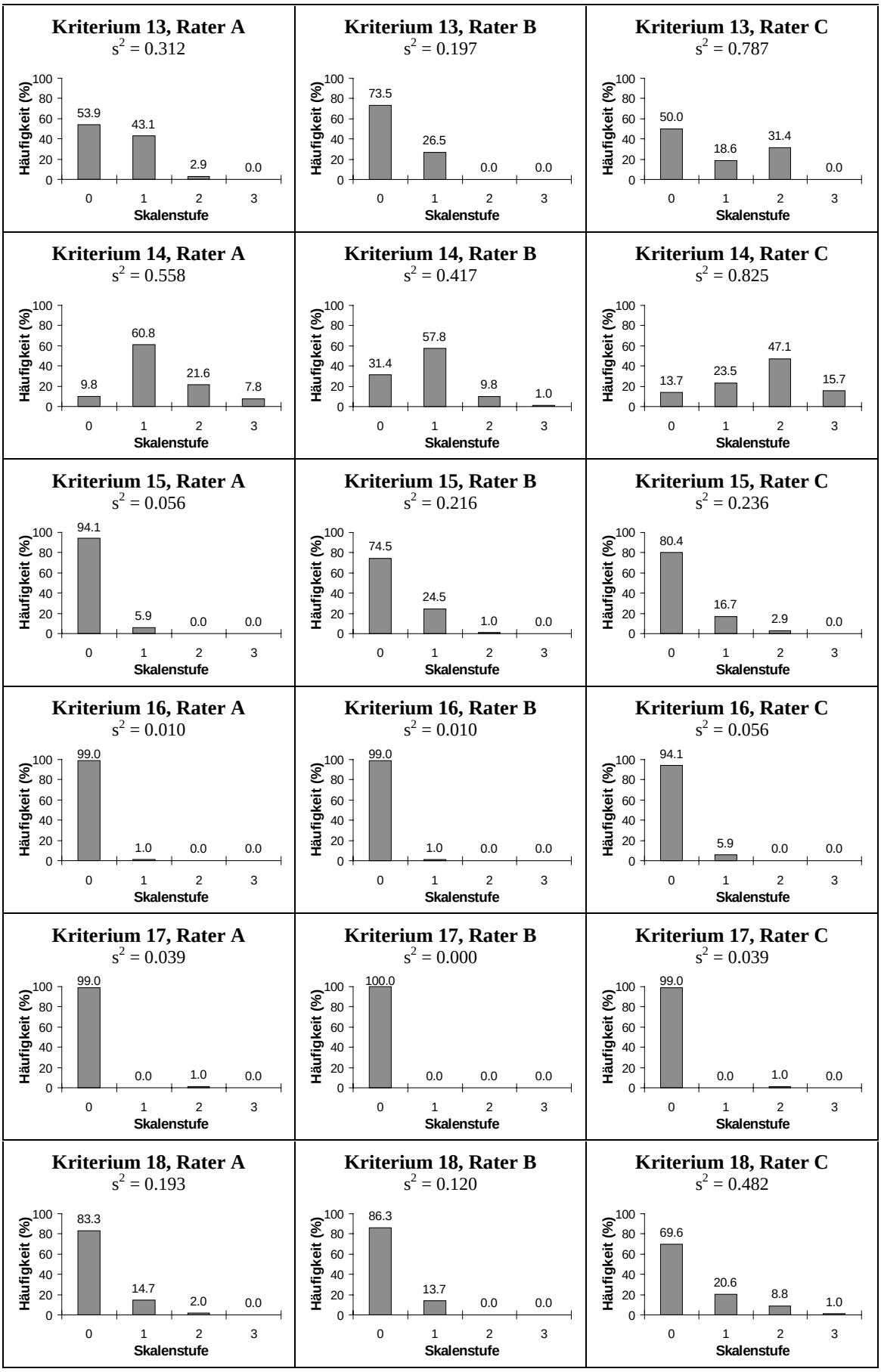


Abbildung F.1 (Fortsetzung). Häufigkeitsverteilungen der Ratings.

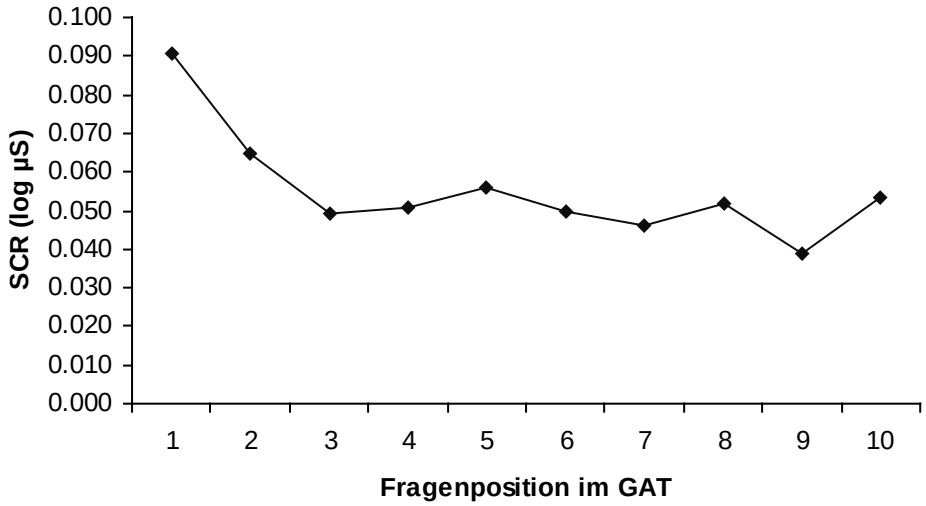


Abbildung F.2. SCR-Magnituden in Abhängigkeit von der Fragenposition im *GAT* (SCR-Quantifizierungsmethode A).

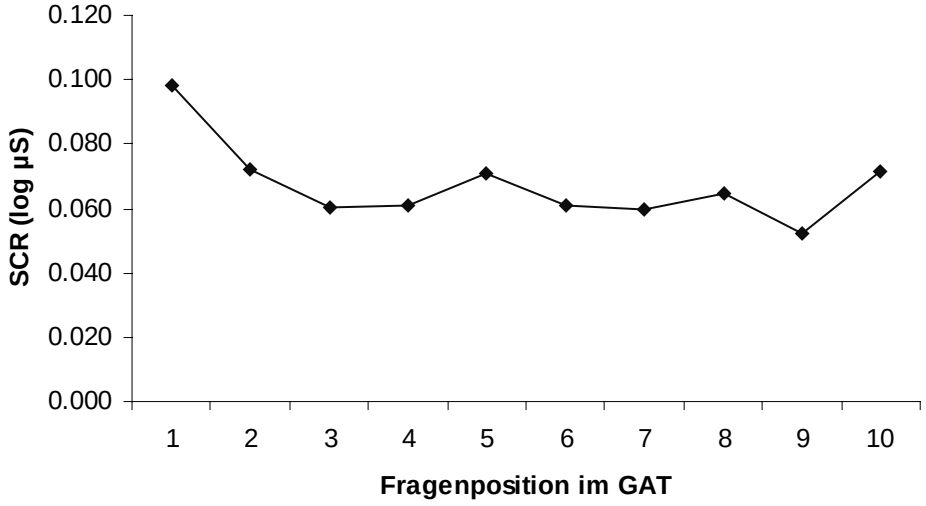


Abbildung F.3. SCR-Magnituden in Abhängigkeit von der Fragenposition im *GAT* (SCR-Quantifizierungsmethode B).

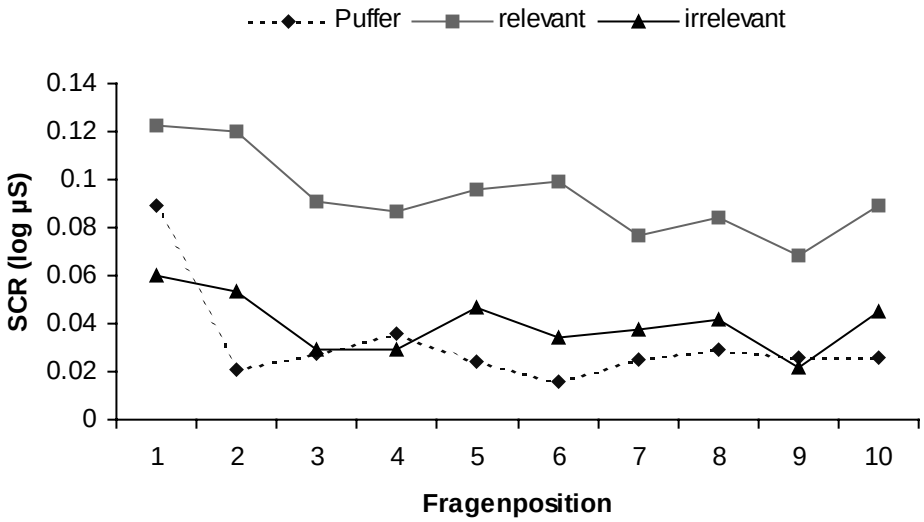


Abbildung F.4. SCR-Magnituden in Abhängigkeit von der Fragenposition und dem Itemtyp im *GAT* (SCR-Quantifizierungsmethode A).

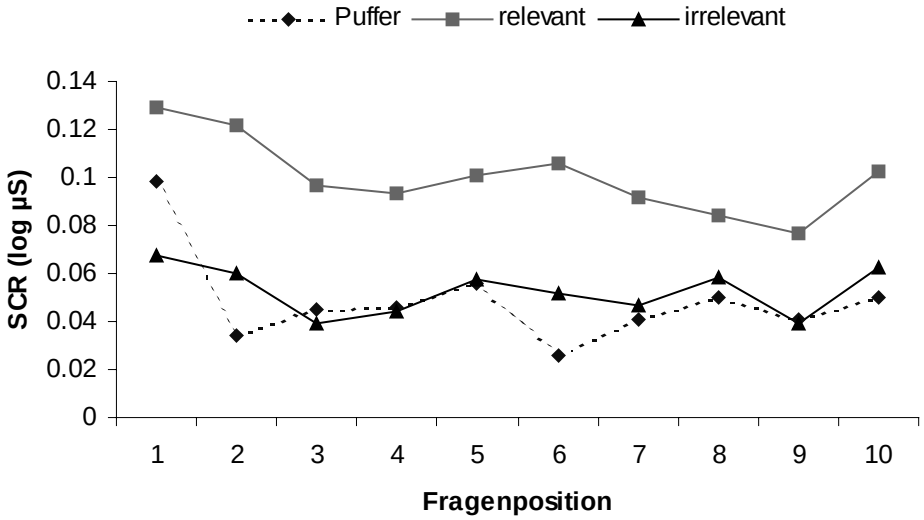


Abbildung F.5. SCR-Magnituden in Abhängigkeit von der Fragenposition und dem Itemtyp im *GAT* (SCR-Quantifizierungsmethode B).

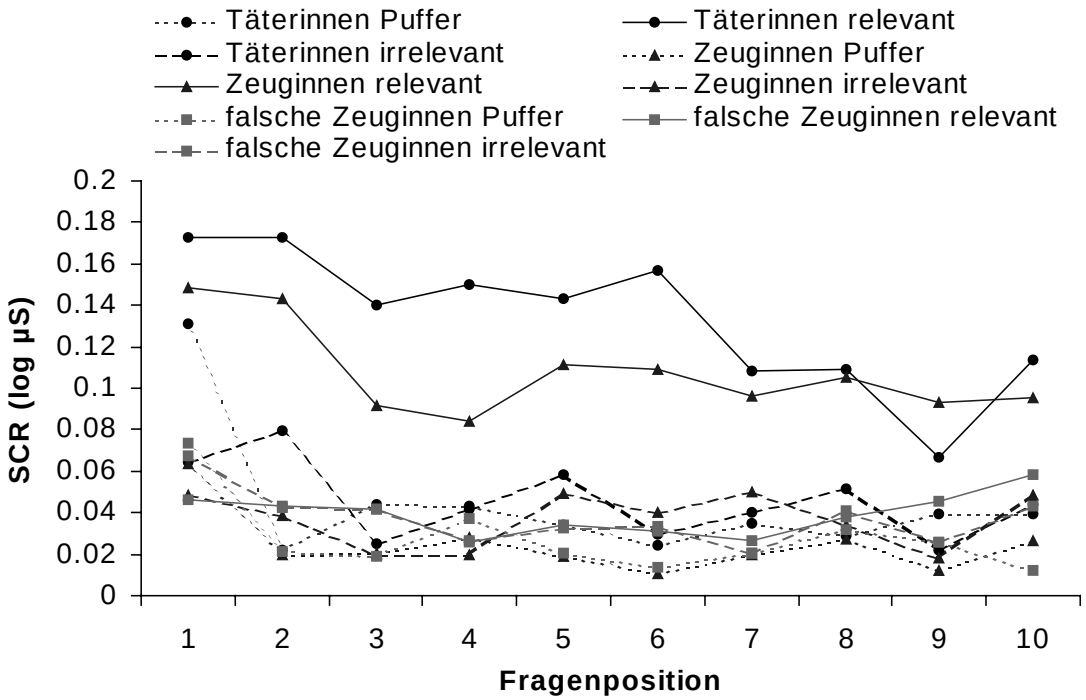


Abbildung F.6. SCR-Magnituden in Abhängigkeit von der experimentellen Gruppenzugehörigkeit, der Fragenposition und dem Itemtyp im *GAT* (SCR-Quantifizierungsmethode A).

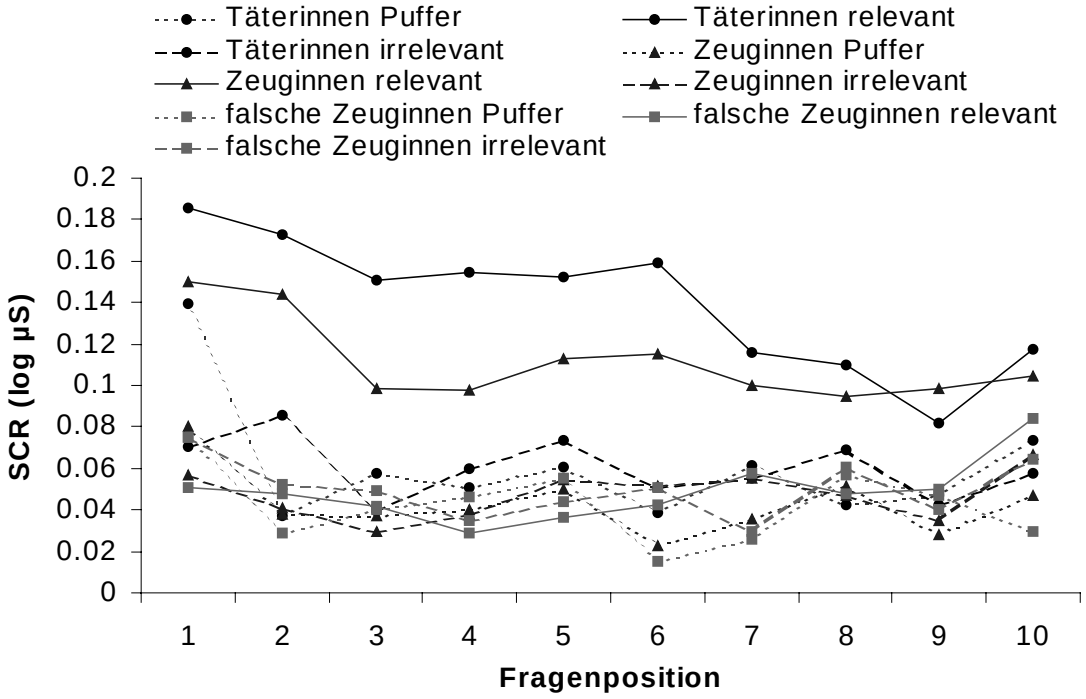


Abbildung F.7. SCR-Magnituden in Abhängigkeit von der experimentellen Gruppenzugehörigkeit, der Fragenposition und dem Itemtyp im *GAT* (SCR-Quantifizierungsmethode B).

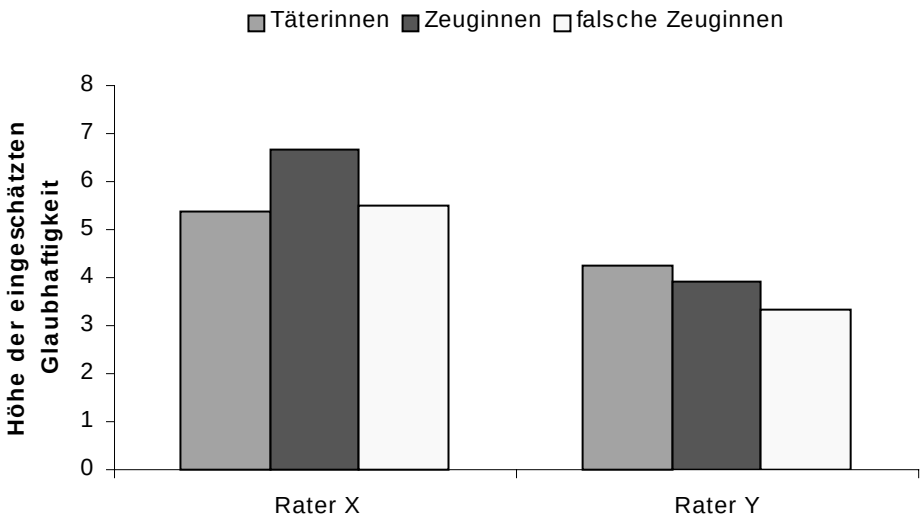


Abbildung F.8. Naive Beurteilung der Glaubhaftigkeit, getrennt nach Ratern und experimentellen Aussagegruppen (ergänzende Datenerhebung).

Tabelle F.1. Einfaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person als UV und der Anzahl elektrodermaler Spontanfluktuationen als AV

Q.d.V.	QS	df	MQ	F	p
Gruppe	848.020	2	424.010	1.223	.299
Fehler	34320.853	99	346.675		
Gesamt	35168.873	101			

Tabelle F.2. Einfaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person als UV und der Menge frei erinnerter kritischer Tatortdetails als AV

Q.d.V.	QS	df	MQ	F	p
Gruppe	2071.314	2	1035.657	4390.455	.000
Fehler	23.353	99	0.236		
Gesamt	2094.667	101			

Tabelle F.3. Scheffé-Tests für paarweise Vergleiche der experimentellen Gruppen bezüglich der Menge frei erinnerter kritischer Tatortdetails

Gruppenvergleich			Mittlere Differenz	Standard-fehler	p	95%-Konfidenzintervall Untergrenze	Obergrenze
Täterinnen	vs.	*Zeuginnen*	-0.118	0.118	.609	-0.410	0.175
Täterinnen	vs.	*falsche Zeuginnen*	9.500	0.118	.000	9.207	9.793
Zeuginnen	vs.	*falsche Zeuginnen*	9.618	0.118	.000	9.325	9.910

Tabelle F.4. Einfaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person als UV und der Menge wiedererkannter kritischer Tatortdetails als AV

Q.d.V.	QS	df	MQ	F	p
Gruppe	2014.137	2	1007.069	1910.819	.000
Fehler	52.176	99	0.527		
Gesamt	2066.314	101			

Tabelle F.5. Scheffé-Tests für paarweise Vergleiche der experimentellen Gruppen bezüglich der Menge wiedererkannter kritischer Tatortdetails

Gruppenvergleich			Mittlere Differenz	Standard-fehler	p	95%-Konfidenzintervall Untergrenze	Obergrenze
Täterinnen	vs.	*Zeuginnen*	-0.029	0.176	.986	-0.467	0.408
Täterinnen	vs.	*falsche Zeuginnen*	9.412	0.176	.000	8.974	9.849
Zeuginnen	vs.	*falsche Zeuginnen*	9.441	0.176	.000	9.004	9.879

Tabelle F.6. Einfaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person als UV und der Motivation im *GAT* als AV

Q.d.V.	QS	df	MQ	F	p
Gruppe	4.471	2	2.235	1.846	.163
Fehler	119.853	99	1.211		
Gesamt	124.324	101			

Tabelle F.7. Einfaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person als UV und der Motivation beim Ablegen der Zeugenaussage als AV

Q.d.V.	QS	df	MQ	F	p
Gruppe	0.314	2	0.157	0.220	.803
Fehler	70.559	99	0.713		
Gesamt	70.873	101			

Tabelle F.8. Einfaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person als UV und der subjektiv eingeschätzten persönlichen Erfolgswahrscheinlichkeit im *GAT* als AV

Q.d.V.	QS	df	MQ	F	p
Gruppe	24060.784	2	12030.392	22.020	.000
Fehler	54088.235	99	546.346		
Gesamt	78149.020	101			

Tabelle F.9. Scheffé-Tests für paarweise Vergleiche der experimentellen Gruppen bezüglich der subjektiv eingeschätzten persönlichen Erfolgswahrscheinlichkeit im *GAT*

Gruppenvergleich			Mittlere Differenz	Standard-fehler	p	95%-Konfidenzintervall Untergrenze	Obergrenze
Täterinnen	vs.	*Zeuginnen*	-33.824	5.669	.000	-47.913	-19.735
Täterinnen	vs.	*falsche Zeuginnen*	-31.176	5.669	.000	-45.265	-17.087
Zeuginnen	vs.	*falsche Zeuginnen*	2.647	5.669	.897	-11.442	16.736

Tabelle F.10. Einfaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person als UV und der subjektiv eingeschätzten persönlichen Erfolgswahrscheinlichkeit beim Ablegen der Zeugenaussage als AV

Q.d.V.	QS	df	MQ	F	p
Gruppe	13013.725	2	6506.863	15.590	.000
Fehler	41320.588	99	417.380		
Gesamt	54334.314	101			

Tabelle F.11. Scheffé-Tests für paarweise Vergleiche der experimentellen Gruppen bezüglich der subjektiv eingeschätzten persönlichen Erfolgswahrscheinlichkeit beim Ablegen der Zeugenaussage

Gruppenvergleich			Mittlere Differenz	Standard-fehler	p	95%-Konfidenzintervall Untergrenze	Obergrenze
Täterinnen	vs.	*Zeuginnen*	-27.059	4.955	.000	-39.373	-14.744
Täterinnen	vs.	*falsche Zeuginnen*	-8.529	4.955	.232	-20.844	3.785
Zeuginnen	vs.	*falsche Zeuginnen*	18.529	4.955	.001	6.215	30.844

Tabelle F.12. Einfaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person als UV und der Durchführung von Manipulationsmaßnahmen im *GAT* als AV

Q.d.V.	QS	df	MQ	F	p
Gruppe	3.196	2	1.598	7.249	.001
Fehler	21.824	99	0.220		
Gesamt	25.020	101			

Tabelle F.13. Scheffé-Tests für paarweise Vergleiche der experimentellen Gruppen bezüglich der Durchführung von Manipulationsmaßnahmen im *GAT*

Gruppenvergleich			Mittlere Differenz	Standard-fehler	p	95%-Konfidenzintervall Untergrenze	Obergrenze
Täterinnen	vs.	*Zeuginnen*	0.412	0.114	.002	0.129	0.695
Täterinnen	vs.	*falsche Zeuginnen*	0.324	0.114	.021	0.041	0.607
Zeuginnen	vs.	*falsche Zeuginnen*	-0.088	0.114	.741	-0.371	0.195

Tabelle F.14. Einfaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person als UV und der Durchführung von Manipulationsmaßnahmen beim Ablegen der Zeugenaussage als AV

Q.d.V.	QS	df	MQ	F	p
Gruppe	0.373	2	0.186	0.734	.482
Fehler	25.118	99	0.254		
Gesamt	25.490	101			

Tabelle F.15. Einfaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person als UV und der Leistung im *Mehrfachwahl-Wortschatz-Intelligenztest* als AV

Q.d.V.	QS	df	MQ	F	p
Gruppe	9.431	2	4.716	0.317	.729
Fehler	1470.647	99	14.855		
Gesamt	1480.078	101			

Tabelle F.16. Einfaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person als UV und der Variable Vorerfahrung mit Diebstahl als AV

Q.d.V.	QS	df	MQ	F	p
Gruppe	0.608	2	0.304	3.577	.032
Fehler	8.412	99	0.085		
Gesamt	9.020	101			

Tabelle F.17. Scheffé-Tests für paarweise Vergleiche der experimentellen Gruppen bezüglich der Variable Vorerfahrung mit Diebstahl

Gruppenvergleich			Mittlere Differenz	Standard-fehler	p	95%-Konfidenzintervall Untergrenze	Obergrenze
Täterinnen	vs.	*Zeuginnen*	-0.147	0.071	.120	-0.323	0.029
Täterinnen	vs.	*falsche Zeuginnen*	0.029	0.071	.917	-0.146	0.205
Zeuginnen	vs.	*falsche Zeuginnen*	0.176	0.071	.049	0.001	0.352

Tabelle F.18. Einfaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person als UV und der Variable Vorkenntnisse zur Glaubhaftigkeitsbeurteilung als AV

Q.d.V.	QS	df	MQ	F	p
Gruppe	0.373	2	0.186	1.057	.351
Fehler	17.441	99	0.176		
Gesamt	17.814	101			

Tabelle F.19. Einfaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person als UV und der Variable frühere Teilnahme an Untersuchungen zur Glaubhaftigkeitsbeurteilung als AV

Q.d.V.	QS	df	MQ	F	p
Gruppe	0.078	2	0.039	2.063	.133
Fehler	1.882	99	0.019		
Gesamt	1.961	101			

Tabelle F.20. Einfache prozentuale Übereinstimmung zwischen den Ratern A, B und C

	Paarweise Übereinstimmung			Gesamtübereinstimmung	
Kriterium	A - B	A - C	B - C	Mittelwert	A - B - C
1. *Konsistenz*	92.16	90.20	88.24	90.20	85.29
2. *Unordnung*	66.67	58.82	64.71	63.40	48.04
3. *Details*	31.37	29.41	60.78	40.52	23.53
4. *Verknüpfungen*	53.92	50.98	40.20	48.37	26.47
5. *Interaktionen*	66.67	68.63	79.41	71.57	60.78
6. *Gespräche*	59.80	71.57	56.86	62.75	45.10
7. *Komplikationen*	84.31	88.24	92.16	88.24	82.35
8. *Ausgefallenes*	38.24	58.82	71.57	56.21	35.29
9. *Nebensächliches*	51.96	33.33	33.33	39.54	19.61
10. *Unverstandenes*	96.08	97.06	97.06	96.73	95.10
11. *Indirektes*	100.00	97.06	97.06	98.04	97.06
12. *Eigenseelisches*	50.00	55.88	33.33	46.41	25.49
13. *Fremdseelisches*	76.47	52.94	52.94	60.78	45.10
14. *Verbesserungen*	45.10	40.20	25.49	36.93	15.69
15. *Erinnerungslücken*	74.51	82.35	74.51	77.12	67.65
16. *Selbsteinwände*	98.04	93.14	93.14	94.77	92.16
17. *Eigenbelastung*	99.02	98.04	99.02	98.69	98.04
18. *Fremdentlastung*	70.59	77.45	66.67	71.57	60.78
Durchschnitt:	69.72	69.12	68.14	68.99	56.86

Tabelle F.21. Erweiterte prozentuale Übereinstimmung zwischen den Ratern A, B und C

	Paarweise Übereinstimmung			Gesamtübereinstimmung	
Kriterium	A - B	A - C	B - C	Mittelwert	A - B - C
1. *Konsistenz*	98.04	99.02	97.06	98.04	97.06
2. *Unordnung*	98.04	95.10	92.16	95.10	90.20
3. *Details*	94.12	71.57	93.14	86.27	68.63
4. *Verknüpfungen*	94.12	96.08	93.14	94.44	88.24
5. *Interaktionen*	94.12	91.18	94.12	93.14	87.25
6. *Gespräche*	99.02	100.00	99.02	99.35	98.04
7. *Komplikationen*	97.06	97.06	100.00	98.04	97.06
8. *Ausgefallenes*	100.00	98.04	100.00	99.35	98.04
9. *Nebensächliches*	78.43	69.61	74.51	74.18	55.88
10. *Unverstandenes*	100.00	99.02	99.02	99.35	99.02
11. *Indirektes*	100.00	100.00	100.00	100.00	100.00
12. *Eigenseelisches*	95.10	100.00	89.22	94.77	86.27
13. *Fremdseelisches*	99.02	97.06	92.16	96.08	90.20
14. *Verbesserungen*	87.25	89.22	71.57	82.68	65.69
15. *Erinnerungslücken*	99.02	97.06	100.00	98.69	96.08
16. *Selbsteinwände*	100.00	100.00	100.00	100.00	100.00
17. *Eigenbelastung*	99.02	98.04	99.02	98.69	98.04
18. *Fremdentlastung*	99.02	96.08	92.16	95.75	90.20
Durchschnitt:	96.19	94.12	93.68	94.66	89.22

Tabelle F.22. Produkt-Moment-Korrelationen zwischen den Ratern A, B und C

	Raterpaare												Mittlere Korrrelation	
	A – B				A - C				B - C					
	r		p	Fishers Z	r		p	Fishers Z	r		p	Fishers Z	Fishers Z	r
1. *Konsistenz*	-.037		.355	-.037	.125		.106	.126	.029		.386	.029	.039	.039
2. *Unordnung*	.395	**	.000	.418	.407	**	.000	.432	.328	**	.000	.341	.397	.377
3. *Details*	.703	**	.000	.873	.460	**	.000	.497	.631	**	.000	.743	.705	.607
4. *Verknüpfungen*	.433	**	.000	.464	.654	**	.000	.782	.389	**	.000	.411	.552	.502
5. *Interaktionen*	.152		.063	.153	.366	**	.000	.384	.487	**	.000	.532	.356	.342
6. *Gespräche*	.818	**	.000	1.151	.866	**	.000	1.317	.813	**	.000	1.136	1.201	.834
7. *Komplikationen*	-.077		.220	-.077	.060		.274	.060	.164	*	.050	.165	.049	.049
8. *Ausgefallenes*	-.026		.399	-.026	.148		.068	.149	.541	**	.000	.606	.243	.238
9. *Nebensächliches*	.209	*	.018	.212	.156		.059	.157	.231	**	.010	.235	.202	.199
10. *Unverstandenes*	-.020		.421	-.020	-.014		.444	-.014	-.014		.444	-.014	-.016	-.016
11. *Indirektes*	[a]				[a]				[a]					
12. *Eigenseelisches*	.600	**	.000	.693	.608	**	.000	.706	.510	**	.000	.563	.654	.574
13. *Fremdseelisches*	.591	**	.000	.679	.646	**	.000	.768	.731	**	.000	.931	.793	.660
14. *Verbesserungen*	.215	*	.015	.218	.436	**	.000	.467	.168	*	.046	.170	.285	.278
15. *Erinnerungslücken*	.127		.102	.128	.228	*	.011	.232	.434	**	.000	.465	.275	.268
16. *Selbsteinwände*	-.010		.461	-.010	-.025		.402	-.025	-.025		.402	-.025	-.020	-.020
17. *Eigenbelastung*	[a]				-.010		.461	-.010	[a]					
18. *Fremdentlastung*	-.040		.346	-.040	.623	**	.000	.730	.216	*	.015	.219	.303	.294

Anmerkung: * $p < .05$ (einseitig); ** $p < .01$ (einseitig); [a] kann nicht berechnet werden, da mindestens eine der Variablen konstant ist.

Tabelle F.23. Gewichtete Kappa-Koeffizienten (κ_w) zwischen den Ratern A, B und C

	Raterpaare																							
	A – B							A - C							B - C									
				95%-Konf.-int.							95%-Konf.-int.								95%-Konf.-int.					
Kriterium	κ_w	SE	SE (0)	Untergrenze	Obergrenze	z (κ_w)		κ_w	SE	SE (0)	Untergrenze	Obergrenze	z (κ_w)		κ_w	SE	SE (0)	Untergrenze	Obergrenze	z (κ_w)		Mittl. κ_w
1	-0.03	0.44	0.44	-0.90	0.84	-0.07		0.10	0.34	0.34	-0.56	0.76	0.29		0.03	0.34	0.33	-0.63	0.69	0.09		0.03
2	0.39	0.11	0.18	0.19	0.60	2.24	*	0.38	0.12	0.18	0.14	0.61	2.14	*	0.30	0.15	0.19	-0.01	0.60	1.60		0.36
3	0.59	0.04	0.12	0.50	0.67	4.99	**	0.31	0.08	0.11	0.15	0.47	2.78	**	0.58	0.07	0.11	0.44	0.72	5.30	**	0.49
4	0.39	0.09	0.13	0.21	0.57	3.02	**	0.63	0.05	0.12	0.53	0.73	5.20	**	0.34	0.09	0.11	0.16	0.53	2.97	**	0.45
5	0.12	0.19	0.21	-0.26	0.50	0.56		0.37	0.12	0.18	0.12	0.61	2.05	*	0.38	0.18	0.20	0.03	0.72	1.89	*	0.29
6	0.77	0.03	0.12	0.71	0.83	6.57	**	0.86	0.02	0.12	0.81	0.90	7.09	**	0.74	0.03	0.12	0.68	0.81	6.22	**	0.79
7	-0.07	0.32	0.31	-0.68	0.55	-0.21		0.05	0.33	0.33	-0.60	0.69	0.14		0.16	0.28	0.31	-0.40	0.72	0.52		0.05
8	-0.01	0.08	0.08	-0.17	0.14	-0.15		0.09	0.13	0.14	-0.17	0.35	0.65		0.50	0.08	0.11	0.35	0.66	4.76	**	0.20
9	0.14	0.14	0.16	-0.15	0.42	0.86		0.08	0.11	0.11	-0.14	0.30	0.71		0.22	0.10	0.11	0.02	0.42	1.89	*	0.15
10	-0.02	0.50	0.50	-1.00	0.96	-0.04		-0.01	0.71	0.70	-1.40	1.38	-0.02		-0.01	0.71	0.70	-1.40	1.38	-0.02		-0.02
11	—[a]	—	—	—	—	—	—	0.00	0.57	0.57	-1.12	1.12	0.00		0.00	0.57	0.57	-1.12	1.12	0.00		—
12	0.46	0.07	0.15	0.31	0.61	3.06	**	0.59	0.05	0.13	0.50	0.68	4.47	**	0.33	0.08	0.12	0.18	0.48	2.85	**	0.46
13	0.52	0.10	0.13	0.33	0.72	4.11	**	0.53	0.06	0.12	0.41	0.66	4.55	**	0.45	0.08	0.12	0.28	0.61	3.70	**	0.50
14	0.17	0.12	0.15	-0.07	0.41	1.17		0.39	0.08	0.12	0.24	0.54	3.21	**	0.10	0.10	0.11	-0.10	0.30	0.89		0.22
15	0.09	0.18	0.18	-0.26	0.44	0.49		0.16	0.23	0.23	-0.29	0.62	0.71		0.43	0.10	0.17	0.24	0.62	2.52	**	0.23
16	-0.01	0.71	0.70	-1.40	1.38	-0.01		-0.02	0.37	0.37	-0.74	0.71	-0.05		-0.02	0.37	0.37	-0.74	0.71	-0.05		-0.02
17	0.00	1.00	1.00	-1.95	1.95	0.00		-0.01	0.71	0.70	-1.40	1.38	-0.01		0.00	1.00	1.00	-1.95	1.95	0.00		0.00
18	-0.04	0.18	0.21	-0.40	0.32	-0.19		0.52	0.11	0.19	0.30	0.75	2.79	**	0.15	0.16	0.20	-0.16	0.47	0.78		0.21

Anmerkung: SE = Standardfehler von κ_w; SE (0) = Standardfehler von κ_w, falls κ_w in Population 0 beträgt; z(κ_w) = z-Wert von κ_w; * p < .05; ** p < .01; [a] κ_w konnte nicht berechnet werden, da beide Rater durchgängig Skalenstufe 0 wählten.

Tabelle F.24. Übereinstimmungsmatrizen der Rater A, B und C

Kriterium 1: *Konsistenz*

Raterpaar A - B

	0	1	2	3	Σ
0	0	0	0	0	0
1	0	0	0	0	0
2	0	0	0	3	3
3	0	2	3	94	99
Σ	0	2	3	97	102

Raterpaar A - C

	0	1	2	3	Σ
0	0	0	0	0	0
1	0	0	0	0	0
2	0	0	1	2	3
3	0	1	7	91	99
Σ	0	1	8	93	102

Raterpaar B - C

	0	1	2	3	Σ
0	0	0	0	0	0
1	0	0	0	2	2
2	0	0	1	2	3
3	0	1	7	89	97
Σ	0	1	8	93	102

Kriterium 2: *Unordnung*

Raterpaar A - B

	0	1	2	3	Σ
0	58	12	0	0	70
1	14	10	2	1	27
2	1	4	0	0	5
3	0	0	0	0	0
Σ	73	26	2	1	102

Raterpaar A - C

	0	1	2	3	Σ
0	51	17	1	1	70
1	12	8	5	2	27
2	1	2	1	1	4
3	0	0	0	0	0
Σ	64	27	7	4	102

Raterpaar B - C

	0	1	2	3	Σ
0	54	15	3	1	73
1	8	12	4	2	26
2	1	0	0	1	2
3	1	0	0	0	1
Σ	64	27	7	4	102

Kriterium 3: *Details*

Raterpaar A - B

	0	1	2	3	Σ
0	0	27	2	0	29
1	0	9	17	2	28
2	0	4	10	13	27
3	0	2	3	13	18
Σ	0	42	32	28	102

Raterpaar A - C

	0	1	2	3	Σ
0	0	13	12	4	29
1	1	4	11	12	28
2	0	3	11	13	27
3	0	1	2	15	18
Σ	1	21	36	44	102

Raterpaar B - C

	0	1	2	3	Σ
0	0	0	0	0	0
1	0	20	16	6	42
2	1	1	17	13	32
3	0	0	3	25	28
Σ	1	21	36	44	102

Kriterium 4: *Verknüpfungen*

Raterpaar A - B

	0	1	2	3	Σ
0	6	19	4	0	29
1	1	36	9	1	47
2	0	9	13	0	22
3	0	1	3	0	4
Σ	7	65	29	1	102

Raterpaar A - C

	0	1	2	3	Σ
0	21	6	2	0	29
1	10	17	20	0	47
2	2	2	12	6	22
3	0	0	2	2	4
Σ	33	25	36	8	102

Raterpaar B - C

	0	1	2	3	Σ
0	6	1	0	0	7
1	21	21	22	1	65
2	5	3	14	7	29
3	1	0	0	0	1
Σ	33	25	36	8	102

Kriterium 5: *Interaktionen*

Raterpaar A - B

	0	1	2	3	Σ
0	61	5	1	0	67
1	23	7	0	0	30
2	3	0	0	0	3
3	1	1	0	0	2
Σ	88	13	1	0	102

Raterpaar A - C

	0	1	2	3	Σ
0	57	4	6	0	67
1	16	12	1	1	30
2	2	0	1	0	3
3	0	0	2	0	2
Σ	75	16	10	1	102

Raterpaar B - C

	0	1	2	3	Σ
0	73	9	5	1	88
1	2	7	4	0	13
2	0	0	1	0	1
3	0	0	0	0	0
Σ	75	16	10	1	102

Kriterium 6: *Gespräche*

Raterpaar A - B

	0	1	2	3	Σ
0	11	4	0	0	15
1	0	24	0	0	24
2	0	14	5	3	32
3	0	1	19	11	31
Σ	11	43	34	14	102

Raterpaar A - C

	0	1	2	3	Σ
0	8	7	0	0	15
1	0	18	6	0	24
2	0	4	21	7	32
3	0	0	5	26	31
Σ	8	29	32	33	102

Raterpaar B - C

	0	1	2	3	Σ
0	7	4	0	0	11
1	1	25	16	1	43
2	0	0	14	20	34
3	0	0	2	12	14
Σ	8	29	32	33	102

Tabelle F.24 (Fortsetzung). Übereinstimmungsmatrizen der Rater A, B und C

Kriterium 7: *Komplikationen*

Raterpaar A - B

	0	1	2	3	Σ
0	86	6	0	0	92
1	7	0	0	0	7
2	3	0	0	0	3
3	0	0	0	0	0
Σ	96	6	0	0	102

Raterpaar A - C

	0	1	2	3	Σ
0	89	3	0	0	92
1	6	1	0	0	7
2	3	0	0	0	3
3	0	0	0	0	0
Σ	98	4	0	0	102

Raterpaar B - C

	0	1	2	3	Σ
0	93	3	0	0	96
1	5	1	0	0	6
2	0	0	0	0	0
3	0	0	0	0	0
Σ	98	4	0	0	102

Kriterium 8: *Ausgefallenes*

Raterpaar A - B

	0	1	2	3	Σ
0	33	59	0	0	92
1	4	6	0	0	10
2	0	0	0	0	0
3	0	0	0	0	0
Σ	37	65	0	0	102

Raterpaar A - C

	0	1	2	3	Σ
0	55	35	2	0	92
1	4	5	1	0	10
2	0	0	0	0	0
3	0	0	0	0	0
Σ	59	40	3	0	102

Raterpaar B - C

	0	1	2	3	Σ
0	35	2	0	0	37
1	24	38	3	0	65
2	0	0	0	0	0
3	0	0	0	0	0
Σ	59	40	3	0	102

Kriterium 9: *Nebensächliches*

Raterpaar A - B

	0	1	2	3	Σ
0	50	16	8	9	83
1	8	1	2	4	15
2	1	1	1	0	3
3	0	0	0	1	1
Σ	59	18	11	14	102

Raterpaar A - C

	0	1	2	3	Σ
0	29	26	25	3	83
1	1	4	9	1	15
2	1	0	1	1	3
3	1	0	0	0	1
Σ	32	30	35	5	102

Raterpaar B - C

	0	1	2	3	Σ
0	20	21	17	1	59
1	6	6	6	0	18
2	2	1	6	2	11
3	4	2	6	2	14
Σ	32	30	35	5	102

Kriterium 10: *Unverstandenes*

Raterpaar A - B

	0	1	2	3	Σ
0	98	2	0	0	100
1	2	0	0	0	2
2	0	0	0	0	0
3	0	0	0	0	0
Σ	100	2	0	0	102

Raterpaar A - C

	0	1	2	3	Σ
0	99	0	1	0	100
1	2	0	0	0	2
2	0	0	0	0	0
3	0	0	0	0	0
Σ	101	0	1	0	102

Raterpaar B - C

	0	1	2	3	Σ
0	99	0	1	0	100
1	2	0	0	0	2
2	0	0	0	0	0
3	0	0	0	0	0
Σ	101	0	1	0	102

Kriterium 11: *Indirektes*

Raterpaar A - B

	0	1	2	3	Σ
0	102	0	0	0	102
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
Σ	102	0	0	0	102

Raterpaar A - C

	0	1	2	3	Σ
0	99	3	0	0	102
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
Σ	99	3	0	0	102

Raterpaar B - C

	0	1	2	3	Σ
0	99	3	0	0	102
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
Σ	99	3	0	0	102

Kriterium 12: *Eigenseelisches*

Raterpaar A - B

	0	1	2	3	Σ
0	6	0	0	0	6
1	27	38	1	0	66
2	0	17	5	0	22
3	0	5	1	2	8
Σ	33	60	7	2	102

Raterpaar A - C

	0	1	2	3	Σ
0	6	0	0	0	6
1	4	31	31	0	66
2	0	5	16	1	22
3	0	0	4	4	8
Σ	10	36	51	5	102

Raterpaar B - C

	0	1	2	3	Σ
0	9	15	9	0	33
1	1	20	37	2	60
2	0	1	4	2	7
3	0	0	1	1	2
Σ	10	36	51	5	102

Tabelle F.24 (Fortsetzung). Übereinstimmungsmatrizen der Rater A, B und C

Kriterium 13: *Fremdseelisches*

Raterpaar A - B

	0	1	2	3	Σ
0	54	1	0	0	55
1	20	24	0	0	44
2	1	2	0	0	3
3	0	0	0	0	0
Σ	75	27	0	0	102

Raterpaar A - C

	0	1	2	3	Σ
0	44	8	3	0	55
1	7	9	28	0	44
2	0	2	1	0	3
3	0	0	0	0	0
Σ	51	19	32	0	102

Raterpaar B - C

	0	1	2	3	Σ
0	51	16	8	0	35
1	0	3	24	0	27
2	0	0	0	0	0
3	0	0	0	0	0
Σ	51	19	32	0	102

Kriterium 14: *Verbesserungen*

Raterpaar A - B

	0	1	2	3	Σ
0	6	3	1	0	10
1	20	36	6	0	62
2	5	14	3	0	22
3	1	6	0	1	8
Σ	32	59	10	1	102

Raterpaar A - C

	0	1	2	3	Σ
0	5	2	3	0	10
1	9	18	27	8	62
2	0	4	14	4	22
3	0	0	4	4	8
Σ	14	24	48	16	102

Raterpaar B - C

	0	1	2	3	Σ
0	6	10	13	3	32
1	8	12	26	13	59
2	0	2	8	0	10
3	0	0	1	0	1
Σ	14	24	48	16	102

Kriterium 15: *Erinnerungslücken*

Raterpaar A - B

	0	1	2	3	Σ
0	73	22	1	0	96
1	3	3	0	0	6
2	0	0	0	0	0
3	0	0	0	0	0
Σ	76	25	1	0	102

Raterpaar A - C

	0	1	2	3	Σ
0	80	13	3	0	96
1	2	4	0	0	6
2	0	0	0	0	0
3	0	0	0	0	0
Σ	82	17	3	0	102

Raterpaar B - C

	0	1	2	3	Σ
0	68	8	0	0	76
1	14	8	3	0	25
2	0	1	0	0	1
3	0	0	0	0	0
Σ	82	17	3	0	102

Kriterium 16: *Selbsteinwände*

Raterpaar A - B

	0	1	2	3	Σ
0	100	1	0	0	101
1	1	0	0	0	1
2	0	0	0	0	0
3	0	0	0	0	0
Σ	101	1	0	0	102

Raterpaar A - C

	0	1	2	3	Σ
0	95	6	0	0	101
1	1	0	0	0	1
2	0	0	0	0	0
3	0	0	0	0	0
Σ	96	6	0	0	102

Raterpaar B - C

	0	1	2	3	Σ
0	95	6	0	0	101
1	1	0	0	0	1
2	0	0	0	0	0
3	0	0	0	0	0
Σ	96	6	0	0	102

Kriterium 17: *Eigenbelastung*

Raterpaar A - B

	0	1	2	3	Σ
0	101	0	0	0	101
1	0	0	0	0	0
2	1	0	0	0	1
3	0	0	0	0	0
Σ	102	0	0	0	102

Raterpaar A - C

	0	1	2	3	Σ
0	100	0	1	0	101
1	0	0	0	0	0
2	1	0	0	0	1
3	0	0	0	0	0
Σ	101	0	1	0	102

Raterpaar B - C

	0	1	2	3	Σ
0	101	0	1	0	102
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
Σ	101	0	1	0	102

Kriterium 18: *Fremdentlastung*

Raterpaar A - B

	0	1	2	3	Σ
0	72	13	0	0	85
1	15	0	0	0	15
2	1	1	0	0	2
3	0	0	0	0	0
Σ	88	14	0	0	102

Raterpaar A - C

	0	1	2	3	Σ
0	69	12	4	0	85
1	2	9	4	0	15
2	0	0	1	1	2
3	0	0	0	0	0
Σ	71	21	9	1	102

Raterpaar B - C

	0	1	2	3	Σ
0	64	17	7	0	88
1	7	4	2	1	14
2	0	0	0	0	0
3	0	0	0	0	0
Σ	71	21	9	1	102

Tabelle F.25. Ergebnisse der varianzanalytischen Bestimmung der Auswertungsobjektivität

	ANOVA								
Kriterium	Quelle der Variation	df	MQ	$s^{2\ a}$	F	p	r_{AA} (single meas.)	$r_{AA}{}^{b}$ (average meas.)	Auswerter- / Gesamtvarianz
1. *Konsistenz*	Aussagen	101	0.086	0.003	1.103	n.s.	.033	.093	.005
	Auswerter	2	0.121	0.000	1.550	n.s.			
	Rest (Fehler)	202	0.078	0.078					
2. *Unordnung*	Aussagen	101	0.735	0.154	2.692	< .01	.354	.622	.019
	Auswerter	2	1.098	0.008	4.023	< .05			
	Rest (Fehler)	202	0.273	0.273					
3. *Details*	Aussagen	101	1.777	0.477	5.135	< .01	.471	.728	.187
	Auswerter	2	19.709	0.190	56.944	< .01			
	Rest (Fehler)	202	0.346	0.346					
4. *Verknüpfungen*	Aussagen	101	1.290	0.317	3.808	< .01	.476	.731	.016
	Auswerter	2	1.435	0.011	4.233	< .05			
	Rest (Fehler)	202	0.339	0.339					
5. *Interaktionen*	Aussagen	101	0.580	0.112	2.388	< .01	.300	.563	.050
	Auswerter	2	2.147	0.019	8.844	< .01			
	Rest (Fehler)	202	0.243	0.243					
6. *Gespräche*	Aussagen	101	2.434	0.758	15.330	< .01	.795	.921	.039
	Auswerter	2	3.964	0.037	24.967	< .01			
	Rest (Fehler)	202	0.159	0.159					
7. *Komplikationen*	Aussagen	101	0.092	0.002	1.056	n.s.	.018	.052	.014
	Auswerter	2	0.219	0.001	2.518	n.s.			
	Rest (Fehler)	202	0.087	0.087					
8. *Ausgefallenes*	Aussagen	101	0.322	0.055	2.071	< .01	.195	.421	.259
	Auswerter	2	7.650	0.073	49.269	< .01			
	Rest (Fehler)	202	0.155	0.155					
9. *Nebensächliches*	Aussagen	101	1.076	0.146	1.685	< .01	.148	.343	.201
	Auswerter	2	20.807	0.198	32.569	< .01			
	Rest (Fehler)	202	0.639	0.639					

Tabelle F.25 (Fortsetzung). Ergebnisse der varianzanalytischen Bestimmung der Auswertungsobjektivität

Kriterium	ANOVA Quelle der Variation	df	MQ	s^2 [a]	F	p	r_{AA} (single meas.)	r_{AA} [b] (average meas.)	Auswerter- / Gesamtvarianz
10. *Unverstandenes*	Aussagen	101	0.025	0.000	0.956	n.s.	-.015	-.047	-.010
	Auswerter	2	0.000	0.000	0.000	n.s.			
	Rest (Fehler)	202	0.026	0.026					
11. *Indirektes*	Aussagen	101	0.010	0.000	1.000	n.s.	.000	.000	.020
	Auswerter	2	0.029	0.000	3.061	< .05			
	Rest (Fehler)	202	0.010	0.010					
12. *Eigenseelisches*	Aussagen	101	1.051	0.280	4.991	< .01	.447	.708	.217
	Auswerter	2	14.062	0.136	66.770	< .01			
	Rest (Fehler)	202	0.211	0.211					
13. *Fremdseelisches*	Aussagen	101	0.934	0.251	5.174	< .01	.496	.747	.147
	Auswerter	2	7.768	0.074	43.032	< .01			
	Rest (Fehler)	202	0.181	0.181					
14. *Verbesserungen*	Aussagen	101	0.932	0.166	2.149	< .01	.215	.450	.225
	Auswerter	2	18.209	0.174	41.998	< .01			
	Rest (Fehler)	202	0.434	0.434					
15. *Erinnerungslücken*	Aussagen	101	0.262	0.046	2.122	< .01	.256	.508	.060
	Auswerter	2	1.219	0.011	9.891	< .01			
	Rest (Fehler)	202	0.123	0.123					
16. *Selbsteinwände*	Aussagen	101	0.024	0.000	0.951	n.s.	-.016	-.051	.021
	Auswerter	2	0.082	0.001	3.192	< .05			
	Rest (Fehler)	202	0.026	0.026					
17. *Eigenbelastung*	Aussagen	101	0.026	0.000	0.985	n.s.	-.005	-.015	-.005
	Auswerter	2	0.013	0.000	0.498	n.s.			
	Rest (Fehler)	202	0.026	0.026					
18. *Fremdentlastung*	Aussagen	101	0.422	0.079	2.265	< .01	.276	.534	.069
	Auswerter	2	2.186	0.020	11.737	< .01			
	Rest (Fehler)	202	0.186	0.186					

Anmerkung: [a] Die Varianzkomponenten werden folgendermaßen geschätzt (vgl. Iseler, 1967, S. 137): $s^2_{Aussagen} = (MQ_{Aussagen} - MQ_{Rest}) / n$; $s^2_{Auswerter} = (MQ_{Auswerter} - MQ_{Rest}) / k$; $s^2_{Rest} = MQ_{Rest}$. Dabei sind n = Anzahl der Auswerter und k = Anzahl der Aussagen.

[b] $r_{AA(average\ measure)} = s^2_{Aussagen} / (s^2_{Aussagen} + ((s^2_{Auswerter} + s^2_{Rest}) / n))$ (vgl. Crocker & Algina, 1986, S. 167).

Tabelle F.26. Finn-Statistiken

Kriterium	Varianz innerhalb Aussagen	Finn-Statistik QS	df	MQ	F	p	Finnkoeff. [a]
1. *Konsistenz*	beobachtet	16.00	204	0.078	14.167	< .01	.929
	erwartet	340.00	306	1.111			(.937)
2. *Unordnung*	beobachtet	57.33	204	0.281	3.953	< .01	.747
	erwartet	340.00	306	1.111			(.775)
3. *Details*	beobachtet	109.33	204	0.536	2.073	< .01	.518
	erwartet	340.00	306	1.111			(.571)
4. *Verknüpfungen*	beobachtet	71.33	204	0.350	3.178	< .01	.685
	erwartet	340.00	306	1.111			(.720)
5. *Interaktionen*	beobachtet	53.33	204	0.261	4.250	< .01	.765
	erwartet	340.00	306	1.111			(.791)
6. *Gespräche*	beobachtet	40.00	204	0.196	5.667	< .01	.824
	erwartet	340.00	306	1.111			(.843)
7. *Komplikationen*	beobachtet	18.00	204	0.088	12.593	< .01	.921
	erwartet	340.00	306	1.111			(.929)
8. *Ausgefallenes*	beobachtet	46.67	204	0.229	4.857	< .01	.794
	erwartet	340.00	306	1.111			(.817)
9. *Nebensächliches*	beobachtet	170.67	204	0.837	1.328	< .05	.247
	erwartet	340.00	306	1.111			(.331)
10. *Unverstandenes*	beobachtet	5.33	204	0.026	42.500	< .01	.976
	erwartet	340.00	306	1.111			(.979)
11. *Indirektes*	beobachtet	2.00	204	0.010	113.333	< .01	.991
	erwartet	340.00	306	1.111			(.992)
12. *Eigenseelisches*	beobachtet	70.67	204	0.346	3.208	< .01	.688
	erwartet	340.00	306	1.111			(.723)
13. *Fremdseelisches*	beobachtet	52.00	204	0.255	4.359	< .01	.771
	erwartet	340.00	306	1.111			(.796)
14. *Verbesserungen*	beobachtet	124.00	204	0.608	1.828	< .01	.453
	erwartet	340.00	306	1.111			(.514)
15. *Erinnerungslücken*	beobachtet	27.33	204	0.134	8.293	< .01	.879
	erwartet	340.00	306	1.111			(.893)
16. *Selbsteinwände*	beobachtet	5.33	204	0.026	42.500	< .01	.976
	erwartet	340.00	306	1.111			(.979)
17. *Eigenbelastung*	beobachtet	5.33	204	0.026	42.500	< .01	.976
	erwartet	340.00	306	1.111			(.979)
18. *Fremdentlastung*	beobachtet	42.00	204	0.206	5.397	< .01	.815
	erwartet	340.00	306	1.111			(.835)

Anmerkung: [a] Berechnung nach der Formel von Finn (1970, S. 72); Werte in Klammern wurden nach der bei Asendorpf und Wallbott (1979, S. 245) aufgeführten Formel berechnet.

Tabelle F.27. Produkt-Moment-Korrelationen zwischen den Glaubhaftigkeitskriterien

		2. *Unordnung*	3. *Details*	4. *Verknüpfungen*	5. *Interaktionen*	6. *Gespräche*	7. *Komplikationen*	8. *Ausgefallenes*	9. *Nebensächliches*	10. *Unverstandenes*	11. *Indirektes*	12. *Eigenseelisches*	13. *Fremdseelisches*	14. *Verbesserungen*	15. *Erinnerungslücken*	16. *Selbsteinwände*	17. *Eigenbelastung*	18. *Fremdentlastung*
1.	*Konsistenz*	.015	.068	.135	.189	-.028	.130	-.065	.014	.012	-.047	.197*	.074	-.064	.044	.041	.055	.047
2.	*Unordnung*		.371*	.047	-.065	.276*	-.162	.116	.310*	.164	.211*	.287*	-.012	.358*	.222*	-.016	.267*	.181
3.	*Details*			.258*	.105	.602*	-.059	.303 *	.525*	.243*	.197*	.397*	.260*	.634*	.225*	-.067	.068	.018
4.	*Verknüpfungen*				.101	.058	.106	-.242*	.016	.044	-.038	.280*	.135	.223*	-.097	-.120	.186	.115
5.	*Interaktionen*					.130	.363*	-.259*	-.025	.037	.008	.142	.195*	.015	.113	.041	.060	-.084
6.	*Gespräche*						-.130	.175	.247*	.201*	.098	.180	.087	.376*	.195*	-.003	.150	.125
7.	*Komplikationen*							-.005	-.208*	-.024	.036	.077	.112	-.121	.029	-.056	-.061	-.015
8.	*Ausgefallenes*								.280*	.069	.086	-.002	-.077	.158	.086	-.056	-.099	.009
9.	*Nebensächliches*									.060	.081	.155	.182	.417*	.166	.095	.105	-.007
10.	*Unverstandenes*										.598*	.029	.013	.165	.029	-.063	.228*	.051
11.	*Indirektes*											.105	.045	.098	.089	.165	-.025	-.114
12.	*Eigenseelisches*												.241*	.289*	.091	-.057	.072	.209*
13.	*Fremdseelisches*													.226*	.081	-.034	.037	-.061
14.	*Verbesserungen*														.096	-.127	.066	.093
15.	*Erinnerungslücken*															.067	-.088	-.141
16.	*Selbsteinwände*																.222*	.036
17.	*Eigenbelastung*																	.286*

Anmerkung: * $p < .05$ (zweiseitig).

Tabelle F.28. Rangkorrelationen (nach Spearman) zwischen den Glaubhaftigkeitskriterien

		2. *Unordnung*	3. *Details*	4. *Verknüpfungen*	5. *Interaktionen*	6. *Gespräche*	7. *Komplikationen*	8. *Ausgefallenes*	9. *Nebensächliches*	10. *Unverstandenes*	11. *Indirektes*	12. *Eigenseelisches*	13. *Fremdseelisches*	14. *Verbesserungen*	15. *Erinnerungslücken*	16. *Selbsteinwände*	17. *Eigenbelastung*	18. *Fremdentlastung*
1.	*Konsistenz*	.008	.123	.118	.209*	.038	.126	-.030	.019	-.026	-.083	.219*	.062	-.030	.110	.023	.059	.041
2.	*Unordnung*		.381*	.072	-.019	.267*	-.169	.145	.260*	.093	.223*	.308*	.039	.356*	.277*	-.015	.193	.045
3.	*Details*			.269*	.158	.608*	-.051	.303*	.506*	.275*	.196*	.425*	.282*	.633*	.255*	-.071	.076	-.020
4.	*Verknüpfungen*				.073	.072	.122	-.264*	.020	.049	-.015	.336*	.159	.221*	-.165	-.115	.181	.078
5.	*Interaktionen*					.165	.373*	-.197*	-.038	.086	.046	.161	.143	.015	.151	.008	.050	-.092
6.	*Gespräche*						-.141	.206*	.218*	.221*	.099	.165	.105	.382*	.220*	-.021	.153	.059
7.	*Komplikationen*							-.064	-.208*	.007	.064	.146	.157	-.100	.115	-.044	-.065	-.007
8.	*Ausgefallenes*								.298*	.094	.075	-.052	-.091	.172	.131	-.049	-.102	.035
9.	*Nebensächliches*									.100	.086	.171	.167	.390*	.205*	.115	.127	-.048
10.	*Unverstandenes*										.506*	.044	.038	.204*	.081	-.066	.292*	-.053
11.	*Indirektes*											.131	.046	.121	.185	.165	-.025	-.135
12.	*Eigenseelisches*												.250*	.302*	.093	-.047	.105	.184
13.	*Fremdseelisches*													.238*	.058	-.034	.024	-.067
14.	*Verbesserungen*														.090	-.146	.081	.081
15.	*Erinnerungslücken*															.080	-.102	-.096
16.	*Selbsteinwände*																.222*	.006
17.	*Eigenbelastung*																	.084

Anmerkung: * $p < .05$ (zweiseitig).

Tabelle F.29. Einfaktorielle MANOVA mit dem Gruppenfaktor Status der aussagenden Person (*Täterinnen*, *Zeuginnen*, *falsche Zeuginnen*) als UV und den 18 Glaubhaftigkeitskriterien als AVn

Teststatistik	Wert	F	Hypothese df	Fehler df	p
Pillais Spurkriterium PS	0.872	3.565	36	166	.000
Wilks Likelihood-Quotient Λ	0.290	3.908	36	164	.000
Hotellings Spurkriterium T	1.893	4.260	36	162	.000
Roys größter Eigenwert	1.528	7.045	18	83	.000

Tabelle F.30. Einfaktorielle MANOVA mit dem Gruppenfaktor Status der aussagenden Person (*Täterinnen* vs. *Zeuginnen*) als UV und den 18 Glaubhaftigkeitskriterien als AVn

Teststatistik	Wert	F	Hypothese df	Fehler df	p
Pillais Spurkriterium PS	0.545	3.263	18	49	.001
Wilks Likelihood-Quotient Λ	0.455	3.263	18	49	.001
Hotellings Spurkriterium T	1.199	3.263	18	49	.001
Roys größter Eigenwert	1.199	3.263	18	49	.001

Tabelle F.31. Einfaktorielle MANOVA mit dem Gruppenfaktor Status der aussagenden Person (*Täterinnen* vs. *falsche Zeuginnen*) als UV und den 18 Glaubhaftigkeitskriterien als AVn

Teststatistik	Wert	F	Hypothese df	Fehler df	p
Pillais Spurkriterium PS	0.440	2.137	18	49	.018
Wilks Likelihood-Quotient Λ	0.560	2.137	18	49	.018
Hotellings Spurkriterium T	0.785	2.137	18	49	.018
Roys größter Eigenwert	0.785	2.137	18	49	.018

Tabelle F.32. Einfaktorielle MANOVA mit dem Gruppenfaktor Status der aussagenden Person (*Zeuginnen* vs. *falsche Zeuginnen*) als UV und den 18 Glaubhaftigkeitskriterien als AVn

Teststatistik	Wert	F	Hypothese df	Fehler df	p
Pillais Spurkriterium PS	0.761	8.690	18	49	.000
Wilks Likelihood-Quotient Λ	0.239	8.690	18	49	.000
Hotellings Spurkriterium T	3.192	8.690	18	49	.000
Roys größter Eigenwert	3.192	8.690	18	49	.000

Tabelle F.33. Einfaktorielle ANOVAs mit dem Gruppenfaktor Status der aussagenden Person als UV und dem Ausprägungsgrad des jeweiligen Glaubhaftigkeitskriteriums als AV

	Kriterium	Q.d.V.	QS	df	MQ	F	p
1.	*Konsistenz*	Gruppe	0.002	2	0.001	0.037	.96345
		Fehler	2.895	99	0.029		
		Gesamt	2.898	101			
2.	*Unordnung*	Gruppe	1.020	2	0.510	2.128	.12454
		Fehler	23.722	99	0.240		
		Gesamt	24.742	101			
3.	*Details*	Gruppe	13.505	2	6.753	14.429	.00000
		Fehler	46.330	99	0.468		
		Gesamt	59.836	101			
4.	*Verknüpfungen*	Gruppe	5.309	2	2.655	6.891	.00158
		Fehler	38.137	99	0.385		
		Gesamt	43.447	101			
5.	*Interaktionen*	Gruppe	0.725	2	0.363	1.911	.15333
		Fehler	18.791	99	0.190		
		Gesamt	19.516	101			
6.	*Gespräche*	Gruppe	15.845	2	7.923	11.866	.00002
		Fehler	66.098	99	0.668		
		Gesamt	81.943	101			
7.	*Komplikationen*	Gruppe	0.041	2	0.021	0.672	.51299
		Fehler	3.049	99	0.031		
		Gesamt	3.090	101			
8.	*Ausgefallenes*	Gruppe	2.904	2	1.452	18.140	.00000
		Fehler	7.925	99	0.080		
		Gesamt	10.829	101			
9.	*Nebensächliches*	Gruppe	7.590	2	3.795	13.114	.00001
		Fehler	28.650	99	0.289		
		Gesamt	36.241	101			

Tabelle F.33 (Fortsetzung). Einfaktorielle ANOVAs mit dem Gruppenfaktor Status der aussagenden Person als UV und dem Ausprägungsgrad des jeweiligen Glaubhaftigkeitskriteriums als AV

	Kriterium	Q.d.V.	QS	df	MQ	F	p
10.	*Unverstandenes*	Gruppe	0.026	2	0.013	1.571	.21289
		Fehler	0.824	99	0.008		
		Gesamt	0.850	101			
11.	*Indirektes*	Gruppe	0.007	2	0.003	1.021	.36413
		Fehler	0.317	99	0.003		
		Gesamt	0.324	101			
12.	*Eigenseelisches*	Gruppe	0.590	2	0.295	0.840	.43485
		Fehler	34.801	99	0.352		
		Gesamt	35.391	101			
13.	*Fremdseelisches*	Gruppe	0.532	2	0.266	0.851	.43002
		Fehler	30.915	99	0.312		
		Gesamt	31.447	101			
14.	*Verbesserungen*	Gruppe	2.643	2	1.321	4.554	.01282
		Fehler	28.725	99	0.290		
		Gesamt	31.368	101			
15.	*Erinnerungslücken*	Gruppe	0.590	2	0.295	3.557	.03222
		Fehler	8.216	99	0.083		
		Gesamt	8.806	101			
16.	*Selbsteinwände*	Gruppe	0.002	2	0.001	0.132	.87649
		Fehler	0.817	99	0.008		
		Gesamt	0.819	101			
17.	*Eigenbelastung*	Gruppe	0.009	2	0.004	0.500	.60805
		Fehler	0.863	99	0.009		
		Gesamt	0.871	101			
18.	*Fremdentlastung*	Gruppe	0.712	2	0.356	2.613	.07833
		Fehler	13.493	99	0.136		
		Gesamt	14.206	101			

Tabelle F.34. Scheffé-Anschlußtests an signifikante ANOVA-Effekte bezüglich der Ausprägung der Glaubhaftigkeitskriterien in den experimentellen Gruppen

Kriterium	Gruppenvergleich			Mittl. Diff.	Stand.-fehler	p	95%-Konf.-int. Untergr.	Obergr.
3. *Details*	Tät.	vs.	Zeug.	-0.804	0.166	.00003	-1.216	-0.392
	Tät.	vs.	fal. Zeug.	-0.069	0.166	.91808	-0.481	0.344
	Zeug.	vs.	fal. Zeug.	0.735	0.166	.00013	0.323	1.148
4. *Verknüpfungen*	Tät.	vs.	Zeug.	0.284	0.151	.17339	-0.090	0.658
	Tät.	vs.	fal. Zeug.	-0.275	0.151	.19487	-0.649	0.100
	Zeug.	vs.	fal. Zeug.	-0.559	0.151	.00158	-0.933	-0.185
6. *Gespräche*	Tät.	vs.	Zeug.	-0.794	0.198	.00059	-1.287	-0.302
	Tät.	vs.	fal. Zeug.	0.078	0.198	.92473	-0.414	0.571
	Zeug.	vs.	fal. Zeug.	0.873	0.198	.00014	0.380	1.365
8. *Ausgefallenes*	Tät.	vs.	Zeug.	-0.118	0.069	.23497	-0.288	0.053
	Tät.	vs.	fal. Zeug.	0.284	0.069	.00037	0.114	0.455
	Zeug.	vs.	fal. Zeug.	0.402	0.069	.00000	0.231	0.572
9. *Nebensächliches*	Tät.	vs.	Zeug.	-0.373	0.130	.01989	-0.697	-0.048
	Tät.	vs.	fal. Zeug.	0.294	0.130	.08393	-0.030	0.618
	Zeug.	vs.	fal. Zeug.	0.667	0.130	.00001	0.342	0.991

Tabelle F.35. Zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Rater als UVn und dem Ausprägungsgrad von Kriterium 9 (*Nebensächliches*) als AV

Q.d.V.		QS	df	MQ	F	p
Gruppe		22.771	2	11.386	13.114	.000
Fehler (Gruppe)		85.951	99	0.868		
Rater	unkorrigiert	41.614	2	20.807	39.786	.000
	Greenh.-Geisser *		1.743	23.873	39.786	.000
Rater × Gruppe	unkorrigiert	25.503	4	6.376	12.191	.000
	Greenh.-Geisser *		3.486	7.315	12.191	.000
Fehler (Rater)	unkorrigiert	103.549	198	0.523		
	Greenh.-Geisser *		172.573	0.600		

Anmerkung: * $\varepsilon = .872$.

Tabelle F.36. Anschlußtests (geringste signifikante Differenz) für paarweise Vergleiche der Rater bezüglich der eingeschätzten Ausprägungsgrade von Glaubhaftigkeitskriterium 9 (*Nebensächliches*)

Gruppenvergleich			Mittlere Differenz	Standard-fehler	p	95%-Konfidenzintervall Untergrenze	Obergrenze
Rater A	vs.	Rater B	-0.569	0.087	.000	-0.742	-0.395
Rater A	vs.	Rater C	-0.892	0.095	.000	-1.081	-0.704
Rater B	vs.	Rater C	-0.324	0.119	.008	-0.559	-0.088

Tabelle F.37. Einfaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person als UV und dem Gesamtscore über alle 18 Glaubhaftigkeitskriterien als AV

Q.d.V.	QS	df	MQ	F	p
Gruppe	121.734	2	60.867	5.515	.005
Fehler	1092.601	99	11.036		
Gesamt	1214.336	101			

Tabelle F.38. Scheffé-Tests für paarweise Vergleiche der experimentellen Gruppen bezüglich der Gesamtscores über alle 18 Glaubhaftigkeitskriterien

Gruppenvergleich			Mittlere Differenz	Standard-fehler	p	95%-Konfidenzintervall Untergrenze	Obergrenze
Täterinnen	vs.	*Zeuginnen*	-2.059	0.806	.042	-4.061	-0.056
Täterinnen	vs.	*falsche Zeuginnen*	0.451	0.806	.855	-1.551	2.453
Zeuginnen	vs.	*falsche Zeuginnen*	2.510	0.806	.010	0.507	4.512

Tabelle F.39. Einfaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person als UV und dem Gesamtscore über alle 18 Glaubhaftigkeitskriterien als AV, wobei anstelle des mittleren Ratings der drei Rater bei Kriterium 9 das Rating von Rater A bei Kriterium 9 in den Gesamtscore einging

Q.d.V.	QS	df	MQ	F	p
Gruppe	61.595	2	30.797	2.829	.064
Fehler	1077.866	99	10.888		
Gesamt	1139.461	101			

Tabelle F.40. Scheffé-Tests für paarweise Vergleiche der experimentellen Gruppen bezüglich der Gesamtscores über alle 18 Glaubhaftigkeitskriterien, wobei anstelle des mittleren Ratings der drei Rater bei Kriterium 9 das Rating von Rater A bei Kriterium 9 in den Gesamtscore einging

Gruppenvergleich			Mittlere Differenz	Standard-fehler	p	95%-Konfidenzintervall Untergrenze	Obergrenze
Täterinnen	vs.	*Zeuginnen*	-1.333	0.800	.254	-3.322	0.656
Täterinnen	vs.	*falsche Zeuginnen*	0.510	0.800	.817	-1.479	2.499
Zeuginnen	vs.	*falsche Zeuginnen*	1.843	0.800	.076	-0.146	3.832

Tabelle F.41. Einfaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person als UV und dem Gesamtscore über alle 18 Glaubhaftigkeitskriterien als AV, wobei anstelle des mittleren Ratings der drei Rater bei Kriterium 9 das Rating von Rater B bei Kriterium 9 in den Gesamtscore einging

Q.d.V.	QS	df	MQ	F	p
Gruppe	218.046	2	109.023	8.630	.000
Fehler	1250.709	99	12.633		
Gesamt	1468.755	101			

Tabelle F.42. Scheffé-Tests für paarweise Vergleiche der experimentellen Gruppen bezüglich der Gesamtscores über alle 18 Glaubhaftigkeitskriterien, wobei anstelle des mittleren Ratings der drei Rater bei Kriterium 9 das Rating von Rater B bei Kriterium 9 in den Gesamtscore einging

Gruppenvergleich			Mittlere Differenz	Standard-fehler	p	95%-Konfidenzintervall Untergrenze	Obergrenze
Täterinnen	vs.	*Zeuginnen*	-2.833	0.862	.006	-4.976	-0.691
Täterinnen	vs.	*falsche Zeuginnen*	0.480	0.862	.856	-1.662	2.623
Zeuginnen	vs.	*falsche Zeuginnen*	3.314	0.862	.001	1.171	5.456

Tabelle F.43. Einfaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person als UV und dem Gesamtscore über alle 18 Glaubhaftigkeitskriterien als AV, wobei anstelle des mittleren Ratings der drei Rater bei Kriterium 9 das Rating von Rater C bei Kriterium 9 in den Gesamtscore einging

Q.d.V.	QS	df	MQ	F	p
Gruppe	111.065	2	55.533	5.222	.007
Fehler	1052.778	99	10.634		
Gesamt	1163.843	101			

Tabelle F.44. Scheffé-Tests für paarweise Vergleiche der experimentellen Gruppen bezüglich der Gesamtscores über alle 18 Glaubhaftigkeitskriterien, wobei anstelle des mittleren Ratings der drei Rater bei Kriterium 9 das Rating von Rater C bei Kriterium 9 in den Gesamtscore einging

Gruppenvergleich			Mittlere Differenz	Standard-fehler	p	95%-Konfidenzintervall Untergrenze	Obergrenze
Täterinnen	vs.	*Zeuginnen*	-2.010	0.791	.044	-3.975	-0.044
Täterinnen	vs.	*falsche Zeuginnen*	0.363	0.791	.900	-1.603	2.328
Zeuginnen	vs.	*falsche Zeuginnen*	2.373	0.791	.013	0.407	4.338

Tabelle F.45. Zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Rater als UVn und dem klinisch-intuitiven Gesamturteil als AV

Q.d.V.		QS	df	MQ	F	p
Gruppe		245.431	2	122.716	14.250	.000
Fehler (Gruppe)		852.569	99	8.612		
Rater	unkorrigiert	169.588	2	84.794	24.108	.000
	Greenh.-Geisser *		1.749	96.987	24.108	.000
Rater × Gruppe	unkorrigiert	115.980	4	28.995	8.243	.000
	Greenh.-Geisser *		3.497	33.165	8.243	.000
Fehler (Rater)	unkorrigiert	696.431	198	3.517		
	Greenh.-Geisser *		173.107	4.023		

Anmerkung: * $\varepsilon = .874$.

Tabelle F.46. Einfaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person als UV und dem klinisch-intuitiven Gesamturteil von Rater A als AV

Q.d.V.	QS	df	MQ	F	p
Gruppe	38.843	2	19.422	4.943	.00899
Fehler	389.000	99	3.929		
Gesamt	427.843	101			

Tabelle F.47. Scheffé-Tests für paarweise Vergleiche der experimentellen Gruppen bezüglich der klinisch-intuitiven Gesamtbeurteilung durch Rater A

Gruppenvergleich			Mittlere Differenz	Standard-fehler	p	95%-Konfidenzintervall Untergrenze	Obergrenze
Täterinnen	vs.	*Zeuginnen*	-1.294	0.481	.03031	-2.489	-0.09930
Täterinnen	vs.	*falsche Zeuginnen*	-1.324	0.481	.02596	-2.518	-0.12871
Zeuginnen	vs.	*falsche Zeuginnen*	-0.029	0.481	.99813	-1.224	1.16541

Tabelle F.48. Einfaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person als UV und dem klinisch-intuitiven Gesamturteil von Rater B als AV

Q.d.V.	QS	df	MQ	F	p
Gruppe	124.373	2	62.186	11.590	.00003
Fehler	531.206	99	5.366		
Gesamt	655.578	101			

Tabelle F.49. Scheffé-Tests für paarweise Vergleiche der experimentellen Gruppen bezüglich der klinisch-intuitiven Gesamtbeurteilung durch Rater B

Gruppenvergleich			Mittlere Differenz	Standard-fehler	p	95%-Konfidenzintervall Untergrenze	Obergrenze
Täterinnen	vs.	*Zeuginnen*	-2.676	0.562	.00004	-4.073	-1.28023
Täterinnen	vs.	*falsche Zeuginnen*	-1.000	0.562	.21028	-2.396	0.39624
Zeuginnen	vs.	*falsche Zeuginnen*	1.676	0.562	.01408	0.280	3.07271

Tabelle F.50. Einfaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person als UV und dem klinisch-intuitiven Gesamturteil von Rater C als AV

Q.d.V.	QS	df	MQ	F	p
Gruppe	198.196	2	99.098	15.602	.00000
Fehler	628.794	99	6.351		
Gesamt	826.990	101			

Tabelle F.51. Scheffé-Tests für paarweise Vergleiche der experimentellen Gruppen bezüglich der klinisch-intuitiven Gesamtbeurteilung durch Rater C

Gruppenvergleich			Mittlere Differenz	Standard-fehler	p	95%-Konfidenzintervall Untergrenze	Obergrenze
Täterinnen	vs.	*Zeuginnen*	-2.235	0.611	.00189	-3.754	-0.71621
Täterinnen	vs.	*falsche Zeuginnen*	1.118	0.611	.19319	-0.401	2.63673
Zeuginnen	vs.	*falsche Zeuginnen*	3.353	0.611	.00000	1.834	4.87203

Tabelle F.52. Anschlußtests (geringste signifikante Differenz) für paarweise Vergleiche der Rater bezüglich der klinisch-intuitiven Gesamtbeurteilung der Aussagen

verglichene Itemtypen			Mittlere Differenz	Standard-fehler	p	95%-Konfidenzintervall Untergrenze	Obergrenze
Rater A	vs.	Rater B	1.735	0.208	.000	1.322	2.149
Rater A	vs.	Rater C	0.382	0.294	.196	-0.201	0.966
Rater B	vs.	Rater C	-1.353	0.278	.000	-1.904	-0.802

Tabelle F.53. Scheffé-Tests für paarweise Vergleiche der experimentellen Gruppen bezüglich der klinisch-intuitiven Gesamtbeurteilung

verglichene Itemtypen			Mittlere Differenz	Standard-fehler	p	95%-Konfidenzintervall Untergrenze	Obergrenze
Täterinnen	vs.	*Zeuginnen*	-2.069	0.411	.000	-3.090	-1.047
Täterinnen	vs.	*falsche Zeuginnen*	-0.402	0.411	.621	-1.423	0.619
Zeuginnen	vs.	*falsche Zeuginnen*	1.667	0.411	.000	0.645	2.688

Tabelle F.54. Einfaktorielle MANCOVA mit dem Gruppenfaktor Status der aussagenden Person (*Täterinnen*, *Zeuginnen*, *falsche Zeuginnen*) als UV, den Kontrollvariablen „Motivation, beim Ablegen der Zeugenaussage glaubhaft zu erscheinen", „subjektiv eingeschätzte persönliche Erfolgswahrscheinlichkeit beim Ablegen der Zeugenaussage", „Manipulationsmaßnahmen beim Ablegen der Zeugenaussage", „intellektuelle Begabung", „Vorerfahrung mit Diebstahl", „Vorkenntnisse zur Glaubhaftigkeitsbeurteilung" sowie „frühere Teilnahme an Untersuchungen zur Glaubhaftigkeitsbeurteilung" als Kovariaten und den 18 Glaubhaftigkeitskriterien als AVn

Teststatistik	Wert	F	Hypothese df	Fehler df	p	η^2	η^2 *
Pillais Spurkriterium PS	0.772	2.656	36	152	.000	.386	.436
Wilks Likelihood-Quotient Λ	0.345	2.927	36	150	.000	.413	.462
Hotellings Spurkriterium T	1.558	3.203	36	148	.000	.438	.486
Roys größter Eigenwert	1.296	5.473	18	76	.000	.564	.604

Anmerkung: η^2 * = Effektstärke bei Nicht-Herauspartialisierung der Kovariaten (MANOVA, s. Tabelle F.29).

Tabelle F.55. Einfaktorielle ANCOVAs mit dem Gruppenfaktor Status der aussagenden Person als UV, den Kontrollvariablen „Motivation, beim Ablegen der Zeugenaussage glaubhaft zu erscheinen“, „subjektiv eingeschätzte persönliche Erfolgswahrscheinlichkeit beim Ablegen der Zeugenaussage“, „Manipulationsmaßnahmen beim Ablegen der Zeugenaussage“, „intellektuelle Begabung“, „Vorerfahrung mit Diebstahl“, „Vorkenntnisse zur Glaubhaftigkeitsbeurteilung“ sowie „frühere Teilnahme an Untersuchungen zur Glaubhaftigkeitsbeurteilung“ als Kovariaten und dem Ausprägungsgrad des jeweiligen Glaubhaftigkeitskriteriums als AV

	Kriterium	Q.d.V.	QS	df	MQ	F	p	η^2	η^2 *
1.	*Konsistenz*	Gruppe	0.000	2	0.000	0.005	.99477	.000	.001
		Fehler	2.803	92	0.030				
2.	*Unordnung*	Gruppe	1.100	2	0.550	2.462	.09087	.051	.041
		Fehler	20.559	92	0.223				
3.	*Details*	Gruppe	9.061	2	4.531	11.420	.00004	.199	.226
		Fehler	36.501	92	0.397				
4.	*Verknüpfungen*	Gruppe	2.958	2	1.479	3.979	.02201	.080	.122
		Fehler	34.198	92	0.372				
5.	*Interaktionen*	Gruppe	0.437	2	0.218	1.223	.29910	.026	.037
		Fehler	16.429	92	0.179				
6.	*Gespräche*	Gruppe	10.954	2	5.477	8.517	.00040	.156	.193
		Fehler	59.163	92	0.643				
7.	*Komplikationen*	Gruppe	0.031	2	0.016	0.548	.57998	.012	.013
		Fehler	2.624	92	0.029				
8.	*Ausgefallenes*	Gruppe	2.611	2	1.305	18.239	.00000	.284	.268
		Fehler	6.585	92	0.072				
9.	*Nebensächliches*	Gruppe	7.987	2	3.994	14.474	.00000	.239	.209
		Fehler	25.385	92	0.276				

Tabelle F.55 (Fortsetzung). Einfaktorielle ANCOVAs mit dem Gruppenfaktor Status der aussagenden Person als UV, den Kontrollvariablen „Motivation, beim Ablegen der Zeugenaussage glaubhaft zu erscheinen“, „subjektiv eingeschätzte persönliche Erfolgswahrscheinlichkeit beim Ablegen der Zeugenaussage“, „Manipulationsmaßnahmen beim Ablegen der Zeugenaussage“, „intellektuelle Begabung“, „Vorerfahrung mit Diebstahl“, „Vorkenntnisse zur Glaubhaftigkeitsbeurteilung“ sowie „frühere Teilnahme an Untersuchungen zur Glaubhaftigkeitsbeurteilung“ als Kovariaten und dem Ausprägungsgrad des jeweiligen Glaubhaftigkeitskriteriums als AV

	Kriterium	Q.d.V.	QS	df	MQ	F	p	η^2	η^2 *
10.	*Unverstandenes*	Gruppe	0.017	2	0.008	0.975	.38108	.021	.031
		Fehler	0.786	92	0.009				
11.	*Indirektes*	Gruppe	0.006	2	0.003	0.853	.42944	.018	.020
		Fehler	0.304	92	0.003				
12.	*Eigenseelisches*	Gruppe	0.279	2	0.140	0.409	.66565	.009	.017
		Fehler	31.440	92	0.342				
13.	*Fremdseelisches*	Gruppe	0.408	2	0.204	0.648	.52551	.014	.017
		Fehler	28.953	92	0.315				
14.	*Verbesserungen*	Gruppe	1.735	2	0.868	3.125	.04865	.064	.084
		Fehler	25.544	92	0.278				
15.	*Erinnerungslücken*	Gruppe	0.940	2	0.470	5.859	.00403	.113	.067
		Fehler	7.382	92	0.080				
16.	*Selbsteinwände*	Gruppe	0.018	2	0.009	1.046	.35560	.022	.003
		Fehler	0.777	92	0.008				
17.	*Eigenbelastung*	Gruppe	0.004	2	0.002	0.239	.78753	.005	.010
		Fehler	0.766	92	0.008				
18.	*Fremdentlastung*	Gruppe	0.423	2	0.211	1.486	.23157	.031	.050
		Fehler	13.081	92	0.142				

Anmerkung: η^2 * = Effektstärke bei Nicht-Herauspartialisierung der Kovariaten (ANOVAs, s. Tabelle F.33).

Tabelle F.56. Einfaktorielle ANCOVA mit dem Gruppenfaktor Status der aussagenden Person als UV, den Kontrollvariablen „Motivation, beim Ablegen der Zeugenaussage glaubhaft zu erscheinen“, „subjektiv eingeschätzte persönliche Erfolgswahrscheinlichkeit beim Ablegen der Zeugenaussage“, „Manipulationsmaßnahmen beim Ablegen der Zeugenaussage“, „intellektuelle Begabung“, „Vorerfahrung mit Diebstahl“, „Vorkenntnisse zur Glaubhaftigkeitsbeurteilung“ sowie „frühere Teilnahme an Untersuchungen zur Glaubhaftigkeitsbeurteilung“ als Kovariaten und der Höhe des über alle 18 Glaubhaftigkeitskriterien aufsummierten Gesamtscores als AV

Q.d.V.	QS	df	MQ	F	p	η^2	η^2 *
Gruppe	122.670	2	61.335	6.493	.002	.124	.100
Fehler	869.093	92	9.447				

Anmerkung: η^2 * = Effektstärke bei Nicht-Herauspartialisierung der Kovariaten (ANOVA, s. Tabelle F.37).

Tabelle F.57. Diskriminanzanalyse mit den 18 Glaubhaftigkeitskriterien als Prädiktoren und der tatsächlichen Glaubhaftigkeit der Aussagen (glaubhaft vs. unglaubhaft) als Kriterium (Berechnung anhand einer Zufallsstichprobe von jeweils 17 *Täterinnen*, *Zeuginnen* und *falschen Zeuginnen*)

Faktor	Eigenwert λ	% der Varianz	Kumulierte %	Kanonische Korrelation
1	1.600	100	100	.784

Signifikanzprüfung

Faktor(en)	Wilks Likelihood-Quotient Λ	χ^2	df	p
1	.385	38.221	18	.004

Kriterium	Diskriminanzkoeffizienten	Faktorladungen
1. *Konsistenz*	-0.103	-.042
2. *Unordnung*	-0.165	.049
3. *Details*	0.342	.356
4. *Verknüpfungen*	-0.331	-.127
5. *Interaktionen*	0.243	.105
6. *Gespräche*	0.615	.418
7. *Komplikationen*	-0.124	-.021
8. *Ausgefallenes*	0.777	.369
9. *Nebensächliches*	0.211	.206
10. *Unverstandenes*	-1.057	-.114
11. *Indirektes*	0.248	.057
12. *Eigenseelisches*	-0.576	-.087
13. *Fremdseelisches*	0.280	.054
14. *Verbesserungen*	-0.018	.154
15. *Erinnerungslücken*	0.327	.200
16. *Selbsteinwände*	0.093	.103
17. *Eigenbelastung*	1.009	-.079
18. *Fremdentlastung*	-0.342	-.095

Tabelle F.58. Diskriminanzanalyse mit dem über alle 18 Glaubhaftigkeitskriterien aufsummierten Gesamtscore als einzigem Prädiktor und der tatsächlichen Glaubhaftigkeit der Aussagen (glaubhaft vs. unglaubhaft) als Kriterium (Berechnung anhand einer Zufallsstichprobe von jeweils 17 *Täterinnen*, *Zeuginnen* und *falschen Zeuginnen*)

Faktor	Eigenwert λ	% der Varianz	Kumulierte %	Kanonische Korrelation
1	.092	100	100	.290

	Signifikanzprüfung			
Faktor(en)	Wilks Likelihood-Quotient Λ	χ^2	df	p
1	.916	4.253	1	.039

Prädiktor	Diskriminanzkoeffizient	Faktorladung
Gesamtscore	1.000	1.000

Tabelle F.59. Einfaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person als UV und dem numerischen *GAT*-Score als AV (SCR-Quantifizierungsmethode A)

Q.d.V.	QS	df	MQ	F	p
Gruppe	863.118	2	431.559	17.362	.000
Fehler	2460.735	99	24.856		
Gesamt	3323.853	101			

Tabelle F.60. Scheffé-Tests für paarweise Vergleiche der experimentellen Gruppen bezüglich der numerischen Scores im *GAT* (SCR-Quantifizierungsmethode A)

Gruppenvergleich			Mittlere Differenz	Standard-fehler	p	95%-Konfidenzintervall Untergrenze	Obergrenze
Täterinnen	vs.	*Zeuginnen*	0.794	1.209	.806	-2.211	3.799
Täterinnen	vs.	*falsche Zeuginnen*	6.529	1.209	.000	3.524	9.535
Zeuginnen	vs.	*falsche Zeuginnen*	5.735	1.209	.000	2.730	8.740

Tabelle F.61. Gesamt-*GAT*: Zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Itemtyp als UVn und den SCR-Amplituden als AV (SCR-Quantifizierungsmethode A)

Q.d.V.		QS	df	MQ	F	p
Gruppe		0.081	2	0.040	2.845	.063
Fehler (Gruppe)		1.402	99	0.014		
Itemtyp	Unkorrigiert	0.227	2	0.114	38.949	.000
	Greenh.-Geisser *		1.169	0.194	38.949	.000
Itemtyp × Gruppe	Unkorrigiert	0.090	4	0.023	7.733	.000
	Greenh.-Geisser *		2.339	0.039	7.733	.000
Fehler (Itemtyp)	Unkorrigiert	0.577	198	0.003		
	Greenh.-Geisser *		115.772	0.005		

Anmerkung: * ε = .585.

Tabelle F.62. t-Test für unabhängige Stichproben zur Überprüfung des Unterschieds zwischen *Täterinnen* und *Zeuginnen* in der SCR-Magnitude bei den irrelevanten Items (SCR-Quantifizierungsmethode A)

					95%-Konfidenzintervall	
t	df	p *	Mittlere Differenz	Stand.-fehler der Differenz	Untergrenze	Obergrenze
0.587	66.000	.559	0.009	0.015	-0.021	0.039

Anmerkung: * zweiseitiger Test.

Tabelle F.63. t-Test für unabhängige Stichproben zur Überprüfung des Unterschieds zwischen *Täterinnen* und *falschen Zeuginnen* in der SCR-Magnitude bei den irrelevanten Items (SCR-Quantifizierungsmethode A)

					95%-Konfidenzintervall	
t	df	p *	Mittlere Differenz	Stand.-fehler der Differenz	Untergrenze	Obergrenze
0.806	66.000	.423	0.008	0.010	-0.012	0.029

Anmerkung: * zweiseitiger Test.

Tabelle F.64. t-Test für unabhängige Stichproben zur Überprüfung des Unterschieds zwischen *Zeuginnen* und *falschen Zeuginnen* in der SCR-Magnitude bei den irrelevanten Items (SCR-Quantifizierungsmethode A)

					95%-Konfidenzintervall	
t	df	p *	Mittlere Differenz	Stand.-fehler der Differenz	Untergrenze	Obergrenze
-0.042	66.000	.967	-0.001	0.014	-0.028	0.027

Anmerkung: * zweiseitiger Test.

Tabelle F.65. t-Test für unabhängige Stichproben zur Überprüfung des Unterschieds zwischen *Täterinnen* und *Zeuginnen* in der SCR-Magnitude bei den relevanten Items (SCR-Quantifizierungsmethode A)

					95%-Konfidenzintervall	
t	df	p *	Mittlere Differenz	Stand.-fehler der Differenz	Untergrenze	Obergrenze
0.721	66.000	.474	0.026	0.035	-0.045	0.096

Anmerkung: * zweiseitiger Test.

Tabelle F.66. t-Test für unabhängige Stichproben zur Überprüfung des Unterschieds zwischen *Täterinnen* und *falschen Zeuginnen* in der SCR-Magnitude bei den relevanten Items (SCR-Quantifizierungsmethode A)

					95%-Konfidenzintervall	
t	df	p *	Mittlere Differenz	Stand.-fehler der Differenz	Untergrenze	Obergrenze
4.257	66.000	.000	0.094	0.022	0.050	0.138

Anmerkung: * zweiseitiger Test.

Tabelle F.67. t-Test für unabhängige Stichproben zur Überprüfung des Unterschieds zwischen *Zeuginnen* und *falschen Zeuginnen* in der SCR-Magnitude bei den relevanten Items (SCR-Quantifizierungsmethode A)

t	df	p *	Mittlere Differenz	Stand.-fehler der Differenz	95%-Konfidenzintervall Untergrenze	95%-Konfidenzintervall Obergrenze
2.297	66.000	.025	0.069	0.030	0.009	0.128

Anmerkung: * zweiseitiger Test.

Tabelle F.68. Anschlußtests (geringste signifikante Differenz) für paarweise Vergleiche der Itemtypen (SCR-Quantifizierungsmethode A)

verglichene Itemtypen			Mittlere Differenz	Standard-fehler	p	95%-Konfidenzintervall Untergrenze	95%-Konfidenzintervall Obergrenze
Puffer	vs.	relevant	-0.061	0.010	.000	-0.081	-0.042
Puffer	vs.	irrelevant	-0.008	0.004	.034	-0.015	-0.001
relevant	vs.	irrelevant	0.053	0.008	.000	0.038	0.069

Tabelle F.69. *GAT*-Frage 1: Zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Itemtyp als UVn und den SCR-Amplituden als AV (SCR-Quantifizierungsmethode A)

Q.d.V.		QS	df	MQ	F	p
Gruppe		0.189	2	0.094	3.610	.031
Fehler (Gruppe)		2.591	99	0.026		
Itemtyp	Unkorrigiert	0.199	2	0.099	14.174	.000
	Greenh.-Geisser *	0.199	1.510	0.132	14.174	.000
Itemtyp × Gruppe	Unkorrigiert	0.217	4	0.054	7.731	.000
	Greenh.-Geisser *	0.217	3.020	0.072	7.731	.000
Fehler (Itemtyp)	Unkorrigiert	1.388	198	0.007		
	Greenh.-Geisser *	1.388	149.508	0.009		

Anmerkung: * $\varepsilon = .755$.

Tabelle F.70. *GAT*-Frage 2: Zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Itemtyp als UVn und den SCR-Amplituden als AV (SCR-Quantifizierungsmethode A)

Q.d.V.		QS	df	MQ	F	p
Gruppe		0.160	2	0.080	3.638	.030
Fehler (Gruppe)		2.174	99	0.022		
Itemtyp	Unkorrigiert	0.517	2	0.259	30.023	.000
	Greenh.-Geisser *	0.517	1.365	0.379	30.023	.000
Itemtyp × Gruppe	Unkorrigiert	0.186	4	0.046	5.398	.000
	Greenh.-Geisser *	0.186	2.731	0.068	5.398	.002
Fehler (Itemtyp)	Unkorrigiert	1.705	198	0.009		
	Greenh.-Geisser *	1.705	135.173	0.013		

Anmerkung: * $\varepsilon = ..683$.

Tabelle F.71. *GAT*-Frage 3: Zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Itemtyp als UVn und den SCR-Amplituden als AV (SCR-Quantifizierungsmethode A)

Q.d.V.		QS	df	MQ	F	p
Gruppe		0.069	2	0.034	2.128	.125
Fehler (Gruppe)		1.602	99	0.016		
Itemtyp	Unkorrigiert	0.270	2	0.135	16.809	.000
	Greenh.-Geisser *	0.270	1.186	0.228	16.809	.000
Itemtyp × Gruppe	Unkorrigiert	0.117	4	0.029	3.642	.007
	Greenh.-Geisser *	0.117	2.372	0.049	3.642	.023
Fehler (Itemtyp)	Unkorrigiert	1.589	198	0.008		
	Greenh.-Geisser *	1.589	117.418	0.014		

Anmerkung: * $\varepsilon = .593$.

Tabelle F.72. *GAT*-Frage 4: Zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Itemtyp als UVn und den SCR-Amplituden als AV (SCR-Quantifizierungsmethode A)

Q.d.V.		QS	df	MQ	F	p
Gruppe		0.125	2	0.063	3.436	.036
Fehler (Gruppe)		1.801	99	0.018		
Itemtyp	Unkorrigiert	0.199	2	0.099	16.813	.000
	Greenh.-Geisser *	0.199	1.708	0.116	16.813	.000
Itemtyp × Gruppe	Unkorrigiert	0.150	4	0.038	6.354	.000
	Greenh.-Geisser *	0.150	3.416	0.044	6.354	.000
Fehler (Itemtyp)	Unkorrigiert	1.169	198	0.006		
	Greenh.-Geisser *	1.169	169.089	0.007		

Anmerkung: * $\varepsilon = .854$.

Tabelle F.73. *GAT*-Frage 5: Zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Itemtyp als UVn und den SCR-Amplituden als AV (SCR-Quantifizierungsmethode A)

Q.d.V.		QS	df	MQ	F	p
Gruppe		0.126	2	0.063	2.445	.092
Fehler (Gruppe)		2.548	99	0.026		
Itemtyp	Unkorrigiert	0.273	2	0.137	18.550	.000
	Greenh.-Geisser *	0.273	1.261	0.217	18.550	.000
Itemtyp × Gruppe	Unkorrigiert	0.103	4	0.026	3.508	.009
	Greenh.-Geisser *	0.103	2.523	0.041	3.508	.024
Fehler (Itemtyp)	Unkorrigiert	1.458	198	0.007		
	Greenh.-Geisser *	1.458	124.865	0.012		

Anmerkung: * $\varepsilon = .631$.

Tabelle F.74. *GAT*-Frage 6: Zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Itemtyp als UVn und den SCR-Amplituden als AV (SCR-Quantifizierungsmethode A)

Q.d.V.		QS	df	MQ	F	p
Gruppe		0.101	2	0.051	2.580	.081
Fehler (Gruppe)		1.947	99	0.020		
Itemtyp	Unkorrigiert	0.389	2	0.194	27.915	.000
	Greenh.-Geisser *	0.389	1.321	0.294	27.915	.000
Itemtyp × Gruppe	Unkorrigiert	0.179	4	0.045	6.413	.000
	Greenh.-Geisser *	0.179	2.641	0.068	6.413	.001
Fehler (Itemtyp)	Unkorrigiert	1.379	198	0.007		
	Greenh.-Geisser *	1.379	130.744	0.011		

Anmerkung: * ε = .660.

Tabelle F.75. *GAT*-Frage 7: Zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Itemtyp als UVn und den SCR-Amplituden als AV (SCR-Quantifizierungsmethode A)

Q.d.V.		QS	df	MQ	F	p
Gruppe		0.088	2	0.044	2.164	.120
Fehler (Gruppe)		2.011	99	0.020		
Itemtyp	Unkorrigiert	0.152	2	0.076	17.594	.000
	Greenh.-Geisser *	0.152	1.826	0.083	17.594	.000
Itemtyp × Gruppe	Unkorrigiert	0.064	4	0.016	3.700	.006
	Greenh.-Geisser *	0.064	3.652	0.017	3.700	.008
Fehler (Itemtyp)	Unkorrigiert	0.853	198	0.004		
	Greenh.-Geisser *	0.853	180.792	0.005		

Anmerkung: * ε = .913.

Tabelle F.76. *GAT*-Frage 8: Zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Itemtyp als UVn und den SCR-Amplituden als AV (SCR-Quantifizierungsmethode A)

Q.d.V.		QS	df	MQ	F	p
Gruppe		0.037	2	0.019	1.152	.320
Fehler (Gruppe)		1.608	99	0.016		
Itemtyp	Unkorrigiert	0.166	2	0.083	11.206	.000
	Greenh.-Geisser *	0.166	1.395	0.119	11.206	.000
Itemtyp × Gruppe	Unkorrigiert	0.076	4	0.019	2.567	.039
	Greenh.-Geisser *	0.076	2.790	0.027	2.567	.061
Fehler (Itemtyp)	Unkorrigiert	1.470	198	0.007		
	Greenh.-Geisser *	1.470	138.091	0.011		

Anmerkung: * ε = .697.

Tabelle F.77. *GAT*-Frage 9: Zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Itemtyp als UVn und den SCR-Amplituden als AV (SCR-Quantifizierungsmethode A)

Q.d.V.		QS	df	MQ	F	p
Gruppe		0.006	2	0.003	0.226	.798
Fehler (Gruppe)		1.411	99	0.014		
Itemtyp	Unkorrigiert	0.137	2	0.068	10.912	.000
	Greenh.-Geisser *	0.137	1.659	0.083	10.912	.000
Itemtyp × Gruppe	Unkorrigiert	0.046	4	0.012	1.845	.122
	Greenh.-Geisser *	0.046	3.319	0.014	1.845	.135
Fehler (Itemtyp)	Unkorrigiert	1.242	198	0.006		
	Greenh.-Geisser *	1.242	164.273	0.008		

Anmerkung: * $\varepsilon = .830$.

Tabelle F.78. *GAT*-Frage 10: Zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Itemtyp als UVn und den SCR-Amplituden als AV (SCR-Quantifizierungsmethode A)

Q.d.V.		QS	df	MQ	F	p
Gruppe		0.042	2	0.021	1.170	.314
Fehler (Gruppe)		1.787	99	0.018		
Itemtyp	Unkorrigiert	0.216	2	0.108	12.738	.000
	Greenh.-Geisser *	0.216	1.477	0.146	12.738	.000
Itemtyp × Gruppe	Unkorrigiert	0.026	4	0.007	0.779	.540
	Greenh.-Geisser *	0.026	2.953	0.009	0.779	.506
Fehler (Itemtyp)	Unkorrigiert	1.676	198	0.008		
	Greenh.-Geisser *	1.676	146.186	0.011		

Anmerkung: * $\varepsilon = .738$.

Tabelle F.79. Einfaktorielle ANCOVA mit dem Gruppenfaktor Status der aussagenden Person als UV, den Kontrollvariablen „Elektrodermale Labilität“, „Motivation, im *GAT* einen unschuldigen Eindruck zu hinterlassen“, „subjektiv eingeschätzte persönliche Erfolgswahrscheinlichkeit im *GAT*“, „Manipulationsmaßnahmen im *GAT*“, „Vorkenntnisse zur Glaubhaftigkeitsbeurteilung“ sowie „frühere Teilnahme an Untersuchungen zur Glaubhaftigkeitsbeurteilung“ als Kovariaten und der Höhe der numerischen *GAT*-Scores als AV (SCR-Quantifizierungsmethode A)

Q.d.V.	QS	df	MQ	F	p	η^2	η^2 *
Gruppe	677.586	2	338.793	15.882	.000	.255	.260
Fehler	1983.830	93	21.332				

Anmerkung: η^2 * = Effektstärke bei Nicht-Herauspartialisierung der Kovariaten (ANOVA, s. Tabelle F.59).

Tabelle F.80. Gesamt-*GAT*: Zweifaktorielle ANCOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Itemtyp als UVn, den Kontrollvariablen „Elektrodermale Labilität“, „Motivation, im *GAT* einen unschuldigen Eindruck zu hinterlassen“, „subjektiv eingeschätzte persönliche Erfolgswahrscheinlichkeit im *GAT*“, „Manipulationsmaßnahmen im *GAT*“, „Vorkenntnisse zur Glaubhaftigkeitsbeurteilung“ sowie „frühere Teilnahme an Untersuchungen zur Glaubhaftigkeitsbeurteilung“ als Kovariaten und den SCR-Amplituden als AV (SCR-Quantifizierungsmethode A)

Q.d.V.		QS	df	MQ	F	p	η^2	η^2 *
Gruppe		0.059	2	0.029	2.547	.084	.052	.054
Fehler (Gruppe)		1.069	93	0.011				
Itemtyp	Unkorrigiert	0.011	2	0.005	1.916	.150	.020	.282
	Greenh.-Geisser [a]	0.011	1.180	0.009	1.916	.167	.020	.282
Itemtyp × Gruppe	Unkorrigiert	0.068	4	0.017	6.228	.000	.118	.135
	Greenh.-Geisser [a]	0.068	2.360	0.029	6.228	.002	.118	.135
Fehler (Itemtyp)	Unkorrigiert	0.511	186	0.003				
	Greenh.-Geisser [a]	0.511	109.732	0.005				

Anmerkung: [a] $\varepsilon = .590$; η^2 * = Effektstärke bei Nicht-Herauspartialisierung der Kovariaten (ANOVA, s. Tabelle F.61).

Tabelle F.81. Diskriminanzanalyse mit dem numerischen *GAT*-Score als einzigem Prädiktor und der experimentellen Gruppenzugehörigkeit als Kriterium (Berechnung anhand einer Zufallsstichprobe von jeweils 17 *Täterinnen*, *Zeuginnen* und *falschen Zeuginnen*) (SCR-Quantifizierungsmethode A)

Faktor	Eigenwert λ	% der Varianz	Kumulierte %	Kanonische Korrelation
1	.443	100	100	.554

Signifikanzprüfung

Faktor	Wilks Likelihood-Quotient Λ	χ^2	df	p
1	.693	17.619	2	.000

Prädiktor	Diskriminanzkoeffizient	Faktorladung
Numerischer Score	1.000	1.000

Tabelle F.82. Diskriminanzanalyse mit der intraindividuellen SCR-Magnituden-Differenz zwischen relevanten und irrelevanten Items als Prädiktor und der experimentellen Gruppenzugehörigkeit als Kriterium (Berechnung anhand einer Zufallsstichprobe von jeweils 17 *Täterinnen*, *Zeuginnen* und *falschen Zeuginnen*) (SCR-Quantifizierungsmethode A)

Faktor	Eigenwert λ	% der Varianz	Kumulierte %	Kanonische Korrelation
1	.229	100	100	.431
		Signifikanzprüfung		
Faktor	Wilks Likelihood-Quotient Λ	χ^2	df	p
1	.814	9.879	2	.007
Prädiktor		Diskriminanzkoeffizient		Faktorladung
SCR-Magnituden-Differenz		1.000		1.000

Tabelle F.83. Einfaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person als UV und dem numerischen *GAT*-Score als AV (SCR-Quantifizierungsmethode B)

Q.d.V.	QS	df	MQ	F	p
Gruppe	863.118	2	431.559	17.362	.000
Fehler	2460.735	99	24.856		
Gesamt	3323.853	101			

Tabelle F.84. Scheffé-Tests für paarweise Vergleiche der experimentellen Gruppen bezüglich der numerischen Scores im *GAT* (SCR-Quantifizierungsmethode B)

Gruppenvergleich			Mittlere Differenz	Standard-fehler	p	95%-Konfidenzintervall Untergrenze	Obergrenze
Täterinnen	vs.	*Zeuginnen*	0.588	1.189	.885	-2.366	3.543
Täterinnen	vs.	*falsche Zeuginnen*	5.941	1.189	.000	2.987	8.896
Zeuginnen	vs.	*falsche Zeuginnen*	5.353	1.189	.000	2.398	8.307

Tabelle F.85. Gesamt-*GAT*: Zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Itemtyp als UVn und den SCR-Amplituden als AV (SCR-Quantifizierungsmethode B)

Q.d.V.		QS	df	MQ	F	p
Gruppe		0.082	2	0.041	2.230	.113
Fehler (Gruppe)		1.824	99	0.018		
Itemtyp	Unkorrigiert	.168	2	0.084	37.486	.000
	Greenh.-Geisser *		1.183	0.142	37.486	.000
Itemtyp × Gruppe	Unkorrigiert	.077	4	0.019	8.567	.000
	Greenh.-Geisser *		2.367	0.032	8.567	.000
Fehler (Itemtyp)	Unkorrigiert	.442	198	0.002		
	Greenh.-Geisser *		117.158	0.004		

Anmerkung: * ε = .592.

Tabelle F.86. t-Test für unabhängige Stichproben zur Überprüfung des Unterschieds zwischen *Täterinnen* und *Zeuginnen* in der SCR-Magnitude bei den irrelevanten Items (SCR-Quantifizierungsmethode B)

					95%-Konfidenzintervall	
t	df	p *	Mittlere Differenz	Stand.-fehler der Differenz	Untergrenze	Obergrenze
0.744	66	.459	0.013	0.017	-0.022	0.048

Anmerkung: * zweiseitiger Test.

Tabelle F.87. t-Test für unabhängige Stichproben zur Überprüfung des Unterschieds zwischen *Täterinnen* und *falschen Zeuginnen* in der SCR-Magnitude bei den irrelevanten Items (SCR-Quantifizierungsmethode B)

					95%-Konfidenzintervall	
t	df	p *	Mittlere Differenz	Stand.-fehler der Differenz	Untergrenze	Obergrenze
0.828	66	.411	0.010	0.013	-0.015	0.036

Anmerkung: * zweiseitiger Test.

Tabelle F.88. t-Test für unabhängige Stichproben zur Überprüfung des Unterschieds zwischen *Zeuginnen* und *falschen Zeuginnen* in der SCR-Magnitude bei den irrelevanten Items (SCR-Quantifizierungsmethode B)

					95%-Konfidenzintervall	
t	df	p *	Mittlere Differenz	Stand.-fehler der Differenz	Untergrenze	Obergrenze
-0.154	66	.878	-0.002	0.016	-0.035	0.030

Anmerkung: * zweiseitiger Test.

Tabelle F.89. t-Test für unabhängige Stichproben zur Überprüfung des Unterschieds zwischen *Täterinnen* und *Zeuginnen* in der SCR-Magnitude bei den relevanten Items (SCR-Quantifizierungsmethode B)

					95%-Konfidenzintervall	
t	df	p *	Mittlere Differenz	Stand.-fehler der Differenz	Untergrenze	Obergrenze
0.803	66	.425	0.028	0.035	-0.042	0.099

Anmerkung: * zweiseitiger Test.

Tabelle F.90. t-Test für unabhängige Stichproben zur Überprüfung des Unterschieds zwischen *Täterinnen* und *falschen Zeuginnen* in der SCR-Magnitude bei den relevanten Items (SCR-Quantifizierungsmethode B)

					95%-Konfidenzintervall	
t	df	p *	Mittlere Differenz	Stand.-fehler der Differenz	Untergrenze	Obergrenze
4.006	66	.000	0.091	0.023	0.046	0.137

Anmerkung: * zweiseitiger Test.

Tabelle F.91. t-Test für unabhängige Stichproben zur Überprüfung des Unterschieds zwischen *Zeuginnen* und *falschen Zeuginnen* in der SCR-Magnitude bei den relevanten Items (SCR-Quantifizierungsmethode B)

t	df	p *	Mittlere Differenz	Stand.-fehler der Differenz	95%-Konfidenzintervall Untergrenze	95%-Konfidenzintervall Obergrenze
2.087	66	.041	0.063	0.030	0.003	0.123

Anmerkung: * zweiseitiger Test.

Tabelle F.92. Anschlußtests (geringste signifikante Differenz) für paarweise Vergleiche der Itemtypen (SCR-Quantifizierungsmethode B)

verglichene Itemtypen			Mittlere Differenz	Standard-fehler	p	95%-Konfidenzintervall Untergrenze	95%-Konfidenzintervall Obergrenze
Puffer	vs.	relevant	-0.052	0.008	.000	-0.068	-0.035
Puffer	vs.	irrelevant	-0.004	0.003	.154	-0.010	0.002
relevant	vs.	irrelevant	0.047	0.007	.000	0.033	0.062

Tabelle F.93. *GAT*-Frage 1: Zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Itemtyp als UVn und den SCR-Amplituden als AV (SCR-Quantifizierungsmethode B)

Q.d.V.		QS	df	MQ	F	p
Gruppe		0.216	2	0.108	3.897	.023
Fehler (Gruppe)		2.743	99	0.028		
Itemtyp	Unkorrigiert	0.191	2	0.095	14.140	.000
	Greenh.-Geisser *	0.191	1.530	0.125	14.140	.000
Itemtyp × Gruppe	Unkorrigiert	0.209	4	0.052	7.752	.000
	Greenh.-Geisser *	0.209	3.060	0.068	7.752	.000
Fehler (Itemtyp)	Unkorrigiert	1.334	198	0.007		
	Greenh.-Geisser *	1.334	151.459	0.009		

Anmerkung: * $\varepsilon = .765$.

Tabelle F.94. *GAT*-Frage 2: Zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Itemtyp als UVn und den SCR-Amplituden als AV (SCR-Quantifizierungsmethode B)

Q.d.V.		QS	df	MQ	F	p
Gruppe		0.159	2	0.080	3.448	.036
Fehler (Gruppe)		2.284	99	0.023		
Itemtyp	Unkorrigiert	0.409	2	0.205	21.847	.000
	Greenh.-Geisser *	0.409	1.413	0.290	21.847	.000
Itemtyp × Gruppe	Unkorrigiert	0.171	4	0.043	4.569	.001
	Greenh.-Geisser *	0.171	2.825	0.061	4.569	.005
Fehler (Itemtyp)	Unkorrigiert	1.854	198	0.009		
	Greenh.-Geisser *	1.854	139.844	0.013		

Anmerkung: * $\varepsilon = .706$.

Tabelle F.95. *GAT*-Frage 3: Zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Itemtyp als UVn und den SCR-Amplituden als AV (SCR-Quantifizierungsmethode B)

Q.d.V.		QS	df	MQ	F	p
Gruppe		0.080	2	0.040	1.845	.163
Fehler (Gruppe)		2.147	99	0.022		
Itemtyp	Unkorrigiert	0.204	2	0.102	14.170	.000
	Greenh.-Geisser *	0.204	1.376	0.148	14.170	.000
Itemtyp × Gruppe	Unkorrigiert	0.134	4	0.034	4.656	.001
	Greenh.-Geisser *	0.134	2.751	0.049	4.656	.005
Fehler (Itemtyp)	Unkorrigiert	1.427	198	0.007		
	Greenh.-Geisser *	1.427	136.190	0.010		

Anmerkung: * ε = .688.

Tabelle F.96. *GAT*-Frage 4: Zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Itemtyp als UVn und den SCR-Amplituden als AV (SCR-Quantifizierungsmethode B)

Q.d.V.		QS	df	MQ	F	p
Gruppe		0.139	2	0.069	3.298	.041
Fehler (Gruppe)		2.080	99	0.021		
Itemtyp	Unkorrigiert	0.161	2	0.081	13.113	.000
	Greenh.-Geisser *	0.161	1.772	0.091	13.113	.000
Itemtyp × Gruppe	Unkorrigiert	0.147	4	0.037	5.980	.000
	Greenh.-Geisser *	0.147	3.543	0.041	5.980	.000
Fehler (Itemtyp)	Unkorrigiert	1.216	198	0.006		
	Greenh.-Geisser *	1.216	175.396	0.007		

Anmerkung: * ε = .886.

Tabelle F.97. *GAT*-Frage 5: Zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Itemtyp als UVn und den SCR-Amplituden als AV (SCR-Quantifizierungsmethode B)

Q.d.V.		QS	df	MQ	F	p
Gruppe		0.130	2	0.065	1.574	.212
Fehler (Gruppe)		4.081	99	0.041		
Itemtyp	Unkorrigiert	0.132	2	0.066	9.964	.000
	Greenh.-Geisser *	0.132	1.764	0.075	9.964	.000
Itemtyp × Gruppe	Unkorrigiert	0.124	4	0.031	4.669	.001
	Greenh.-Geisser *	0.124	3.529	0.035	4.669	.002
Fehler (Itemtyp)	Unkorrigiert	1.312	198	0.007		
	Greenh.-Geisser *	1.312	174.670	0.008		

Anmerkung: * ε = .882.

Tabelle F.98. *GAT*-Frage 6: Zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Itemtyp als UVn und den SCR-Amplituden als AV (SCR-Quantifizierungsmethode B)

Q.d.V.		QS	df	MQ	F	p
Gruppe		0.112	2	0.056	2.529	.085
Fehler (Gruppe)		2.193	99	0.022		
Itemtyp	Unkorrigiert	0.342	2	0.171	24.967	.000
	Greenh.-Geisser *	0.342	1.392	0.246	24.967	.000
Itemtyp × Gruppe	Unkorrigiert	0.134	4	0.033	4.883	.001
	Greenh.-Geisser *	0.134	2.784	0.048	4.883	.004
Fehler (Itemtyp)	Unkorrigiert	1.355	198	0.007		
	Greenh.-Geisser *	1.355	137.788	0.010		

Anmerkung: * $\varepsilon = .696$.

Tabelle F.99. *GAT*-Frage 7: Zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Itemtyp als UVn und den SCR-Amplituden als AV (SCR-Quantifizierungsmethode B)

Q.d.V.		QS	df	MQ	F	p
Gruppe		0.083	2	0.041	1.623	.203
Fehler (Gruppe)		2.522	99	0.025		
Itemtyp	Unkorrigiert	0.155	2	0.078	16.144	.000
	Greenh.-Geisser *	0.155	1.583	0.098	16.144	.000
Itemtyp × Gruppe	Unkorrigiert	0.016	4	0.004	0.831	.507
	Greenh.-Geisser *	0.016	3.166	0.005	0.831	.484
Fehler (Itemtyp)	Unkorrigiert	0.951	198	0.005		
	Greenh.-Geisser *	0.951	156.723	0.006		

Anmerkung: * $\varepsilon = .792$.

Tabelle F.100. *GAT*-Frage 8: Zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Itemtyp als UVn und den SCR-Amplituden als AV (SCR-Quantifizierungsmethode B)

Q.d.V.		QS	df	MQ	F	p
Gruppe		0.018	2	0.009	0.368	.693
Fehler (Gruppe)		2.407	99	0.024		
Itemtyp	Unkorrigiert	0.064	2	0.032	4.683	.010
	Greenh.-Geisser *	0.064	1.723	0.037	4.683	.014
Itemtyp × Gruppe	Unkorrigiert	0.065	4	0.016	2.397	.052
	Greenh.-Geisser *	0.065	3.447	0.019	2.397	.061
Fehler (Itemtyp)	Unkorrigiert	1.350	198	0.007		
	Greenh.-Geisser *	1.350	170.616	0.008		

Anmerkung: * $\varepsilon = .862$.

Tabelle F.101. *GAT*-Frage 9: Zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Itemtyp als UVn und den SCR-Amplituden als AV (SCR-Quantifizierungsmethode B)

Q.d.V.		QS	df	MQ	F	p
Gruppe		0.007	2	0.004	0.159	.853
Fehler (Gruppe)		2.204	99	0.022		
Itemtyp	Unkorrigiert	0.094	2	0.047	7.944	.000
	Greenh.-Geisser *	0.094	1.840	0.051	7.944	.001
Itemtyp × Gruppe	Unkorrigiert	0.043	4	0.011	1.811	.128
	Greenh.-Geisser *	0.043	3.680	0.012	1.811	.134
Fehler (Itemtyp)	Unkorrigiert	1.175	198	0.006		
	Greenh.-Geisser *	1.175	182.144	0.006		

Anmerkung: * $\varepsilon = .920$.

Tabelle F.102. *GAT*-Frage 10: Zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Itemtyp als UVn und den SCR-Amplituden als AV (SCR-Quantifizierungsmethode B)

Q.d.V.		QS	df	MQ	F	p
Gruppe		0.028	2	0.014	0.551	.578
Fehler (Gruppe)		2.510	99	0.025		
Itemtyp	Unkorrigiert	0.151	2	0.076	8.172	.000
	Greenh.-Geisser *	0.151	1.610	0.094	8.172	.001
Itemtyp × Gruppe	Unkorrigiert	0.025	4	0.006	0.678	.608
	Greenh.-Geisser *	0.025	3.221	0.008	0.678	.577
Fehler (Itemtyp)	Unkorrigiert	1.835	198	0.009		
	Greenh.-Geisser *	1.835	159.433	0.012		

Anmerkung: * $\varepsilon = .805$.

Tabelle F.103. Einfaktorielle ANCOVA mit dem Gruppenfaktor Status der aussagenden Person als UV, den Kontrollvariablen „Elektrodermale Labilität“, „Motivation, im *GAT* einen unschuldigen Eindruck zu hinterlassen“, „subjektiv eingeschätzte persönliche Erfolgswahrscheinlichkeit im *GAT*“, „Manipulationsmaßnahmen im *GAT*“, „Vorkenntnisse zur Glaubhaftigkeitsbeurteilung“ sowie „frühere Teilnahme an Untersuchungen zur Glaubhaftigkeitsbeurteilung“ als Kovariaten und der Höhe der numerischen *GAT*-Scores als AV (SCR-Quantifizierungsmethode B)

Q.d.V.	QS	df	MQ	F	p	η^2	η^2 *
Gruppe	630.045	2	315.023	15.452	.000	.249	.235
Fehler	1896.064	93	20.388				

Anmerkung: η^2 * = Effektstärke bei Nicht-Herauspartialisierung der Kovariaten (ANOVA, s. Tabelle F.83).

Tabelle F.104. Gesamt-*GAT*: Zweifaktorielle ANCOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Itemtyp als UVn, den Kontrollvariablen „Elektrodermale Labilität“, „Motivation, im *GAT* einen unschuldigen Eindruck zu hinterlassen“, „subjektiv eingeschätzte persönliche Erfolgswahrscheinlichkeit im *GAT*“, „Manipulationsmaßnahmen im *GAT*“, „Vorkenntnisse zur Glaubhaftigkeitsbeurteilung“ sowie „frühere Teilnahme an Untersuchungen zur Glaubhaftigkeitsbeurteilung“ als Kovariaten und den SCR-Amplituden als AV (SCR-Quantifizierungsmethode B)

Q.d.V.		QS	df	MQ	F	p	η^2	η^2 *
Gruppe		0.061	2	0.031	2.194	.117	.045	.043
Fehler (Gruppe)		1.296	93	0.014				
Itemtyp	Unkorrigiert	0.004	2	0.002	0.822	.441	.009	.275
	Greenh.-Geisser [a]	0.004	1.198	0.003	0.822	.387	.009	.275
Itemtyp × Gruppe	Unkorrigiert	0.058	4	0.015	6.707	.000	.126	.148
	Greenh.-Geisser [a]	0.058	2.395	0.024	6.707	.001	.126	.148
Fehler (Itemtyp)	Unkorrigiert	0.403	186	0.002				
	Greenh.-Geisser [a]	0.403	111.372	0.004				

Anmerkung: [a] $\varepsilon = .599$; η^2 * = Effektstärke bei Nicht-Herauspartialisierung der Kovariaten (ANOVA, s. Tabelle F.85).

Tabelle F.105. Diskriminanzanalyse mit dem numerischen *GAT*-Score als einzigem Prädiktor und der experimentellen Gruppenzugehörigkeit als Kriterium (Berechnung anhand einer Zufallsstichprobe von jeweils 17 *Täterinnen*, *Zeuginnen* und *falschen Zeuginnen*) (SCR-Quantifizierungsmethode B)

Faktor	Eigenwert λ	% der Varianz	Kumulierte %	Kanonische Korrelation
1	.450	100	100	.557

Signifikanzprüfung

Faktor(en)	Wilks Likelihood-Quotient Λ	χ^2	df	p
1	.690	17.833	2	.000

Prädiktor	Diskriminanzkoeffizient	Faktorladung
Numerischer Score	1.000	1.000

Tabelle F.106. Diskriminanzanalyse mit der intraindividuellen SCR-Magnituden-Differenz zwischen relevanten und irrelevanten Items als Prädiktor und der experimentellen Gruppenzugehörigkeit als Kriterium (Berechnung anhand einer Zufallsstichprobe von jeweils 17 *Täterinnen*, *Zeuginnen* und *falschen Zeuginnen*) (SCR-Quantifizierungsmethode B)

Faktor	Eigenwert λ	% der Varianz	Kumulierte %	Kanonische Korrelation
1	.225	100	100	.429

Signifikanzprüfung

Faktor	Wilks Likelihood-Quotient Λ	χ^2	df	p
1	.816	9.742	2	.008

Prädiktor	Diskriminanzkoeffizient	Faktorladung
SCR-Magnituden-Differenz	1.000	1.000

Tabelle F.107.: Naive Glaubhaftigkeitsbeurteilung: Zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Rater (Aussagenstichprobe I)

Q.d.V.		QS	df	MQ	F	p
Gruppe		40.533	2	20.267	1.203	.334
Fehler (Gruppe)		202.200	12	16.850		
Rater	Unkorrigiert	210.783	15	14.052	3.341	.000
	Greenh.-Geisser *	210.783	6.330	33.298	3.341	.005
Rater × Gruppe	Unkorrigiert	75.467	30	2.516	0.598	.952
	Greenh.-Geisser *	75.467	12.660	5.961	0.598	.845
Fehler (Rater)	Unkorrigiert	757.000	180	4.206		
	Greenh.-Geisser *	757.000	75.962	9.966		

Anmerkung: * ε = .422.

Tabelle F.108.: Naive Glaubhaftigkeitsbeurteilung: Zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Rater (Aussagenstichprobe II)

Q.d.V.		QS	df	MQ	F	p
Gruppe		33.908	2	16.954	1.094	.366
Fehler (Gruppe)		185.950	12	15.496		
Rater	Unkorrigiert	52.250	15	3.483	1.241	.245
	Greenh.-Geisser *	52.250	6.336	8.246	1.241	.294
Rater × Gruppe	Unkorrigiert	50.625	30	1.688	0.601	.950
	Greenh.-Geisser *	50.625	12.673	3.995	0.601	.842
Fehler (Rater)	Unkorrigiert	505.250	180	2.807		
	Greenh.-Geisser *	505.250	76.037	6.645		

Anmerkung: * ε = .422.

Tabelle F.109.: Naive Glaubhaftigkeitsbeurteilung: Zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Rater (Aussagenstichproben I und II zusammengefaßt)

Q.d.V.		QS	df	MQ	F	p
Gruppe		11.354	2	5.677	0.295	.747
Fehler (Gruppe)		518.738	27	19.213		
Rater	Unkorrigiert	155.800	15	10.387	2.977	.000
	Greenh.-Geisser *	155.800	8.953	17.402	2.977	.002
Rater × Gruppe	Unkorrigiert	82.713	30	2.757	0.790	.780
	Greenh.-Geisser *	82.713	17.906	4.619	0.790	.710
Fehler (Rater)	Unkorrigiert	1412.863	405	3.489		
	Greenh.-Geisser *	1412.863	241.726	5.845		

Anmerkung: * $\varepsilon = .597$.

Tabelle F.110. Einfaktorielle ANOVAs mit dem Gruppenfaktor Status der aussagenden Person als UV und der Höhe der durch den jeweiligen naiven Beurteiler eingeschätzten Glaubhaftigkeit als AV

ANOVA	Q.d.V.	QS	df	MQ	F	p
Rater 1	Gruppe	4.800	2	2.400	0.373	.696
	Fehler	77.200	12	6.433		
Rater 2	Gruppe	9.733	2	4.867	1.825	.203
	Fehler	32.000	12	2.667		
Rater 3	Gruppe	2.133	2	1.067	0.171	.845
	Fehler	74.800	12	6.233		
Rater 4	Gruppe	0.400	2	0.200	0.040	.961
	Fehler	60.000	12	5.000		
Rater 5	Gruppe	22.933	2	11.467	5.134	.024
	Fehler	26.800	12	2.233		
Rater 6	Gruppe	0.533	2	0.267	0.042	.959
	Fehler	76.400	12	6.367		
Rater 7	Gruppe	2.533	2	1.267	0.208	.815
	Fehler	73.200	12	6.100		
Rater 8	Gruppe	0.133	2	0.067	0.008	.992
	Fehler	97.600	12	8.133		
Rater 9	Gruppe	14.800	2	7.400	2.552	.119
	Fehler	34.800	12	2.900		
Rater 10	Gruppe	17.200	2	8.600	1.583	.245
	Fehler	65.200	12	5.433		
Rater 11	Gruppe	12.933	2	6.467	1.162	.346
	Fehler	66.800	12	5.567		

Tabelle F.110 (Fortsetzung). Einfaktorielle ANOVAs mit dem Gruppenfaktor Status der aussagenden Person als UV und der Höhe der durch den jeweiligen naiven Beurteiler eingeschätzten Glaubhaftigkeit als AV

ANOVA	Q.d.V.	QS	df	MQ	F	p
Rater 12	Gruppe	0.933	2	0.467	0.122	.886
	Fehler	46.000	12	3.833		
Rater 13	Gruppe	17.733	2	8.867	1.520	.258
	Fehler	70.000	12	5.833		
Rater 14	Gruppe	0.933	2	0.467	0.156	.858
	Fehler	36.000	12	3.000		
Rater 15	Gruppe	3.333	2	1.667	0.284	.758
	Fehler	70.400	12	5.867		
Rater 16	Gruppe	4.933	2	2.467	0.569	.581
	Fehler	52.000	12	4.333		
Rater 17	Gruppe	10.000	2	5.000	1.027	.387
	Fehler	58.400	12	4.867		
Rater 18	Gruppe	0.933	2	0.467	0.184	.834
	Fehler	30.400	12	2.533		
Rater 19	Gruppe	2.533	2	1.267	0.220	.806
	Fehler	69.200	12	5.767		
Rater 20	Gruppe	8.400	2	4.200	1.286	.312
	Fehler	39.200	12	3.267		
Rater 21	Gruppe	1.600	2	0.800	0.312	.738
	Fehler	30.800	12	2.567		
Rater 22	Gruppe	8.133	2	4.067	1.470	.269
	Fehler	33.200	12	2.767		
Rater 23	Gruppe	7.600	2	3.800	1.839	.201
	Fehler	24.800	12	2.067		
Rater 24	Gruppe	2.133	2	1.067	0.386	.688
	Fehler	33.200	12	2.767		
Rater 25	Gruppe	6.933	2	3.467	0.904	.431
	Fehler	46.000	12	3.833		
Rater 26	Gruppe	3.733	2	1.867	0.251	.782
	Fehler	89.200	12	7.433		
Rater 27	Gruppe	8.133	2	4.067	0.753	.492
	Fehler	64.800	12	5.400		
Rater 28	Gruppe	2.533	2	1.267	1.152	.349
	Fehler	13.200	12	1.100		

Tabelle F.110 (Fortsetzung). Einfaktorielle ANOVAs mit dem Gruppenfaktor Status der aussagenden Person als UV und der Höhe der durch den jeweiligen naiven Beurteiler eingeschätzten Glaubhaftigkeit als AV

ANOVA	Q.d.V.	QS	df	MQ	F	p
Rater 29	Gruppe	0.400	2	0.200	0.088	.916
	Fehler	27.200	12	2.267		
Rater 30	Gruppe	8.400	2	4.200	1.326	.302
	Fehler	38.000	12	3.167		
Rater 31	Gruppe	8.933	2	4.467	1.457	.271
	Fehler	36.800	12	3.067		
Rater 32	Gruppe	4.133	2	2.067	0.437	.656
	Fehler	56.800	12	4.733		

Tabelle F.111. Einfaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person als UV und der Wortzahl der Aussagen als AV

Q.d.V.	QS	df	MQ	F	p
Gruppe	1384130.725	2	692065.363	7.979	.001
Fehler	8586485.118	99	86732.173		
Gesamt	9970615.843	101			

Tabelle F.112. Scheffé-Tests für paarweise Vergleiche der experimentellen Gruppen bezüglich der Wortzahl der Aussagen

Gruppenvergleich			Mittlere Differenz	Standard-fehler	p	95%-Konfidenzintervall Untergrenze	Obergrenze
Täterinnen	vs.	*Zeuginnen*	-130.971	71.428	.191	-308.486	46.545
Täterinnen	vs.	*falsche Zeuginnen*	154.059	71.428	.103	-23.457	331.574
Zeuginnen	vs.	*falsche Zeuginnen*	285.029	71.428	.001	107.514	462.545

Tabelle F.113. Einfaktorielle ANOVA mit dem Faktor Interviewerin (A vs. B) als UV und der Wortzahl der Aussagen als AV

Q.d.V.	QS	df	MQ	F	p
Interviewerin	450871.031	1	450871.031	4.736	.032
Fehler	9519744.813	100	95197.448		
Gesamt	9970615.843	1010			

Tabelle F.114. Zweifaktorielle ANOVA mit den Faktoren Status der aussagenden Person und Interviewerin (A vs. B) als UVn und der Wortzahl der Aussagen als AV

Q.d.V.	QS	df	MQ	F	p
Gruppe	1279291.918	2	639645.959	7.563	.001
Interviewerin	380678.366	1	380678.366	4.501	.036
Gruppe × Interv.	88219.629	2	44109.814	0.522	.595
Fehler	8119252.727	96	84575.549		
Gesamt	57976896.000	102			

Tabelle F.115. Einfaktorielle MANCOVA mit dem Gruppenfaktor Status der aussagenden Person (*Täterinnen*, *Zeuginnen*, *falsche Zeuginnen*) als UV, der Variable „Interviewerin“ (A vs. B) als Kovariate und den 18 Glaubhaftigkeitskriterien als AVn

Teststatistik	Wert	F	Hypothese df	Fehler df	p	η^2	η^2 *
Pillais Spurkriterium PS	0.884	3.606	36	164	.000	.442	.436
Wilks Likelihood-Quotient Λ	0.282	3.981	36	162	.000	.469	.462
Hotellings Spurkriterium T	1.965	4.367	36	160	.000	.496	.486
Roys größter Eigenwert	1.598	7.280	18	82	.000	.615	.604

Anmerkung: η^2 * = Effektstärke bei Nicht-Herauspartialisierung der Kovariaten (MANOVA, s. Tabelle F.29).

Tabelle F.116. Einfaktorielle ANCOVAs mit dem Gruppenfaktor Status der aussagenden Person als UV, der Variable „Interviewerin“ (A vs. B) als Kovariate und dem Ausprägungsgrad des jeweiligen Glaubhaftigkeitskriteriums als AV

Kriterium	Q.d.V.	QS	df	MQ	F	p	η^2	η^2 *
1. *Konsistenz*	Gruppe	0.003	2	0.001	0.044	.957	.001	.001
	Fehler	2.884	98	0.029				
2. *Unordnung*	Gruppe	0.963	2	0.481	2.046	.135	.040	.041
	Fehler	23.053	98	0.235				
3. *Details*	Gruppe	13.367	2	6.684	14.187	.000	.225	.226
	Fehler	46.168	98	0.471				
4. *Verknüpfungen*	Gruppe	5.362	2	2.681	6.904	.002	.123	.122
	Fehler	38.053	98	0.388				
5. *Interaktionen*	Gruppe	0.726	2	0.363	1.894	.156	.037	.037
	Fehler	18.790	98	0.192				
6. *Gespräche*	Gruppe	16.466	2	8.233	12.881	.000	.208	.193
	Fehler	62.640	98	0.639				

Tabelle F.116 (Fortsetzung). Einfaktorielle ANCOVAs mit dem Gruppenfaktor Status der aussagenden Person als UV, der Variable „Interviewerin“ (A vs. B) als Kovariate und dem Ausprägungsgrad des jeweiligen Glaubhaftigkeitskriteriums als AV

	Kriterium	Q.d.V.	QS	df	MQ	F	p	η^2	η^2 *
7.	*Komplikationen*	Gruppe	0.040	2	0.020	0.644	.528	.013	.013
		Fehler	3.044	98	0.031				
8.	*Ausgefallenes*	Gruppe	2.964	2	1.482	18.735	.000	.277	.268
		Fehler	7.753	98	0.079				
9.	*Nebensächliches*	Gruppe	7.538	2	3.769	12.900	.000	.208	.209
		Fehler	28.633	98	0.292				
10.	*Unverstandenes*	Gruppe	0.027	2	0.014	1.647	.198	.033	.031
		Fehler	0.816	98	0.008				
11.	*Indirektes*	Gruppe	0.007	2	0.003	1.066	.348	.021	.020
		Fehler	0.315	98	0.003				
12.	*Eigenseelisches*	Gruppe	0.590	2	0.295	0.831	.439	.017	.017
		Fehler	34.800	98	0.355				
13.	*Fremdseelisches*	Gruppe	0.535	2	0.267	0.847	.432	.017	.017
		Fehler	30.905	98	0.315				
14.	*Verbesserungen*	Gruppe	2.600	2	1.300	4.445	.014	.083	.084
		Fehler	28.666	98	0.293				
15.	*Erinnerungslücken*	Gruppe	0.566	2	0.283	3.417	.037	.065	.067
		Fehler	8.115	98	0.083				
16.	*Selbsteinwände*	Gruppe	0.002	2	0.001	0.131	.877	.003	.003
		Fehler	0.814	98	0.008				
17.	*Eigenbelastung*	Gruppe	0.009	2	0.004	0.492	.613	.010	.010
		Fehler	0.863	98	0.009				
18.	*Fremdentlastung*	Gruppe	0.713	2	0.356	2.589	.080	.050	.050
		Fehler	13.491	98	0.138				

Anmerkung: η^2 * = Effektstärke bei Nicht-Herauspartialisierung der Kovariaten (ANOVAs, s. Tabelle F.33).

Tabelle F.117. Einfaktorielle ANCOVA mit dem Gruppenfaktor Status der aussagenden Person als UV, der Variable „Interviewerin“ (A vs. B) als Kovariate und der Höhe des über alle 18 Glaubhaftigkeitskriterien aufsummierten Gesamtscores als AV

Q.d.V.	QS	df	MQ	F	p	η^2	η^2 *
Gruppe	121.502	2	60.751	5.449	.006	.100	.100
Fehler	1092.601	98	11.149				

Anmerkung: η^2 * = Effektstärke bei Nicht-Herauspartialisierung der Kovariaten (ANOVA, s. Tabelle F.37).

Tabelle F.118. Gesamt-*GAT*: Dreifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und den Meßwiederholungsfaktoren Fragenposition und Itemtyp als UVn und den SCR-Amplituden als AV (SCR-Quantifizierungsmethode A)

Q.d.V.		QS	df	MQ	F	p
Gruppe		0.806	2	0.403	2.845	.063
Fehler (Gruppe)		14.023	99	0.142		
Fragenposition	Unkorrigiert	0.552	9	0.061	10.020	.000
	Greenh.-Geisser *	0.552	6.348	0.087	10.020	.000
Fragenp. × Gruppe	Unkorrigiert	0.138	18	0.008	1.253	.212
	Greenh.-Geisser *	0.138	12.695	0.011	1.253	.239
Fehler (Fragenp.)	Unkorrigiert	5.456	891	0.006		
	Greenh.-Geisser *	5.456	628.412	0.009		
Itemtyp	Unkorrigiert	2.272	2	1.136	38.949	.000
	Greenh.-Geisser *	2.272	1.169	1.943	38.949	.000
Itemtyp × Gruppe	Unkorrigiert	0.902	4	0.226	7.733	.000
	Greenh.-Geisser *	0.902	2.339	0.386	7.733	.000
Fehler (Itemtyp)	Unkorrigiert	5.774	198	0.029		
	Greenh.-Geisser *	5.774	115.772	0.050		
Fragenp. × Itemtyp	Unkorrigiert	0.245	18	0.014	2.975	.000
	Greenh.-Geisser *	0.245	10.444	0.023	2.975	.001
Fr. × Itemtyp × Gr.	Unkorrigiert	0.262	36	0.007	1.593	.015
	Greenh.-Geisser *	0.262	20.889	0.013	1.593	.044
Fehler (Fr. × It.)	Unkorrigiert	8.155	1782	0.005		
	Greenh.-Geisser *	8.155	1033.995	0.008		

Anmerkung: * ε (Fragenposition) = .705; ε (Itemtyp) = .585; ε (Fragenposition × Itemtyp) = .580.

Tabelle F.119. Gesamt-*GAT*: Dreifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und den Meßwiederholungsfaktoren Fragenposition und Itemtyp als UVn und den SCR-Amplituden als AV (SCR-Quantifizierungsmethode B)

Q.d.V.		QS	df	MQ	F	p
Gruppe		0.822	2	0.411	2.230	.113
Fehler (Gruppe)		18.239	99	0.184		
Fragenposition	Unkorrigiert	0.441	9	0.049	6.300	.000
	Greenh.-Geisser *	0.441	6.320	0.070	6.300	.000
Fragenp. × Gruppe	Unkorrigiert	0.149	18	0.008	1.067	.382
	Greenh.-Geisser *	0.149	12.640	0.012	1.067	.386
Fehler (Fragenp.)	Unkorrigiert	6.931	891	0.008		
	Greenh.-Geisser *	6.931	625.670	0.011		
Itemtyp	Unkorrigiert	1.675	2	0.838	37.486	.000
	Greenh.-Geisser *	1.675	1.183	1.415	37.486	.000
Itemtyp × Gruppe	Unkorrigiert	0.766	4	0.191	8.567	.000
	Greenh.-Geisser *	0.766	2.367	0.323	8.567	.000
Fehler (Itemtyp)	Unkorrigiert	4.424	198	0.022		
	Greenh.-Geisser *	4.424	117.158	0.038		
Fragenp. × Itemtyp	Unkorrigiert	0.228	18	0.013	2.409	.001
	Greenh.-Geisser *	0.228	11.474	0.020	2.409	.005
Fr. × Itemtyp × Gr.	Unkorrigiert	0.302	36	0.008	1.595	.014
	Greenh.-Geisser *	0.302	22.948	0.013	1.595	.037
Fehler (Fr. × It.)	Unkorrigiert	9.385	1782	0.005		
	Greenh.-Geisser *	9.385	1135.938	0.008		

Anmerkung: * ε (Fragenposition) = .702; ε (Itemtyp) = .592; ε (Fragenposition × Itemtyp) = .637.

Tabelle F.120. Diskriminanzanalyse mit dem numerischen *GAT*-Score als einzigem Prädiktor und dem tatbezogenen Kenntnisstand (Tatwissen vorhanden vs. nicht vorhanden) als Kriterium (Berechnung anhand einer Zufallsstichprobe von 34 Personen mit Tatwissen [17 *Täterinnen*, 17 *Zeuginnen*] und 17 Personen ohne Tatwissen [*falsche Zeuginnen*]) (SCR-Quantifizierungsmethode A)

Faktor	Eigenwert λ	% der Varianz	Kumulierte %	Kanonische Korrelation
1	.417	100	100	.543

Signifikanzprüfung

Faktor	Wilks Likelihood-Quotient Λ	χ^2	df	p
1	.706	16.910	1	.000

Prädiktor	Diskriminanzkoeffizient	Faktorladung
Numerischer Score	1.000	1.000

Tabelle F.121. Diskriminanzanalyse mit dem numerischen *GAT*-Score als einzigem Prädiktor und dem tatbezogenen Kenntnisstand (Tatwissen vorhanden vs. nicht vorhanden) als Kriterium (Berechnung anhand einer Zufallsstichprobe von 34 Personen mit Tatwissen [17 *Täterinnen*, 17 *Zeuginnen*] und 17 Personen ohne Tatwissen [*falsche Zeuginnen*]) (SCR-Quantifizierungsmethode B)

Faktor	Eigenwert λ	% der Varianz	Kumulierte %	Kanonische Korrelation
1	.405	100	100	.537

Signifikanzprüfung

Faktor	Wilks Likelihood-Quotient Λ	χ^2	df	p
1	.712	16.499	1	.000

Prädiktor	Diskriminanzkoeffizient	Faktorladung
Numerischer Score	1.000	1.000

Tabelle F.122. Diskriminanzanalyse mit der intraindividuellen SCR-Magnituden-Differenz zwischen relevanten und irrelevanten Items als einzigem Prädiktor und dem tatbezogenen Kenntnisstand (Tatwissen vorhanden vs. nicht vorhanden) als Kriterium (Berechnung anhand einer Zufallsstichprobe von 34 Personen mit Tatwissen [17 *Täterinnen*, 17 *Zeuginnen*] und 17 Personen ohne Tatwissen [*falsche Zeuginnen*]) (SCR-Quantifizierungsmethode A)

Faktor	Eigenwert λ	% der Varianz	Kumulierte %	Kanonische Korrelation
1	.216	100	100	.422

Signifikanzprüfung

Faktor	Wilks Likelihood-Quotient Λ	χ^2	df	p
1	.822	9.499	1	.002

Prädiktor	Diskriminanzkoeffizient	Faktorladung
Numerischer Score	1.000	1.000

Tabelle F.123. Diskriminanzanalyse mit der intraindividuellen SCR-Magnituden-Differenz zwischen relevanten und irrelevanten Items als einzigem Prädiktor und dem tatbezogenen Kenntnisstand (Tatwissen vorhanden vs. nicht vorhanden) als Kriterium (Berechnung anhand einer Zufallsstichprobe von 34 Personen mit Tatwissen [17 *Täterinnen*, 17 *Zeuginnen*] und 17 Personen ohne Tatwissen [*falsche Zeuginnen*]) (SCR-Quantifizierungsmethode B)

Faktor	Eigenwert λ	% der Varianz	Kumulierte %	Kanonische Korrelation
1	.208	100	100	.415

Signifikanzprüfung

Faktor	Wilks Likelihood-Quotient Λ	χ^2	df	p
1	.828	9.181	1	.002

Prädiktor	Diskriminanzkoeffizient	Faktorladung
Numerischer Score	1.000	1.000

Tabelle F.124.: Naive Glaubhaftigkeitsbeurteilung: Zweifaktorielle ANOVA mit dem Gruppenfaktor Status der aussagenden Person und dem Meßwiederholungsfaktor Rater (ergänzende Datenerhebung)

Q.d.V.	QS	df	MQ	F	p
Gruppe	25.657	2	12.828	3.386	.038
Fehler (Gruppe)	375.132	99	3.789		
Rater	206.005	1	206.005	70.337	.000
Rater × Gruppe	23.539	2	11.770	4.019	.021
Fehler (Rater)	289.956	99	2.929		

Tabelle F.125. Scheffé-Tests für paarweise Vergleiche der experimentellen Gruppen bezüglich der durch beide naiven Beurteiler eingeschätzten Glaubhaftigkeit (ergänzende Datenerhebung)

Gruppenvergleich			Mittlere Differenz	Standard-fehler	p	95%-Konfidenzintervall Untergrenze	Obergrenze
Täterinnen	vs.	*Zeuginnen*	-0.471	0.334	.374	-1.300	0.359
Täterinnen	vs.	*falsche Zeuginnen*	0.397	0.334	.495	-0.433	1.227
Zeuginnen	vs.	*falsche Zeuginnen*	0.868	0.334	.038	0.038	1.697

Tabelle F.126. Einfaktorielle ANOVAs mit dem Gruppenfaktor Status der aussagenden Person als UV und der Höhe der durch den jeweiligen naiven Beurteiler eingeschätzten Glaubhaftigkeit als AV (ergänzende Datenerhebung)

ANOVA	Q.d.V.	QS	df	MQ	F	p
Rater X	Gruppe	34.824	2	17.412	7.065	.001
	Fehler	243.971	99	2.464		
Rater Y	Gruppe	14.373	2	7.186	1.689	.190
	Fehler	421.118	99	4.254		

Tabelle F.127. Scheffé-Tests für paarweise Vergleiche der experimentellen Gruppen bezüglich der naiven Glaubhaftigkeitsbeurteilung durch Rater X (ergänzende Datenerhebung)

Gruppenvergleich			Mittlere Differenz	Standard-fehler	p	95%-Konfidenzintervall Untergrenze	Obergrenze
Täterinnen	vs.	*Zeuginnen*	-1.294	0.381	.004	-2.240	-0.348
Täterinnen	vs.	*falsche Zeuginnen*	-0.118	0.381	.953	-1.064	0.829
Zeuginnen	vs.	*falsche Zeuginnen*	1.177	0.381	.010	0.230	2.123

Tabelle F.128. Treffsicherheit der naiven Glaubhaftigkeitsbeurteilung (ergänzende Datenerhebung)

	Status der aussagenden Person													
	Täterinnen (unglaubhaft)				*Zeuginnen* (glaubhaft)				*falsche Zeuginnen* (unglaubhaft)					
	Diagnose				Diagnose				Diagnose					
	unglaubh.		glaubhaft		Unglaubh.		glaubhaft		unglaubh.		glaubhaft		Gesamt-trefferquote	
Rater	n	%	n	%	n	%	n	%	n	%	n	%	%	%*
X	9	26.5	25	73.5	2	5.9	32	94.1	9	26.5	25	73.5	49.0	60.3
Y	18	52.9	16	47.1	20	58.8	14	41.2	21	61.8	13	38.2	52.0	49.3
gesamt		39.7		60.3		32.3		67.7		44.1		55.9	50.5	54.8

Anmerkung: grau unterlegte Felder = Trefferquoten; * Gesamttrefferquote bei Korrektur der Basisraten von glaubhaften und unglaubhaften Aussagen.

Tabelle F.129. Diskriminanzanalyse mit dem durchschnittlichen naiven Glaubhaftigkeitsrating auf der achtstufigen Skala als Prädiktor und der tatsächlichen Glaubhaftigkeit der experimentellen Aussagen (glaubhaft vs. unglaubhaft) als Kriterium (Berechnung anhand von Aussagenstichprobe II)

Faktor	Eigenwert λ	% der Varianz	Kumulierte %	Kanonische Korrelation
1	.107	100	100	.311

Signifikanzprüfung

Faktor	Wilks Likelihood-Quotient Λ	χ^2	df	p
1	.903	1.271	1	.260

Prädiktor	Diskriminanzkoeffizient	Faktorladung
durchschnittl. Glaubh.-rating	1.000	1.000

Abstract

In der forensischen Glaubhaftigkeitsbeurteilung dominieren der inhaltsorientierte und der psychophysiologische Ansatz. Beim inhaltsorientierten Ansatz wird primär analysiert, ob eine Aussage inhaltliche Merkmale aufweist, die für erlebnisbasierende Schilderungen typisch sein sollen. Der Hauptanwendungsbereich ist die Begutachtung vermeintlicher Zeugen. Mit der psychophysiologischen Glaubhaftigkeitsbeurteilung werden in erster Linie die Einlassungen von Beschuldigten begutachtet. Während einer standardisierten Befragung oder Reizdarbietung werden physiologische Veränderungen beim Beschuldigten gemessen, von denen Rückschlüsse auf die Glaubhaftigkeit der Täterschaftsabstreitung gezogen werden.

In der vorliegenden Untersuchung wurde die Validität beider Ansätze verglichen. Einhundertundzwei Probandinnen nahmen als Täterinnen (Schuldige), Zeuginnen (Unschuldige mit Tatwissen) oder falsche Zeuginnen (Unschuldige ohne Tatwissen) an einer experimentellen Verbrechenssimulation teil. Die Täterinnen stritten anschließend die Täterschaft wahrheitswidrig ab und machten eine wahrheitswidrige „Zeugenaussage“. Die Zeuginnen stritten die Täterschaft wahrheitsgemäß ab und machten eine wahrheitsgemäße Zeugenaussage. Die falschen Zeuginnen stritten die Täterschaft wahrheitsgemäß ab und machten eine wahrheitswidrige „Zeugenaussage“. Die Glaubhaftigkeit der vermeintlichen Zeugenaussage wurde jeweils inhaltsorientiert, mit Hilfe der *Kriterienorientierten Inhaltsanalyse* beurteilt. Die Glaubhaftigkeit der Täterschaftsabstreitung beurteilte man jeweils psychophysiologisch, mit dem *Guilty Actions Test*, einem speziellen Verfahren, das zwischen Schuldigen, Unschuldigen mit Tatwissen und Unschuldigen ohne Tatwissen differenzieren soll. Neben den beiden forensischen Methoden wurde als Kontrollbedingung auch noch eine naive, d.h. rein intuitive Glaubhaftigkeitsbeurteilung vorgenommen.

Die diagnostische Differenzierungsfähigkeit der beiden forensischen Methoden war höher als die der naiven Glaubhaftigkeitsbeurteilung, welche nur dem Zufallsniveau entsprach. Erwartungskonträr eignete sich der *Guilty Actions Test* nur zur Differenzierung von Personen mit und solchen ohne Tatwissen, nicht jedoch zur Differenzierung zwischen Schuldigen und Unschuldigen mit Tatwissen. Als Instrument zur Überprüfung von Tatwissen erzielte der *Guilty Actions Test* eine Treffsicherheit, die höher war als die Treffsicherheit der inhaltsorientierten Beurteilung bezüglich der Identifizierung glaubhafter vs. unglaubhafter Aussagen. Die Treffsicherheit der inhaltsorientierten Beurteilung lag zudem nur bei den glaubhaften, nicht jedoch bei den unglaubhaften Aussagen über dem Zufallsniveau. Die Ergebnisse wurden insbesondere im Hinblick auf methodische Aspekte der vorliegenden Untersuchung diskutiert.