Forensische Psychophysiologie:

Ein Beitrag zu den psychologischen und physiologischen Grundlagen neuerer Ansätze der „Lügendetektion“

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

vorgelegt dem Fachbereich 12 Sozialwissenschaften
der Johannes Gutenberg-Universität
Mainz

von
Hans-Georg Rill
aus Hermannstadt

Mainz
2001

Referent:

Korreferent:

Tag des Prüfungskolloquiums: 13. Juli 2001

Ich danke allen, die an der Entstehung dieser Arbeit mitgewirkt haben:

Herrn Professor Dr. [redacted] für die intensive Betreuung der Dissertation und die sowohl fachliche als auch persönliche Unterstützung während unserer gesamten Zusammenarbeit,

Herrn Professor Dr. [redacted], der mir durch die Anstellung als wissenschaftlicher Mitarbeiter die erforderlichen Rahmenbedingungen zur Promotion geschaffen hat,

Herrn Hochschuldozent Dr. [redacted] für die zusätzlichen Hinweise zur Interpretation der elektrodermalen und kardiovaskulären Daten,

Frau [redacted] M. A. für die vielfältigen Hilfestellungen, Anregungen und Korrekturvorschläge während der gesamten Untersuchung und bei der Erstellung des Manuskripts,

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

Herrn [redacted] und Herrn [redacted], die an der Vorbereitung und Durchführung der Untersuchung sowie an der Datenauswertung beteiligt waren,

Herrn Referendar jur. [redacted] für die rechtswissenschaftliche Beratung,

Frau [redacted], Frau [redacted], Frau [redacted] und Frau [redacted] für die Durchsicht von Teilen des Manuskripts,

den Mitarbeitern der Abteilung Allgemeine Experimentelle Psychologie des Psychologischen Instituts und den Mitgliedern der Interdisziplinären Forschungsgruppe Forensische Psychophysiologie,

und nicht zuletzt allen 122 Teilnehmern des Experiments.

Diese Arbeit ist meinen Eltern [redacted] gewidmet.

Inhaltsverzeichnis

> The effectiveness of lie detection procedures is limited by a lack of knowledge of what psychological principles are involved in successful lie detection. (Davis, 1961, S. 160)

1. Einleitung

Die „**psychophysiologische Aussagebeurteilung**“ („forensische Psychophysiologie“, „Lügendetektion“) zielt darauf ab, anhand der körperlichen Reaktionen auf bestimmte Stimuli diagnostische Schlußfolgerungen über die Glaubwürdigkeit bzw. Tatbeteiligung einer Person an einem kriminellen Vergehen zu treffen. Zu diesem Zweck werden in der Regel verschiedene verbale Reize (Fragen oder Items) dargeboten und mehrere physiologische Variablen erfaßt. Dabei handelt es sich überwiegend um Indikatoren für die Aktivität des peripheren vegetativen Nervensystems, wie etwa elektrodermale, kardiovaskuläre und respiratorische Größen. Neben den erhobenen körperlichen Parametern ist die Art der Stimulation, d. h. die Befragungstechnik, von entscheidender Bedeutung. Die gängigen Verfahren bestehen aus einer Kombination von sog. relevanten, direkt auf die kritische Tat (z. B. Verbrechen) bezogenen Stimuli und entsprechenden Vergleichsreizen. Letztere weisen keinen unmittelbaren Tatbezug auf, sie thematisieren aber ähnliche Sachverhalte. Gemäß den zentralen Annahmen sollen sich Unterschiede in der kognitiven Verarbeitung und emotionalen Bewertung der Reize in den physiologischen Variablen manifestieren. Die individualdiagnostischen Entscheidungen über den Wahrheitsgehalt von Aussagen basieren auf dem intraindividuellen Vergleich der Reaktionen der betreffenden Person auf die einzelnen Fragen- bzw. Itemtypen.

Die wichtigsten **Befragungstechniken** sind der Kontrollfragentest (KFT) und der Tatwissentest (TWT). Beide haben aus verschiedenen Gründen Kritik auf sich gezogen. Infolgedessen wurden in den vergangenen Jahren die Anstrengungen intensiviert, die Verfahren zu modifizieren und zu verbessern. Zwei dieser neueren Entwicklungen sind der „Directed Lie Test“ (DLT) und der „Guilty Actions Test“ (GAT). Im Gegensatz zum KFT bietet der DLT eine erhöhte Standardisierbarkeit. Und in Relation zum TWT soll der GAT eine bessere Differenzierung zwischen Schuldigen und Unschuldigen mit Tatwissen ermöglichen.

Trotz der methodischen Fortschritte bleibt eine Reihe von **Problemen** ungelöst. Neben der diagnostischen Güte sind auch die theoretischen Grundlagen der Befragungstechniken bislang weitgehend unklar. Dies hängt unter anderem damit zusammen, daß sich die psychophysiologische Aussageforschung überwiegend mit der kriterienbezogenen Validität, d. h. Treffsicherheit der Verfahren, beschäftigt. Indes werden nur selten die Einflüsse der Stimuli auf die körperlichen Reaktionen und die entsprechenden psychophy-

siologischen Prozesse direkt analysiert. Insbesondere experimentelle Studien unter kontrollierten Bedingungen eignen sich jedoch weniger zur Abschätzung von Trefferquoten, als vielmehr zur Grundlagenforschung. Darüber hinaus sind die Standardisierung und die Objektivität der Testverfahren sowohl in Feld- als auch in Laboruntersuchungen oft nicht gewährleistet, so daß unterschiedliche Studien kaum vergleichbar sind und ihre Ergebnisse keine eindeutigen Schlüsse zulassen.

Ausgehend von dieser Kritik verfolgt die **vorliegende Arbeit** ausdrücklich *keine* primär individualdiagnostische Zielsetzung. Statt dessen sollen im Rahmen eines sog. „Scheinverbrechen-Experiments“ mit standardisierter Durchführung der psychophysiologischen Aussagebeurteilung und objektiver Auswertung der Daten quantitative Reaktionsunterschiede zwischen den Fragen- bzw. Itemtypen des DLT bzw. GAT nachgewiesen und untersucht werden.

Eine **wesentliche Problemstellung** besteht darin, inwieweit relevante Fragen bzw. Items, die sich auf eine verübte Tat beziehen und wahrheitswidrig beantwortet werden, stärkere Reaktionen evozieren als vergleichbare Stimuli, die ein nicht begangenes Delikt thematisieren und somit wahrheitsgemäß verneint werden. Anders als bei den üblichen Validitätsstudien der forensischen Psychophysiologie fußt dieser Vergleich nicht auf einer Gegenüberstellung „schuldiger“ vs. „unschuldiger“ Probanden (Pbn), sondern auf einer intraindividuellen Variation der Täterschaft. Folglich beinhaltet die eigene Untersuchung zwei Scheinverbrechen, die äquivalent gestaltet sind. Jeder Teilnehmer muß nur eines davon durchführen und ist somit hinsichtlich dieses Delikts „schuldig“ und hinsichtlich des anderen „unschuldig“. Während der Tatbegehung werden alle Personen auch mit den kritischen Details der jeweils nicht verübten Tat konfrontiert. Anschließend wird jeder Proband (Pb) entweder mit dem DLT oder dem GAT bezüglich beider Scheinverbrechen getestet, d. h., die relevanten Fragen bzw. Items beziehen sich auf beide Delikte. Beim DLT variiert man zusätzlich den **Wahrheitsgehalt der Antworten auf die Kontrollfragen**, um zu überprüfen, ob dies einen Effekt auf die Reaktionsstärke hat. Darüber hinaus beinhalten die Befragungstechniken irrelevante Stimuli, die keinen direkten Zusammenhang zu den Scheinverbrechen aufweisen.

Als **abhängige Variablen** werden neben den physiologischen Größen (Hautleitfähigkeitsreaktionen und phasische Herzschlagfrequenz) auch Selbsteinstufungen der körperlichen Reaktionsstärke und Einschätzungen der Reizrelevanz erhoben, um zu untersuchen, inwiefern sich die Reaktions- und Bedeutsamkeitsunterschiede zwischen den Fragen- bzw. Itemtypen im subjektiven Erleben widerspiegeln. Durch eine zeitliche **Verzögerung der Antworten** auf die Fragen bzw. Items will man außerdem die körperlichen Reaktionen auf die Darbietung der Reize von den physiologischen Begleit-

erscheinungen der Antworten trennen. Dies soll eine genauere Analyse der Prozesse ermöglichen, die eventuell am Zustandekommen der Reaktionsunterschiede beteiligt sind (z. B. Reizwahrnehmung, Informationsverarbeitung, Reaktionsvorbereitung und Antwortgabe). Ferner werden erstmals für den DLT und den GAT die **Einflüsse der elektrodermalen Labilität** auf die Reaktionsunterschiede zwischen den Fragen- bzw. Itemtypen untersucht, da frühere Studien darauf hindeuteten, daß dieses psychophysiologische Personenmerkmal Effekte auf die Reaktionen und Trefferquoten der konventionellen Verfahren (v. a. Kontrollfragentest) haben kann.

Gewisse **Grundkenntnisse der forensischen Psychophysiologie** sind für die Nachvollziehbarkeit der vorliegenden Untersuchung und ihrer Problemstellungen unabdingbar. Einerseits setzt das Verständnis der neueren Verfahren voraus, daß die konventionellen Befragungstechniken und deren Kritikpunkte bereits bekannt sind. Andererseits nehmen Zielsetzung, Methodik und Diskussion der eigenen Arbeit wiederholt Bezug auf die bisherigen Befunde und Theorien zur psychophysiologischen Aussagebeurteilung.

Auch angesichts der Tatsache, daß dieser Forschungs- und Anwendungsbereich zumindest in Deutschland erst im vergangenen Jahrzehnt zunehmend an Aktualität gewonnen hat, bietet das **zweite Kapitel** einen ausführlichen **Exkurs** zur Einführung in die Thematik und Problematik der „Lügendetektion“. Nach einer Erörterung der gesellschaftspolitischen Relevanz werden die wichtigsten Befragungstechniken dargestellt und ihre Vor- und Nachteile diskutiert. Neben dem Tatwissentest nimmt der Kontrollfragentest dabei einen zentralen Stellenwert ein, zumal keine andere Befragungstechnik derart verbreitet und gleichzeitig so umstritten ist. Jene Leser, die bereits mit den konventionellen Verfahren und deren Problemen vertaut sind, können die entsprechenden Abschnitte überspringen und sich direkt den neueren Ansätzen zuwenden, die jeweils im Anschluß daran beschrieben werden. Der Darstellung der Befragungstechniken folgt eine Zusammenfassung der bisherigen Validitätsforschung zur psychophysiologischen Aussagebeurteilung einschließlich einer kritischen Analyse ihrer Methoden. In diesem Kontext sollen insbesondere auch die Möglichkeiten der laborexperimentellen Forschung thematisiert werden. Ein weiterer Abschnitt ist den theoretischen Entwürfen gewidmet, die zur Erklärung der körperlichen Reaktionsunterschiede herangezogen werden. Dabei spielen emotionale und motivationale Konzepte ebenso eine bedeutende Rolle wie kognitive Prozesse. Allerdings muß man feststellen, daß keine der bisher formulierten Theorien in der Lage ist, die Gesamtheit der empirischen Befunde auch nur annähernd umfassend abzudecken. Insofern zielt die eigene Studie nicht darauf ab, einzelne Ansätze gegenüberzustellen und zu überprüfen. Der theoretische Hintergrund kann allenfalls einen Bezugsrahmen für die Interpretation der erzielten Ergebnisse bieten.

Leser, die über hinreichendes Vorwissen verfügen, können den Exkurs gänzlich auslassen und direkt mit dem **Kapitel 3** beginnen, das die **Zielsetzungen** der eigenen experimentellen Untersuchung ausführlich darlegt. Wie bereits erwähnt, stehen dabei weder die Abschätzung der Treffsicherheit der Verfahren noch eine Überprüfung einzelner Theorien im Vordergrund. Statt dessen soll eine besondere **Methodik** (**Kapitel 4**) neue Erkenntnisse erbringen, inwiefern die körperlichen Reaktionen und Selbsteinschätzungen der Pbn auf die Stimuli des DLT bzw. GAT von der intraindividuellen Variation der Täterschaft, vom Wahrheitsgehalt der Antworten auf die Kontrollfragen des DLT und von der elektrodermalen Labilität der Personen abhängig sind. Darüber hinaus liegt ein besonderer Schwerpunkt auf der standardisierten Durchführung und objektiven Auswertung der psychophysiologischen Aussagebeurteilung. Die Resultate für die erfaßten elektrodermalen, kardiovaskulären und subjektiven Daten werden im **Ergebnisteil** (**Kapitel 5**) getrennt referiert und im **Diskussionsteil** (**Kapitel 6**) interpretiert.

2. Exkurs: Einführung in die Methodik und Problematik der psychophysiologischen Aussagebeurteilung

2.1 Terminologische und definitorische Vorbemerkungen

William Moulton Marston, einer der Pioniere der psychophysiologischen Aussagebeurteilung, veröffentlichte 1938 sein Buch „The Lie Detector Test“ (zur historischen Entwicklung siehe Ben-Shakhar & Furedy, 1990, S. 1ff.; Berning, 1992, S. 3ff.; Lykken, 1998, S. 21ff.; Steller, 1987, S. 16ff.). Sein Werk trug entscheidend zur Verbreitung des Begriffs „**Lügendetektion**“ und dessen Mythos bei. Über Jahrzehnte hinweg etablierte sich insbesondere bei Laien aber auch bei Experten des Forschungsgegenstandes die Annahme, es gäbe charakteristische körperliche Reaktionen oder Symptome der Täuschung (vgl. „deception responses“, „symptoms of deception“, Reid & Inbau, 1977, S. 61ff.). Diese müßte man nur apparativ erfassen, um daraus den Wahrheitsgehalt einer Aussage bestimmen zu können.

Inzwischen herrscht sowohl unter Befürwortern (z. B. Podlesny & Raskin, 1977, S. 784f.; Undeutsch, 1983, S. 401f.) als auch unter kritischeren Vertretern (z. B. Ben-Shakhar & Furedy, 1990, S. 3; Lykken, 1998, S. 63ff.) der psychophysiologischen Aussagebeurteilung weitgehend Einigkeit, daß keine empirischen Evidenzen auf die Existenz einer „**spezifischen Lügenreaktion**“ („specific lie response“) hindeuten. Ein qualitativ distinktes körperliches Reaktionsmuster, das als inter- oder intraindividuell konsistentes Korrelat von Täuschungen auftritt, ist bisher nicht bekannt und möglicherweise auch gar nicht existent. Insofern kann zumindest nach dem jetzigen Stand der Erkenntnisse von „Lügendetektion“ im engeren Sinne keine Rede sein. Deshalb sollte dieser umgangssprachliche Begriff vermieden oder nur in Anführungszeichen gebraucht werden.

Alle Bezeichnungen, die ein Entdecken von Lügen oder Täuschungen nahelegen, sind terminologisch illegitim. Dazu gehört auch der Ausdruck „**Psychophysiological Detection of Deception**“, der sich in der angloamerikanischen Fachliteratur eingebürgert hat (vgl. Yankee, 1995, S. 63). Matte (1996, S. 4) schlägt „Psychophysiological Veracity Examination“ als Alternative vor. Ein deutsches Pendant dazu lautet „Psychophysiologische Glaubhaftigkeitsbeurteilung“ (Greuel et al., 1998, S. 235). Andere Synonyme sind ebenfalls problematisch. Die insbesondere von Juristen verwendeten Begriffe „**Polygraphie**“, „Polygraphentest“ bzw. „polygraphische Untersuchung“ (z. B. Achenbach, 1984, S. 350; Eisenberg, 1993, S. 121ff.; Klimke, 1981, S. 433) implizieren allesamt eine Überbewertung der entsprechenden Meßgeräte zur Erfassung körperlicher Veränderungen (Polygraphen, d. h. Mehrkanalschreiber). Statt dessen sind jedoch die

Art der dargebotenen Stimuli und deren Darbietungsform – also die Befragungstechniken – von entscheidender Bedeutung. Die von Undeutsch (1979, S. 228, 1983, S. 390, 1997, S. 303) favorisierten Ausdrücke „**Psychophysiologische Täterschaftsermittlung bzw. -diagnostik**“ umfassen nicht die ebenfalls praktizierbare Glaubwürdigkeitsbegutachtung von Zeugenaussagen (Steller, 1987, S. 5).

Aus diesen Gründen wird hier gemäß Steller (1983a, b, 1987) an den neutraleren Bezeichnungen „**Psychophysiologische Aussagebeurteilung bzw. Aussagebegutachtung**“ festgehalten, zumal sie aktuell gebräuchlich sind (z. B. Salzgeber, Stadler & Vehrs, 1997, S. 213; Vossel & Zimmer, 1998, S. 191). Als Oberbegriff bietet sich auch „**Forensische Psychophysiologie**“ an (vgl. „Forensic Psychophysiology“, Yankee, 1995, S. 63f.), da er einerseits die notwendige enge Bindung des Forschungsbereichs an die psychophysiologische Grundlagenwissenschaft unterstreicht und andererseits auf die potentiellen Anwendungsfelder im Rahmen der Rechtspsychologie verweist.

In einer ersten **definitorischen Annäherung** ist festzuhalten, daß die psychophysiologische Aussagebeurteilung darauf abzielt, anhand eines intraindividuellen Vergleichs der körperlichen Reaktionen einer Person auf bestimmte Stimuli (in der Regel verbal dargebotene Fragen) mittelbar diagnostische Rückschlüsse auf die Glaubhaftigkeit von Aussagen bzw. die Kenntnis von Tatdetails zu ziehen (vgl. auch Fröhlich, 2000, S. 208). Die Überlegung, ob derartige Schlußfolgerungen überhaupt möglich sind, stellt eine genuin psychophysiologische Problemstellung dar, insofern als die Psychophysiologie „eine humanwissenschaftliche Disziplin [ist], die sich mit den Zusammenhängen von psychischen und körperlichen Vorgängen befaßt“ (Vossel & Zimmer, 1998, S. 18), bzw. – spezifischer – „die körperlichen Auswirkungen psychischer Prozesse zum Gegenstand hat“ (Velden, 1994, S. 17). Auf die im Rahmen der forensischen Psychophysiologie verwendeten psychologischen Reizbedingungen und physiologischen Meßtechniken gehen die nachfolgenden Abschnitte näher ein.

Die „**Psychophysiologische Aussageforschung**“ (Steller, 1987, S. 29ff.) untersucht die psychophysiologischen und psychometrischen Grundlagen der Verfahren. Ihr besonderes Augenmerk gilt dabei der Treffsicherheit. Dazu bedient sie sich im wesentlichen zweier Paradigmen (vgl. Ben-Shakhar & Furedy, 1990, S. 38ff.). Die beiden Vorgehensweisen werden im folgenden kurz erläutert (für eine ausführlichere Darstellung der Methoden und ihrer Vor- und Nachteile vgl. Abschnitt 2.9).

Mit den sog. **Feldstudien** wird eine Stichprobe real durchgeführter Tests untersucht. Deren Ergebnisse vergleicht man mit anderen Indikatoren der Glaubwürdigkeit bzw. Täterschaft der Pbn (z. B. Geständnisse, Gerichtsurteile oder Expertenentscheidungen)

und berechnet daraus die Trefferquoten. Bei den **Analogstudien** handelt es sich um experimentelle Simulationen. Mindestens zwei Bedingungen werden systematisch variiert. Meistens sollen die Versuchspersonen (Vpn) der Experimentalgruppe („Schuldige“, „Täter“) ähnlich wie in einem instruierten Rollenspiel ein „Scheinverbrechen“ („mock crime“) begehen (z. B. simulierter Diebstahl). Die Kontrollgruppe ist hinsichtlich dieser Tat „unschuldig“ („Nichttäter“). Anschließend absolvieren alle Vpn eine psychophysiologische Aussagebeurteilung, und die Gültigkeit der Befunde wird anhand der Gruppenzugehörigkeit überprüft. Steller (1987, S. 38) grenzt die Analogstudien von anderen **Laborexperimenten** ab, die nicht mit fingierten oder induzierten Delikten arbeiten bzw. keine realistische Testdurchführung simulieren. Dabei versucht man in der Regel die Kenntnisse der Pbn hinsichtlich bestimmter verheimlichter Informationen aufzudecken (z. B. gezogene Spielkarten, ausgedachte Zahlen oder autobiographische Daten). Aufgrund seiner geringen Realitätsnähe eignet sich dieser Ansatz aber nicht zur Abschätzung der Trefferquoten der Verfahren. Er erfüllt eine eher heuristische Funktion.

In der pychophysiologischen Aussageforschung hat es sich eingebürgert, sowohl bei Realfällen als auch in experimentellen Arbeiten Täter oder Nichttäter als „**schuldig**“ („guilty“) bzw. „**unschuldig**“ („innocent“) zu bezeichnen (vgl. Patrick & Iacono, 1991, S. 231; Steller, 1987, S. 15). Auch die vorliegende Arbeit hält an dieser Terminologie fest. Es sei aber betont, daß damit keine juristische oder gar moralische Bewertung intendiert ist. Die Begriffe dienen lediglich der vereinfachten Kennzeichnung der Gruppenzugehörigkeit in wissenschaftlichen Untersuchungen. Die Pbn werden anhand der Testergebnisse üblicherweise als „unglaubwürdig“ („**positiver Befund**“) oder „glaubwürdig“ („**negativer Befund**“) klassifiziert („deception indicated“ vs. „no deception indicated“). In Abhängigkeit davon, ob solche dichotomen Diagnosen zutreffen oder nicht, unterscheidet man zwischen wahren (validen) positiven/negativen Befunden und falsch positiven/negativen Befunden (vgl. Tabelle 1; zum Problem der Klassenzuordnung generell siehe Kallus & Janke, 1988, S. 135[1]). Zwei Arten von Trefferquoten geben die relativen Häufigkeiten bzw. die daraus geschätzten bedingten Wahrscheinlichkeiten korrekter Ergebnisse wieder (Lykken, 1991b, S. 216). Die **Sensitivität** bezieht sich auf die Rate valider positiver Entscheidungen unter den „Schuldigen“ (Anteil der korrekt als „unglaubwürdig“ diagnostizierten „schuldigen“ Pbn). Die **Spezifität** bezeichnet analog die Quote valider negativer Befunde unter den „Unschuldigen“ (Anteil der zutreffend als „glaubwürdig“ klassifizierten „unschuldigen“ Pbn).

[1] Im Gegensatz zu der differenzierteren Nomenklatur von Kallus und Janke (1988) werden hier die Begriffe Klassenzuordnung und Klassifikation synonym verwendet.

Tabelle 1. Nomenklatur der Kriteriumsvariable, der Testergebnisse und der Trefferquoten im Rahmen der psychophysiologischen Aussagebeurteilung mit Zahlenbeispielen

	Status hinsichtlich Täterschaft	
Testergebnis	„schuldig“ (z. B.: 100 Probanden)	„unschuldig“ (z. B.: 100 Pbn)
positiv („unglaubwürdig“)	valider positiver Befund (z. B.: 90 Pbn)	falsch positiver Befund (z. B.: 30 Pbn)
negativ („glaubwürdig“)	falsch negativer Befund (z. B.: 10 Pbn)	valider negativer Befund (z. B.: 70 Pbn)
Trefferquoten	Sensitivität (z. B.: 90%)	Spezifität (z. B.: 70%)

Anmerkung. Die Zahlenangaben in der Tabelle sind fiktiv und dienen nur der Veranschaulichung.

Auch die beiden Begriffe „**Täuschung**“ und „**Lüge**“ bedürfen im Rahmen der vorliegenden Arbeit einer näheren Erläuterung. Ungeachtet der Schwierigkeiten, einvernehmliche Definitionen festzulegen, werden in der psychologischen Literatur gemeinhin zwei grundlegende Aspekte der Täuschung in den Vordergrund gestellt: die Vorsätzlichkeit der Handlung und die kommunikative Übermittlung von Informationen, die nach Ansicht des Kommunikators nicht den Tatsachen entsprechen (Berning, 1992, S. 2; Furedy, 1986, S. 684; Zuckerman, DePaulo & Rosenthal, 1981, S. 3). D. h., eine Täuschung kann relativ unabhängig vom objektiven Wahrheitsgehalt der übermittelten Informationen sein. Entscheidend ist die subjektive Überzeugung der kommunizierenden Person sowie ihre Intention. Unter einer Lüge versteht man eine besondere Variante der Täuschung, nämlich deren verbale Form (Köhnken, 1990, S. 3). Im Gegensatz dazu ist „**Glaubwürdigkeit**“ durch das Fehlen einer Täuschungsabsicht definiert (Köhnken, 1990, S. 4). In diesem Fall übermittelt der Kommunikator Informationen, von denen er annimmt, daß sie zutreffen. Undeutsch (1967, S. 51ff.) unterscheidet ferner zwischen der allgemeinen Glaubwürdigkeit eines Person (im Sinne eines relativ stabilen Persönlichkeitsmerkmals) und der speziellen Glaubwürdigkeit (bzw. Glaubhaftigkeit) einer Aussage. In der modernen forensisch-psychologischen Glaubwürdigkeitsforschung ist letzteres Konstrukt von entscheidender Bedeutung, zumal sich der allgemeine personenbezogene Ansatz als relativ unbrauchbar erwiesen hat (Steller & Volbert, 1997, S. 15).

Die o. g. Begriffsbestimmungen umfassen selbstverständlich nur menschliche Interaktionen. Darüber hinaus findet man aber in der Tier- und Pflanzenwelt ebenfalls Phänomene, die über eine enggefaßte Definition hinaus im weitesten Sinne als Täuschungen zu bezeichnen sind (Ben-Shakhar & Furedy, 1990, S. 1). Dazu gehören auch Tarnung

und Mimikry. Entsprechend rät Furedy (1996a, S. 98) davon ab, menschliche Täuschungen einseitig unter ethischen Gesichtspunkten zu analysieren, d. h. als etwas moralisch Verwerfliches, das es möglichst aufzudecken gelte. Aus **evolutionstheoretischer Perspektive** könnte man Täuschung – phylogenetisch betrachtet – als eine menschliche Weiterentwicklung der Versteckreaktion („hiding response“) auffassen, die wiederum als Alternative zur „fight-or-flight response“ eine wichtige Anpassungsreaktion des Organismus darstelle (vgl. auch Ben-Shakhar & Furedy, 1990, S. 142).

2.2 Gesellschaftspolitischer und juristischer Stellenwert der Thematik

Die psychophysiologische Aussagebeurteilung zählt zu den **häufigsten Anwendungen psychophysiologischer Erkenntnisse und Methoden**. In Deutschland ist ihr Einsatz weniger verbreitet, zumal hier die strafrechtliche Verwertung der Testresultate von höchstrichterlicher Seite untersagt wurde. Die „Lügendetektion“ nimmt jedoch in vielen anderen Ländern einen zwar mitunter kontrovers diskutierten, aber dennoch bedeutsamen gesellschaftlichen Stellenwert ein. Statistische Hochrechnungen aus den achtziger Jahren schätzten die Zahl der jährlich in den USA durchgeführten „Lügendetektortests“ auf ein bis vier Millionen (Lykken, 1981, S. 2). Matte (1996, S. 5) geht davon aus, daß die forensische Psychophysiologie in 57 Ländern weltweit praktiziert wird. Er nennt einige Beispiele. Dazu gehören neben Nordamerika (USA und Kanada) auch afrikanische Staaten (z. B. Republik Südafrika) sowie europäische und asiatische Länder wie etwa Rumänien, Kroatien, Rußland, die Türkei, Israel, Indien, China, Japan, Südkorea und Malaysia. Lykken (1998, S. 51) ergänzt die Aufzählung um Mexiko, Pakistan, die Philippinen, Taiwan und Thailand. Salzgeber et al. (1997, S. 215) erwähnen weitere europäische Länder wie Italien, Niederlande und die Schweiz. Darüber hinaus wurde der „Lügendetektor“ bereits in Polen, Schweden und Großbritannien eingesetzt (vgl. Jaworski, 1990, S. 123, 2000, S. 24f.; Thornton, 1988, S. 150ff.; Undeutsch, 1983, S. 406).

Bemerkenswert ist, daß die praktische Anwendung vielfach die Grundlagenforschung überflügelte und hinter sich ließ. Besonders deutlich trat dieser Sachverhalt in den **USA** zutage. Dort konnte sich der „Lügendetektor“ v. a. im beruflichen Umfeld (Privatwirtschaft und öffentlicher Dienst) als diagnostisches Hilfsmittel für Personalentscheidungen etablieren. Bei diesen sog. „preemployment and periodic screenings“ (Honts, 1991, S. 96ff.) handelt es sich um kommerzielle Untersuchungen im Rahmen der Einstellungsauslese und regelmäßige Testungen der Belegschaft, die nur selten der Klärung eines konkreten kriminellen Tatverdachts dienen. Vielmehr zielen sie auf die Integrität, politische Gesinnung oder Loyalität der Arbeitnehmer gegenüber dem Arbeitgeber ab

(vgl. z. B. Furedy & Heslegrave, 1991a, S. 162). Aber auch im strafrechtlich relevanten Kontext wurde die „Lügendetektion“ zur Glaubwürdigkeitsbegutachtung herangezogen, bevor die psychologische Forschung ab den siebziger Jahren des vergangenen Jahrhunderts begann, sich im größeren Umfang systematisch mit der Thematik und ihrer Problematik auseinanderzusetzen (vgl. Steller, 1987, S. 28; Undeutsch, 1997, S. 303). Auf der Basis einer kritischen Analyse (z. B. Office of Technology Assessment [OTA], 1983) wurde die ausufernde Nutzung der psychophysiologischen Aussagebeurteilung im privatwirtschaftlichen Bereich eingedämmt („Employee Polygraph Protection Act“ von 1988; Iacono, 2000, S. 772) und ihre kriminalistische bzw. forensische Verwertung an strenge Restriktionen geknüpft (vgl. Berning, 1992, S. 211f.; Faigman, Kaye, Saks & Sanders, 1997, S. 556ff.; Honts & Perry, 1992, S. 362ff.). Im öffentlichen Dienst hingegen sind die Untersuchungen zur Personalselektion und turnusmäßige Überprüfungen von Angestellten v. a. in Hochsicherheitsbereichen (z. B. Geheimdienste und Rüstungsforschung) zum Zwecke der Spionageabwehr weiterhin zulässig (siehe Beardsley, 1999, S. 11f.; Honts, 1991, 96ff.).

In **Deutschland** hat der Bundesgerichtshof (BGH) 1954 die Verwendung von Polygraphen im Strafverfahren und in den Vorermittlungen mit Verweis auf die verfassungsrechtlich geschützte Menschenwürde des Angeklagten abgelehnt. Inzwischen wurden jedoch diese Bedenken durch ein **aktuelles Urteil** von 1998 wieder aufgehoben (für eine Übersicht zur juristischen Entwicklung vgl. Anhang A). Falls der Beschuldigte freiwillig an einer polygraphischen Testung mitwirkt, verstößt die Untersuchung nicht gegen Verfassungsgrundsätze (BGH, 2000, S. 312). D. h., es handelt sich auch nicht mehr um eine verbotene Vernehmungsmethode im Sinne des § 136 a Strafprozeßordnung (StPO). Statt dessen wurden nun Zweifel an der Zuverlässigkeit der Testergebnisse stärker hervorgehoben. Dem bisher in Deutschland ausschließlich angewandten Kontrollfragentest stellte der Senat jeglichen Beweiswert in Abrede (BGH, 2000, S. 319). In bezug auf den Tatwissentest, der bislang hierzulande noch nicht praktiziert wird, fiel die Ablehnung weniger rigoros aus. Seine Durchführung sei zumindest „zum Zeitpunkt der Hauptverhandlung“ (BGH, 2000, S. 328) nicht beweiskräftig. Daß sich der BGH konträr zu den überwiegenden Erwartungen der Jurisprudenz (vgl. Kommentar zum BGH-Urteil von Martin, 1999, S. 714) gegen eine strafrechtliche Verwertung ausgesprochen hat, ist unter anderem auf die dezidierte Kritik zurückzuführen, die von psychologischer und psychophysiologischer Seite gegenüber dem Kontrollfragentest geäußert wurde (z. B. Fiedler, 1999; Rill & Vossel, 1998).

Trotz der ablehnenden Haltung eröffnet die BGH-Entscheidung der forensischen Psychophysiologie eine **neue Perspektive** (Offe & Offe, 1999). Die Methodenkritik, die zur Zurückweisung führte, bezieht sich im wesentlichen auf eine bestimmte Befra-

gungstechnik (Kontrollfragentest) und steht unter dem Vorbehalt des derzeitigen Forschungsstandes (Hamm, 1999, S. 922f.). Falls es gelingt psychophysiologische Verfahren zu entwickeln, die valide individualdiagnostische Schlußfolgerungen über die Glaubhaftigkeit von Aussagen gestatten, müßte über deren strafrechtliche Verwertbarkeit neu entschieden werden (vgl. Geppert, 1999).

Darüber hinaus ist die Übertragbarkeit des BGH-Urteils auf das **Ermittlungsverfahren** und den **Zivilprozeß** umstritten. So könnten etwa Strafverfolgungsbehörden entsprechende Verfahren einsetzen (Zallmanzig, 1999). Und bereits vor dem aktuellen Urteil, als die strafrechtliche Verwertung der psychophysiologischen Aussagebegutachtung noch als unvereinbar mit der Menschenwürde galt, war deren zivilrechtliche bzw. familiengerichtliche Anwendung nach Ansicht einiger Autoren ungeklärt (z. B. Endres & Scholz, 1994, S. 473) bzw. sogar zulässig (vgl. Interview mit Prof. Siegfried Willutzki, Präsident des deutschen Familiengerichtstags; Höfling, 1998, S. 38ff.; zum weniger restriktiven Beweisrecht in der Freiwilligen Gerichtsbarkeit siehe auch Undeutsch, 1997, S. 304). Demzufolge sind die Zivilgerichte nicht zwingend an die Entscheidungen des BGH-Strafsenats gebunden. Insbesondere Richter untergeordneter Instanzen (z. B. auf Amtgerichtsebene) können die Testergebnisse weiterhin als Beweismittel berücksichtigen (vgl. Hamm, 1999, S. 923). Ferner diskutieren Fachkreise den Einsatz der psychophysiologischen Aussagebeurteilung in der **Therapie und Bewährungskontrolle** rechtskräftig verurteilter Sexualstraftäter (Greuel, 1998; Salzgeber & Stadler, 1997), nachdem erste Ansätze dazu aus Nordamerika vorliegen (vgl. Lalumiere & Quinsey, 1991a, b; Matte, 1996, S. 6).

2.3 Prinzip und Systematik der Verfahren

Bis zum Grundsatzurteil des BGH von 1998 zeigte sich im juristischen Schrifttum Deutschlands oftmals eine **mangelnde Differenzierung** zwischen den unterschiedlichen Methoden der psychophysiologischen Aussagebeurteilung (vgl. Kritik von Berning, 1992, S. 239). Zwar fand man bisweilen noch zu Beginn der Erörterungen Verweise auf die Existenz diverser Befragungstechniken, und in diesem Zusammenhang führten viele Autoren exemplarisch den Kontrollfragen- und den Tatwissentest an (siehe Eisenberg, 1993, S. 193; Frister, 1994, S. 306f.; Volckart, 1998, S. 139). Meistens jedoch wurden die Verfahren unter den Schlagworten „Lügendetektor“- bzw. „Polygraphen“-Tests subsumiert und ihre forensische Verwertbarkeit unabhängig von der jeweiligen Vorgehensweise diskutiert. Dies galt gleichermaßen sowohl bei ablehnender (z. B. Peters, 1975) als auch zustimmender Haltung (z. B. Schwabe, 1979). Mit der aktuellen Entscheidung zeigte der BGH (2000, S. 319ff.) eine differenziertere

Sichtweise, indem er seine methodische Kritik v. a. auf den Kontrollfragentest konzentrierte.

De facto existieren **verschiedene Befragungstechniken**, die jeweils spezifische Vor- und Nachteile aufweisen, wie z. B. hinsichtlich ihrer theoretischen und empirischen Fundierung oder ihrer Einsatzmöglichkeiten sowie Fehleranfälligkeit. Die Frage nach der diagnostischen Validität der psychophysiologischen Aussagebegutachtung kann nicht pauschal beantwortet werden (Ben-Shakhar & Furedy, 1990, S. 33). Dementsprechend ist der Forderung von Berning (1993) zuzustimmen: „Die Unterschiede der dargestellten Testverfahren machen auch in rechtlicher Hinsicht eine z. T. differenzierende Beurteilung erforderlich“ (S. 253).

Trotz der Diskrepanzen lassen sich einige grundlegende **Gemeinsamkeiten** der angewandten Verfahren abstrahieren (vgl. Furedy, 1986, S. 686f.; Honts, 1991, S. 93ff.). Allesamt erfordern sie die Messung körperlicher Veränderungen, während man dem Probanden[2] (Pb) eine Sequenz von Fragen darbietet. Üblicherweise erfaßt man die elektrodermale Aktivität (meist den elektrischen Hautwiderstand, selten die Hautleitfähigkeit) und die Respiration (thorakale und abdominale Atembewegungen). Ferner wird die Aktivität des kardiovaskulären Systems aufgezeichnet (in der Regel der sog. „relative arterielle Blutdruck“ im Oberarm und eventuell zusätzlich photoplethysmographisch die Durchblutung im Finger). Diese Funktionsmaße spiegeln zumindest partiell die Aktivität des peripheren vegetativen Nervensystems wider, das sich einer direkten willentlichen Kontrolle weitgehend entzieht (vgl. Vossel & Zimmer, 1998, S. 41). In experimentellen Untersuchungen werden gelegentlich andere physiologische Größen (mit-)erfaßt, wie etwa das Elektrokardiogramm (z. B. Podlesny & Raskin, 1978), die Pupillenweite (z. B. Bradley & Janisse, 1981) oder ereignisbezogene Potentiale des Elektroenzephallogramms (z. B. Farwell & Donchin, 1991; zu deren Problematik vgl. auch Bashore & Rapp, 1993, S. 14ff.). Darüber hinaus erfordern alle Verfahren der psychophysiologischen Aussagebeurteilung ein mehr oder weniger strukturiertes Interview, das dem eigentlichen Test vorangestellt ist. Während dieses sog. Vortest-Interviews werden die Pbn auf die eigentliche Untersuchung vorbereitet. In der Regel erfolgt unmittelbar im Anschluß an den Test die Auswertung der physiologischen Aufzeichnungen.

Substantielle **Unterschiede** zwischen den Methoden bestehen hinsichtlich der Gestaltung des Vortest-Interviews, der Art der dargebotenen Fragen und der Vorgehensweise

[2] Der sprachlichen Einfachheit halber wird für Probanden und Untersucher die männliche Form gewählt, womit selbstverständlich beide Geschlechter gemeint sind.

bei der Auswertung der Daten. Im großen und ganzen lassen sich zwei Kategorien unterscheiden: **direkte und indirekte Verfahren** (Steller, 1987, S. 6ff.; Tent, 1967, S. 229; Undeutsch, 1983; S. 402ff.).

Anhand der **indirekten Verfahren** („concealed information tests“, vgl. Honts, 1991, S. 93; „information tests“, Podlesny & Raskin, 1977, S. 785), zu denen auch der **Tatwissentest** gehört (vgl. Abschnitt 2.7), versucht man festzustellen, ob eine Person Kenntnisse über tatrelevante Informationen besitzt. D. h., die Fragen beziehen sich auf nähere Begleitumstände des Verbrechens, die neben den Ermittlungsbehörden nur Tatbeteiligte kennen. Ungeachtet der konzeptionellen Vorteile und vielversprechenden laborexperimentellen Validitätsbefunden (vgl. Abschnitt 2.9.2) spielen die indirekten Verfahren im Rahmen der angewandten psychophysiologischen Aussagebeurteilung Nordamerikas nur eine sekundäre Rolle (Honts, 1991, S. 93). Indes wird der Tatwissentest beispielsweise in Israel (vgl. Elaad, 1990, S. 522) und Japan (vgl. Yamamura & Miyata, 1990, S. 263) relativ oft im Zuge polizeilicher Ermittlungen eingesetzt (zusammenfassend Ben-Shakhar & Furedy, 1990, S. 118ff.). Seine Anwendung setzt voraus, daß noch nicht allzu viele Tatdetails an die Öffentlichkeit gedrungen sind. Wenn auch unbeteiligte Personen Tatwissen erlangt haben (etwa über die Medien, polizeiliche Vernehmungen oder Mundpropaganda), sind die Einsatzmöglichkeiten stark eingeschränkt oder nicht mehr gegeben.

Weitaus gängiger in der Praxis sind die **direkten Verfahren** („detection of deception tests“, vgl. Honts, 1991, S. 93; „deception tests“, Podlesny & Raskin, 1977, S. 785; „guilty person tests“, Waid, Orne & Wilson, 1979, S. 16). Sie zeichnen sich dadurch aus, daß man durch den Vergleich der körperlichen Reaktionen auf bestimmte Fragen diagnostische Schlußfolgerungen über die Glaubwürdigkeit einer Person bei der Verneinung des Tatvorwurfs zieht. Die entsprechenden Tests umfassen sowohl relevante Fragen, die direkt auf die Tatbegehung abzielen, als auch unterschiedliche Arten von Vergleichsfragen, die in keinem unmittelbaren Zusammenhang mit dem betreffenden Verbrechen stehen. Je nach Art der Vergleichsfragen kann man zwei weitere Subgruppen gegeneinander abgrenzen. Bei den **Relevant-Irrelevant-Techniken** werden den relevanten Fragen relativ triviale neutrale Fragen gegenübergestellt, die die Pbn wahrheitsgemäß beantworten (z. B.: „Ist Ihr Vorname ...?“). Im Gegensatz dazu kombinieren die **Kontrollfragentests** tatbezogene Items und emotional belastende Vergleichsfragen. Letztere zielen nicht direkt auf die Tat ab, sondern auf andere Normverstöße (z. B.: „Haben Sie jemals gelogen, um sich aus einer prekären Situation zu retten?“). Diese Vergehen sind weniger schwerwiegend als das eigentliche Delikt, das im Zentrum der Ermittlungen steht. Der Pb soll dennoch die Kontrollfragen verneinen und dabei bewußt lügen bzw. am Wahrheitsgehalt seiner Antworten zweifeln. Alle direkten Verfahren

basieren auf der **Prämisse**, daß die schuldigen Pbn den konkreten Tatvorwurf der relevanten Fragen wahrheitswidrig abstreiten und dabei stärkere körperliche Reaktionen zeigen (erhöhte relevante Reaktionen). Im Gegensatz dazu sollen Unschuldige auf die Vergleichsfragen zumindest ähnlich oder stärker reagieren (erhöhte Vergleichsreaktionen).

2.4 Kontrollfragentest (KFT)

2.4.1 Vorgehensweise

Der von Reid (1947) konzipierte „**Kontrollfragentest (KFT)**“ („control question test“ [CQT], Raskin, 1981, S. 13) gilt unter seinen Befürwortern als das Standardverfahren der angewandten forensischen Psychophysiologie (vgl. Honts, Kircher & Raskin, 1995, S. 199). Es sei aber darauf hingewiesen, daß diese Befragungstechnik strenggenommen kein einheitliches Format besitzt. Dahinter verbirgt sich ein umfangreiches Methodeninventar. Dennoch wird hier von „dem Kontrollfragentest“ im Singular gesprochen, da sich diese Bezeichnung als Sammelbegriff für die ganze Gruppe von Verfahren eingebürgert hat. Damit soll allerdings deren Heterogenität nicht außer acht gelassen werden.

Die hier beschriebene Methode ist die sog. „**Zone Comparison Technique**“ nach Backster (1962, zitiert nach Podlesny & Raskin, 1977, S. 786), von deren ursprünglichen Gestaltung wiederum diverse Modifikationen abgeleitet wurden (vgl. Barland & Raskin, 1975a, S. 324; Undeutsch, 1983, S. 405). Da jedoch die spätere Kritik im Prinzip alle Versionen von Kontrollfragentests betrifft, ist an dieser Stelle keine weitere Nuancierung nötig. Die einzige Ausnahme ist der sog. „Directed Lie Test“ (vgl. Lykken, 1998, S. 137ff.; Raskin, Kircher, Horowitz & Honts, 1988, S. 8ff.), auf den der Abschnitt 2.5 gesondert eingeht. Das Prinzip des KFT wird im folgenden anhand der „Utah Zone Comparison Technique“ (vgl. Matte, 1996, S. 368) in der Version von Raskin und Mitarbeitern (z. B. Raskin & Hare, 1978, S. 128) exemplifiziert.

Eine Untersuchung mit dem KFT läßt sich in ungefähr **drei bis vier Phasen** unterteilen: Vortest-Interview, Testphase, Auswertung und ein eventuell durchgeführtes Nachtest-Interview (einen Überblick geben etwa Ben-Shakhar & Furedy, 1990, S. 8f.; Salzgeber et al., 1997, S. 217ff., Steller, 1987, S. 12ff.; Undeutsch & Klein, 1999, S. 67).

Die Hauptfunktion des ca. 30- bis 60minütigen **Vortest-Interviews** bestehen darin, die Fragen für die anschließende Testphase zu generieren. Ausgehend vom Tatvorwurf werden im Einvernehmen mit dem Pbn die relevanten, d. h. tatbezogenen Fragen fest-

gelegt (z. B. bei einem Schmuckdiebstahl: „Haben Sie den betr. Ring aus der Schublade entwendet?“, vgl. Undeutsch, 1983, S. 405). Außerdem formuliert der Untersucher auf der Basis einer biographischen Exploration die Kontrollfragen. Diese beziehen sich zwar nicht direkt auf die Tat, sie thematisieren aber ähnliche, emotional belastende Normverstöße (z. B.: „Haben Sie vor Ihrem 19. Lebensjahr jemals irgendwelches Geld gestohlen?“, Undeutsch, 1983, S. 405). Obwohl die Vergehen in Relation zum Tatbestand der eigentlichen Ermittlungen (Ringdiebstahl) weniger schwerwiegend sind, sollen die Kontrollfragen eine beunruhigende Wirkung nach sich ziehen. Folglich muß der Untersucher deren Bedeutsamkeit für den Pb manipulativ erhöhen. Der Beschuldigte soll zur Überzeugung gelangen, daß die Antworten und Reaktionen auf die Kontrollfragen entscheidend für die Beurteilung seiner Glaubwürdigkeit seien.

Zu diesem Zweck **suggeriert** der Untersucher dem Pb, daß etwaige Zugeständnisse den konkreten Tatverdacht gegen ihn zusätzlich erhärten (Raskin & Kircher, 1991, S. 217). Dies geschieht nach dem Motto: einer Person, die das Vergehen der Kontrollfrage begangen hat, traut man auch das Delikt der relevanten Frage zu (Honts et al., 1995, S. 200). Räumt der Pb dennoch entsprechende Verstöße ein (z. B. Bekennen eines Gelddiebstahls im Alter von 18 Jahren), muß der Untersucher durch weiteres Nachhaken den Pb dazu bringen, die Kontrollfrage zu verneinen. Dabei wird sie jeweils unter Ausschluß der eingestandenen Taten umformuliert (z. B.: „Haben Sie vor Ihrem 18. Lebensjahr ...?“ oder „Abgesehen davon, was Sie bisher erzählt haben, haben Sie ...?“). Außerdem soll der Untersucher seine Bedenken hinsichtlich der Bekenntnisse und der Unschuld des Verdächtigen kundtun. Wichtig ist, daß der Pb nicht sämtliche früheren Vergehen aufzählt. Aufgrund des impliziten Drucks, sich mit jedem weiteren Eingeständnis zunehmend verdächtig zu machen, soll er die Kontrollfrage möglichst bald verneinen. Man nimmt an, daß er dabei entweder bewußt lügt oder zumindest am Wahrheitsgehalt der Antwort zweifelt. Seine Unsicherheit wird durch die vage Formulierung und den breiten Zeitrahmen der Frage zusätzlich gesteigert.

Nach diesem Muster legt der Untersucher drei Paare von relevanten Fragen und Kontrollfragen fest. Eine beispielhafte Sequenz mit den erwarteten Antworten ist in der Tabelle 2 veranschaulicht.

Neben diesen drei Paaren werden noch vier **zusätzliche Items** formuliert, auf die der Pb mit Ja antworten muß. Dabei handelt es sich einerseits um zwei vollständig irrelevante, d. h. neutrale Fragen (z. B. nach dem Vor- und Nachnamen), die v. a. zur Stabilisierung der körperlichen Erregung während der Testphase dienen. Ein anderer Itemtyp bezieht sich zwar auch auf den Gegenstand der Ermittlungen, für die Beurteilung der Glaubwürdigkeit ist er aber ohne Belang. Diese „sacrifice relevant“ (Lykken, 1998, S. 115f.)

bzw. „preparatory relevant question“ (Matte, 1996, S. 325) soll den Pb auf die Fragen nach der Tat vorbereiten (Barland & Raskin, 1975a, S. 324). Starke Reaktionen auf die sog. „outside issue“ (Lykken, 1998, S. 115f.) bzw. „symptomatic question“ (Matte, 1996, S. 324) sollen auf Befürchtungen des Pb hindeuten, daß der Test andere Bereiche ansprechen könnte, die zwar nichts mit dem eigentlichen Verbrechen zu tun haben, ihn aber ebenfalls beunruhigen (z. B. ein anderes bisher ungeklärtes Delikt, das er begangen hat).

Tabelle 2. Beispiel für die Fragen und Antworten eines KFT (nach Undeutsch, 1983, S. 405)

Nr.	Typ	Frage	Antwort
1.	I	Ist Ihr Nachname ...?	J
2.	S	Bezüglich des Ringdiebstahls – haben Sie die Absicht, alle diesbezüglichen Fragen wahrheitsgemäß zu beantworten?	J
3.	O	Glauben Sie mir, daß ich Ihnen nur Fragen stellen werde, die wir zuvor vereinbart haben?	J
4.	K	Haben Sie vor Ihrem 19. Lebensjahr jemals irgendwelches Geld gestohlen?	N
5.	R	Haben Sie den betreffenden Ring genommen?	N
6.	K	Haben Sie während der Schulzeit irgendeinen Gegenstand von Wert entwendet?	N
7.	R	Haben Sie den betreffenden Ring aus der Schublade entwendet?	N
8.	I	Heißen Sie mit Vornamen ...?	J
9.	K	Haben Sie jemals einen Menschen, der berechtigt war, die Wahrheit zu erfahren, belogen, um sich unangenehme Konsequenzen fernzuhalten?	N
10.	R	Waren Sie irgendwie an dem Diebstahl des betreffenden Ringes beteiligt?	N

Anmerkungen. Typ = Fragentyp: I = irrelevante Frage, S = „sacrifice relevant“, O = „outside issue“, K = Kontrollfrage, R = relevante Frage; Antwort: J = Ja, N = Nein.

Während des Vortest-Interviews wird dem Pb außerdem suggeriert, daß starke körperliche Reaktionen auf die Kontrollfragen ein Täterschaftsindiz seien. Eine solche Reaktionstendenz würde ihn als „unglaubwürdig“ entlarven und somit auch seine Glaubhaftigkeit beim Abstreiten des aktuellen Tatvorwurfs diskreditieren (vgl. Raskin, 1982, S. 325, 1986, S. 34). Diese Angabe entspricht nicht den tatsächlichen Gegebenheiten. Das Gegenteil trifft zu. Stärkere Reaktionen auf die Kontrollfragen führen zu einem negativen Befund („glaubwürdig“). Am Ende des Vortest-Interviews wird der genaue Wortlaut und die Reihenfolge der insgesamt zehn Fragen festgelegt. Die Sequenz beginnt mit einem irrelevanten Item, gefolgt von der „sacrifice relevant“- und der „outside issue“-Frage. Danach werden die drei Paare von relevanten und Kontrollfragen positioniert, wobei noch ein irrelevantes Items zwischengeschaltet ist (vgl. Tabelle 2).

Die **Testphase** (Dauer ca. 20 – 30 Minuten, vgl. Barland & Raskin, 1975a, S. 324; Furedy, 1996a, S. 99) beginnt mit dem Anbringen der Signalabnehmer. Dem Pb wird die Funktionsweise des vegetativen Nervensystems und des Polygraphen grob erklärt. Empirische Befunde deuten darauf hin, daß die Treffsicherheit steigt, wenn der Pb von der Unfehlbarkeit des Verfahrens überzeugt ist (vgl. zusammenfassend Ben-Shakhar & Furedy, 1990, S. 62ff.; Steller, 1987, S. 63f.). Darum instruiert ihn der Untersucher, daß Täuschungsversuche körperliche Veränderungen nach sich ziehen würden, die mit Hilfe des Polygraphen registrierbar seien (Undeutsch & Klein, 1999, S. 67).

Zur Demonstration der Treffsicherheit führt man in der Regel sog. **Stimulationstests** durch („stim test“, Furedy, 1996a, S. 99), die auch dazu dienen, die physiologische Reaktivität des Pb abzuschätzen und die Meßapparatur zu justieren. Dabei handelt es sich meist um Zahlen- oder Kartentests (zu den verschiedenen Vorgehensweisen siehe Matte, 1996, S. 307, S. 311). Beispielsweise wird der Pb aufgefordert, eine Spielkarte aus einem Stapel zu ziehen und sich diese einzuprägen, ohne sie dem Untersucher zu zeigen. Unter Aufzeichnung der physiologischen Reaktionen stellt der Untersucher Fragen nach der Farbe und dem Wert der gezogenen Karte, die der Pb stets verneint (eine ausführliche Darstellung des Ablaufs findet sich bei Lykken, 1998, S. 14f.). Obwohl in den meisten Fällen bereits nach wenigen Durchgängen die stärkste Reaktion auf die zutreffende Alternative auftreten dürfte, greifen viele Untersucher auf Tricks zurück, wie z. B. einen Stapel gezinkter bzw. stets gleicher Karten, um dem Pb mit Gewißheit ein richtiges Ergebnis rückmelden zu können (vgl. Barland & Raskin, 1975a, S. 321; Reid & Inbau, 1977, S. 42).

Im Anschluß an den Stimulationstest findet die **eigentliche Befragung** mit kontinuierlicher Messung der physiologischen Variablen statt. Die Darbietung der Sequenz erfolgt genau so, wie im Vortest-Interview besprochen. Dadurch sollen etwaige Überraschungseffekte vermieden werden. Die Kontrollfragen und relevanten Fragen wechseln einander ab, um die potentiellen Einflüsse von Habituations- bzw. Sensitivierungsprozessen auf die Reaktionsunterschiede zu reduzieren (vgl. Podlesny & Raskin, 1977, S. 786). Die Angaben zu den zeitlichen Abständen zwischen den Items schwanken zwischen 10 bis 15 (Barland & Raskin, 1975a, S. 324; Lykken, 1998, S. 14) und 30 bis 35 Sekunden (Furedy, 1996a, S. 99; Raskin & Hare, 1978, S. 128). Die Sequenz wird mehrfach, d. h. in drei bis fünf Durchgängen (sog. „charts“, vgl. Furedy, 1996a, S. 99), dargeboten. Mitunter variiert dabei die Reihenfolge der Kontrollfragen (Lykken, 1998, S. 15; Matte, 1996, S. 368). Die einzelnen Durchgänge sind durch Pausen voneinander abgegrenzt, die auch dazu dienen, die pneumatische Manschette am Oberarm zu lösen, um einen schmerzhaften Blutstau zu verhindern (Iacono & Lykken, 1997, S. 583). Außerdem werden wieder die Kontrollfragen angesprochen und wenn nötig umformu-

liert. Auf diese Weise wird die Aufmerksamkeit des Pb erneut auf diese Fragen gelenkt und deren Relevanz nochmals betont (Raskin, 1979, S. 593). Die Testphase endet mit der Entscheidung des Untersuchers, daß keine weiteren Durchgänge mehr nötig sind.

Eine wichtige **Grundannahme** des KFT besagt, daß Täter erhöhte Reaktionen auf die relevanten Fragen zeigen, während Unschuldige stärker auf die Kontrollfragen reagieren sollen. Hinsichtlich der Gründe, weshalb solche Unterschiede zu erwarten sind, existieren nur wenig fundierte Hypothesen (vgl. Abschnitt 2.10). Ungeachtet der mangelnden theoretischen Fundierung erfolgt die **Auswertung** der physiologischen Aufzeichnungen anhand eines entsprechenden Reaktionsvergleichs (Undeutsch, 1983, S. 405). Stärkere Reaktionen auf die tatbezogenen Fragen (relevante Reaktionen) sind ein Anzeichen für einen positiven Befund („unglaubwürdig“, „deception indicated“). Stärkere Reaktionen auf die Kontrollfragen (Vergleichsreaktionen) führen zu einem negativen Befund („glaubwürdig“, „no deception indicated“). Falls keine eindeutigen Unterschiede resultieren, wird der Test als „unentscheidbar“ („inconclusive“) gewertet.

Zur **Abschätzung der Reaktionsstärken** werden verschiedene Merkmale und Parameter herangezogen (vgl. Barland & Raskin, 1975a, S. 325; Kircher & Raskin, 1988, S. 294; Matte, 1996, S. 371–397; Podlesny & Raskin, 1977, S. 789ff.; Raskin & Hare, 1978, S. 129; Undeutsch & Klein, 1999, S. 70). In bezug auf die Respiration sind v. a. eine verminderte Atemfrequenz und -tiefe sowie ein Anstieg der Baseline von Belang. Bei den elektrodermalen Reaktionen wird besonders auf deren Amplitude, Dauer und/oder Mehrgipfeligkeit (Komplexität) geachtet. Die erwarteten Reaktionen des kardiovaskulären Systems sind Veränderungen der Pulsfrequenz und ein Anstieg der Grundlinie des „relativen Blutdrucks“ sowie eine Abnahme des peripheren Blutvolumens und der Pulsvolumenamplitude im Plethysmogramm.

Bei der Auswertung und Interpretation der Aufzeichnungen kann man **drei Vorgehensweisen** unterscheiden: den klinischen und den numerischen Ansatz sowie die computergestützte Diagnose.

Im Rahmen des **klinischen Ansatzes** wird eine sog. „globale“ Auswertung durchgeführt (Lykken, 1998, S. 93ff.; Matte, 1996, S. 450ff.; zur klinischen Vorgehensweise allgemein vgl. auch Amelang & Zielinski, 1997, S. 146). Der Untersucher entscheidet anhand eines Überblicks über die polygraphischen Aufzeichnungen eher intuitiv, ob stärkere Reaktionen auf die relevanten Fragen oder Kontrollfragen vorliegen. Ferner fordert der klinische Ansatz ausdrücklich die Mitberücksichtigung zusätzlicher Informationsquellen bei der Urteilsbildung (Matte, 1996, S. 450). Neben den polygraphischen Daten fließen auch Bewertungen des verbalen und non-verbalen Verhaltens des Pb sowie die

Kenntnisse des Untersuchers über die Aktenlage in die Diagnose ein. Beispielsweise empfehlen Reid und Inbau (1977, S. 292ff.) auf charakteristische „Verhaltenssymptome“ („behavior symptoms“) von Lügen (z. B. Meiden von Augenkontakt) und Aufrichtigkeit (z. B. kooperatives Verhalten während des Tests) zu achten. Um die Erhebung der Verhaltenssymptome zu erleichtern, wurden Checklisten entwickelt. Diese werden am Ende der Untersuchung mit den polygraphischen Resultaten verglichen. Bei mangelnder Übereinstimmung soll keine Diagnose über die Glaubhaftigkeit des Pb erfolgen („unentscheidbar“), während kongruente Ergebnisse sich gegenseitig stützen und den endgültigen Befund im Sinne eines Doppelbelegs absichern können (Matte, 1996, S. 450).

Die Vertreter des **numerischen Ansatzes** hingegen beharren darauf, daß die Diagnose nur anhand der physiologischen Daten erfolgen soll (vgl. Kircher & Raskin, 1988, S. 292; Matte, 1996, S. 322). Zur Quantifizierung der Reaktionsunterschiede wird ein sog. „semi-objektives“ numerisches Schätzverfahren herangezogen („numerical scoring“, Podlesny & Raskin, 1977, S. 786), das ursprünglich auf Backster (vgl. Undeutsch, 1983, S. 405) zurückgeht und später von der Forschungsgruppe um Raskin (z. B. Raskin & Hare, 1978, S. 129) adaptiert wurde. Für jede einzelne physiologische Variable und jedes Fragenpaar werden die relevanten Reaktionen und Vergleichsreaktionen gegenübergestellt. Wenn keine eindeutigen Unterschiede erkennbar sind, resultieren null Punkte. Andernfalls vergibt der Untersucher in Abhängigkeit davon, ob er die Differenz als „leicht“, „deutlich“ oder „stark“ einschätzt, Werte von ±1, ±2 oder ±3. Die Vorzeichen sind durch die Richtung der Reaktionsunterschiede definiert. Ein negatives Vorzeichen resultiert bei einer stärkeren relevanten Reaktion. Ein positiver Wert bedeutet eine erhöhte Reaktion auf die Kontrollfrage. Die Punktwerte werden unter Berücksichtigung der Vorzeichen über alle Fragenpaare und physiologischen Variablen addiert. Liegt die Summe im Indifferenzbereich von -5 bis +5, so bleibt das Ergebnis „unentscheidbar“. Bei einem Gesamtwert +6 und höher wird der Pb als „glaubwürdig“ klassifiziert (negativer Befund). Beträgt die Summe -6 oder weniger, dann resultiert ein positiver Befund („unglaubwürdig“ bei der Verneinung des Tatvorwurfs)[3]. Die Bereichsgrenzen wurden weniger unter theoretischen oder empirischen Gesichtspunkten, sondern rein pragmatisch festgelegt (Steller, 1987, S. 9). Neben den Trennwerten („cut-offs“) von ±6 (Podlesny & Raskin, 1978, S. 349; Raskin & Hare, 1978, S. 129) empfehlen beispielsweise andere Publikationen ±5 als Entscheidungskriterium (vgl. Barland & Raskin, 1975a, S. 325).

[3] Zur Vermeidung von Mißverständnissen ist zu beachten, daß die Richtung der Befunde und die Vorzeichen der Gesamtwerte genau entgegengesetzt ausfallen.

Die „klinische Lügendetektion“ („clinical lie detection“, Szucko & Kleinmuntz, 1981, S. 488) spielt in der Praxis keine gravierende Rolle mehr. Von größere Bedeutung sind – insbesondere in Deutschland (vgl. Salzgeber et al. 1997, S. 219) – der numerische Ansatz und zunehmend auch **computerisierte Auswertungsmethoden** (Kircher & Raskin, 1981, 1982, 1988; Olsen, Harris, Capps & Ansley, 1997). Durch die Verwendung computergestützter Polygraphen können die digitalisierten physiologischen Aufzeichnungen elektronisch weiterverarbeitet und einer automatischen Kennwertbildung unterzogen werden. Die Reaktionsvergleiche zwischen relevanten Fragen und Kontrollfragen erfolgen dann ebenfalls rechnergesteuert auf der Basis spezieller Algorithmen (Diskriminanzfunktionen, Bayes-Theorem; vgl. Fiedler, 1999, S. 21ff.; Honts, 1994, S. 80f.), die schließlich zu Wahrscheinlichkeitsaussagen über die Glaubwürdigkeit des Pb führen. Die zur Zeit handelsüblichen Computer-Polygraphen gestatten sowohl eine konventionelle als auch automatische Auswertung, so daß man letztere zusätzlich zum Verifizieren der numerischen Scores verwenden kann (Yankee, 1995, S. 65).

Nach der Auswertung und Interpretation der Befunde ist **der weitere Ablauf** vom Ergebnis abhängig. Speziell in der strafrechtlich relevanten Anwendung in Nordamerika, wo der KFT häufig von Ermittlungsbehörden zur Sondierung potentieller Tatverdächtiger eingesetzt wird, gibt es im wesentlichen drei Alternativen (vgl. Patrick & Iacono, 1991, S. 231). Für jene Pbn, die als „glaubwürdig“ eingeschätzt wurden, ist die Untersuchung für gewöhnlich beendet. Meist werden auch die Ermittlungen gegen sie eingestellt. Auf unentscheidbare Tests folgt – sofern nötig und möglich – eine erneute Untersuchung, mitunter auch nach einer zeitlichen Verzögerung von wenigen Wochen. Die Klassifikation „unglaubwürdig“ mündet routinemäßig in ein sog. **Nachtest-Interview**. Dabei handelt es sich im Grunde genommen um ein Verhör, das v. a. dazu dient, den Pb zu einem Geständnis zu bewegen. Auch in Deutschland erfolgen mitunter solche Nachbefragungen, wenn der KFT zur Glaubwürdigkeitsbegutachtung bei Verdacht auf sexuellen Kindesmißbrauch verwendet wird. Nach Salzgeber et al. (1997, S. 219f.) ist dabei ebenfalls mit weitreichenden Eingeständnissen der Pbn zu rechnen.

2.4.2 Kritik am Kontrollfragentest

Der Kontrollfragentest (KFT) ist in vielfältiger Weise sowohl unter **methodischen** als auch **ethischen Gesichtspunkten** kritisiert worden (vgl. Überblick von Rill & Vossel, 1998, S. 483ff.; Vossel, Gödert & Rill, 2001, S. 7f.). Die methodischen Vorbehalte betreffen insbesondere die diagnostische Testgüte, während die ethischen Bedenken v. a. auf konkrete Probleme in der Anwendung abzielen. Die folgenden Abschnitte gehen auf

die wesentlichen allgemeinen Kritikpunkte ein. Zunächst werden die *testtheoretischen Grundlagen der Analyse* erörtert und anschließend die Gütekriterien *Objektivität*, *Reliabilität* und *Validität* im Zusammenhang mit dem KFT diskutiert. Danach stehen *ethische Kritikpunkte* im Zentrum der Betrachtung. Spezielle Problemfelder, wie z. B. der Einsatz des KFT im Kontext des sexuellen Kindesmißbrauchs (Faller, 1997; Raskin & Steller, 1989; Cross & Saxe, 1992; Williams, 1995), im beruflichen Umfeld (Honts, 1991) oder zur Glaubwürdigkeitsbegutachtung etwaiger Opfer (Raskin, 1989, S. 288f.) bzw. Zeugen (Undeutsch & Klein, 1999, S. 125f.), werden an dieser Stelle nicht weiter vertieft. Dazu sei auf die jeweils angegebenen Quellen verwiesen.

Testtheoretische Grundlagen der Analyse:

Bis auf wenige Ausnahmen herrscht in der Literatur Konsens, daß die Evaluation der Methoden der forensischen Psychophysiologie auch aus testtheoretischer Perspektive erfolgen muß (z. B. Ben-Shakhar & Furedy, 1990, S. 33; Berning, 1992, S. 42; Blinkhorn, 1988, S. 30; Fiedler, 1999, S. 10f.; Undeutsch, 1979, S. 231, 1983, S. 408). Dabei sind jene Qualitätsstandards anzulegen, die sich in der Psychodiagnostik etabliert haben. Die grundlegenden Anforderungen, die man an psychometrische Testverfahren stellt, bezeichnet man als Gütekriterien. Drei **Hauptgütekriterien** lassen sich unterscheiden: Objektivität, Reliabilität und Validität (vgl. Amelang & Zielinski, 1997, S. 142ff.; Lienert & Raatz, 1994, S. 7ff.). Die **Objektivität** verweist auf die Unabhängigkeit der Testergebnisse von der Person des Untersuchers. Die **Reliabilität** (Zuverlässigkeit) bezeichnet den Grad der Meßgenauigkeit (im Sinne einer geringen Meßfehlerbehaftetheit), unabhängig davon, was der Test zu messen beansprucht. Die **Validität** (Gültigkeit) bezieht sich darauf, inwiefern ein Testverfahren das erfaßt oder vorhersagt, was es indizieren bzw. prädizieren soll.

Gemäß dem hier vertretenen Standpunkt beanspruchen die diversen Befragungstechniken der forensischen Psychophysiologie durchaus Testcharakter. Wie jedoch gezeigt wird, erfüllt der KFT nicht die Kriterien der modernen Psychodiagnostik. Somit bleibt die Verwendung des Begriffs „Test“ in diesem Zusammenhang äußerst fragwürdig (s. u.). Steller und Dahle (1999, S. 161f.) definieren die psychophysiologische Aussagebeurteilung als eine „**komplexe diagnostische Prozedur**“, die sich einer testtheoretischen Analyse weitgehend entzieht und v. a. auf der Basis ihrer allgemeintheoretischen Fundierung und empirisch ermittelten Trefferquoten beurteilt werden soll. Diese Sichtweise ist allerdings stark verkürzt. Sie verkennt, daß ein entsprechendes diagnostisches Verfahren erst durch die Gewährleistung basaler psychometrischer Anforderungen treffsichere Befunde erbringen kann. Der Wechselbeziehung zwischen den drei Gütekriterien ist nämlich hierarchischer Natur (vgl. Lienert & Raatz, 1994, S. 13f.). Die

Unabhängigkeit eines Verfahrens vom Untersucher (Objektivität) ist eine notwendige, aber keine hinreichende Bedingung für meßgenaue Ergebnisse. Ebenso setzt Validität Reliabilität voraus, wenngleich reliable Tests nicht unbedingt auch valide sein müssen.

Übertragen auf die **diagnostische Zielsetzung der psychophysiologischen Aussagebeurteilung** wäre die Objektivität eines Verfahrens dann hoch, wenn unabhängige Untersucher bei denselben Pbn zu gleichen Ergebnissen hinsichtlich deren aussagenbezogener Glaubwürdigkeit gelangen würden (hohe interpersonale Übereinstimmung der Kategorisierungen in „glaubwürdig“ vs. „unglaubwürdig“ bzw. „unentscheidbar“). Eine hohe Reliabilität würde (zumindest hypothetisch) bedeuten, daß man bei mehrfachen Untersuchungen bestimmter Pbn unter stets äquivalenten Bedingungen konsistent dieselben Befunde replizieren könnte (vgl. auch Ben-Shakhar & Furedy, 1990, S. 33). Die testtheoretische Zuverlässigkeit der psychophysiologischen Aussagebeurteilung muß abgegrenzt werden von der umgangssprachlichen oder juristischen Konnotation des Begriffs. Letztere bezieht sich eher auf die kriterienbezogene Validität, die im wesentlichen durch die Trefferquoten (Sensitivität und Spezifität) definiert ist. Ein anderer Aspekt der Gültigkeit, der bei der „Lügendetektion“ ebenfalls eine wichtige Rolle spielt, ist die Konstruktvalidität. Dort geht es um die Frage, welche psychischen Zustände und Prozesse mit den entsprechenden Verfahren erfaßt werden (Furedy, 1986, S. 687f.).

Insgesamt ist festzustellen, daß eine hohe Gültigkeit auch eine hohe Reliabilität und Objektivität impliziert (Lienert & Raatz, 1994, S. 13). Daher konzentrierte sich die psychophysiologische Aussageforschung bisher weitgehend auf die kriterienbezogene Validität, d. h. Treffsicherheit der Verfahren (vgl. Steller, 1987, S. 29). Die empirische Überprüfung der beiden untergeordneten Gütekriterien stößt bei der „Lügendetektion“ auf erhebliche methodische Schwierigkeiten, denn wiederholte Testungen derselben Pbn sind nicht ohne weiteres miteinander vergleichbar bzw. äquivalent (vgl. Undeutsch, 1979, S. 231). Diese Probleme entbinden jedoch nicht von der Notwendigkeit, zumindest grundsätzliche Überlegungen zu den basalen Gütekriterien anzustellen. Die folgenden Ausführungen werden zeigen, daß die Objektivität und somit auch die Reliabilität des KFT entgegen teils anderslautender Ansichten (z. B. Berning, 1992, S. 44, S. 52f.; Steller, 1987, S. 9; Undeutsch, 1983, S. 408) keinesfalls als gesichert gelten können. Daraus resultieren erhebliche Zweifel an dessen Validität.

Zur Objektivität des KFT:

Zunächst muß das Konzept der Objektivität weiter differenziert werden. Abhängig davon, in welcher Phase einer testgestützten Untersuchung Beeinträchtigungen auftreten

können, unterscheidet man zwischen **Durchführungs-**, **Auswertungs- und Interpretationsobjektivität** (Lienert & Raatz, 1994, S. 8). Die Objektivität steht in einem engen Zusammenhang mit dem Grad an **Standardisierung** eines Testverfahrens, d. h. der Konstanz und Einheitlichkeit der Untersuchungsbedingungen (Amelang & Zielinski, 1997, S. 26). Je weniger Einfluß der Untersucher auf den Testablauf sowie die Quantifizierung und Interpretation der Ergebnisse nehmen kann, desto unabhängiger sollten die diagnostischen Befunde von seiner Person sein.

Die **Durchführungsobjektivität** ist definiert als das Ausmaß, in dem Variationen im Verhalten des Untersuchers und Schwankungen der Untersuchungsbedingungen zu testrelevanten Verhaltensvariationen des Pb führen können (vgl. Amelang & Zielinski, 1997, S. 143). Beim KFT ist die Durchführungsobjektivität bereits aufgrund der Tatsache gefährdet, daß es „den Kontrollfragentest“ per se – im Sinne eines eindeutig spezifizierbaren Verfahrens – gar nicht gibt (Furedy, 1996b, S. 55f.). Vielmehr existieren unterschiedliche Modifikationen, deren wesentliche Gemeinsamkeit nur darin besteht, daß neben tatbezogenen Fragen auch emotional belastende Vergleichsstimuli dargeboten werden. Über die Art der zu verwendenden Kontrollfragen (z. B. hinsichtlich des Zeitraums, den sie umfassen) bestehen divergente Auffassungen (vgl. Horvath, 1988, S. 199; Podlesny & Raskin, 1978, S. 346). Gleiches gilt auch für das zahlenmäßige Verhältnis und die Abfolge von relevanten und Kontrollfragen sowie die Einbeziehung zusätzlicher Items (z. B. irrelevante, „outside issue“- oder „sacrifice relevant“-Fragen). Die professionellen „Lügendetektor“-Untersucher in Nordamerika („Polygraphers“) favorisieren und applizieren je nach Ausbildung unterschiedliche Techniken (eine ausführliche Zusammenstellung bietet Matte, 1996, S. 325–370, S. 431–465).

Zur **Erhöhung der Durchführungsobjektivität** wird allgemein empfohlen, die Untersuchungssituation weitgehend zu standardisieren sowie die soziale Interaktion zwischen Untersucher und Pb auf ein Minimum zu reduzieren (Amelang & Zielinski, 1997, S. 143; Lienert & Raatz, 1994, S. 8). Aufgrund der zentralen Rolle des Vortest-Interviews sind diese Forderungen beim KFT nicht ohne weiteres zu erfüllen. Variationen im Untersuchungsablauf lassen sich kaum vermeiden, da die Formulierung der Fragen individuell an das konkrete Delikt und den Beschuldigten adaptiert werden muß (Lykken, 1998, S. 34). Die Prozedur zeichnet sich also dadurch aus, daß sie entgegen entsprechender Empfehlungen (z. B. Berning, 1992, S. 43f.) nicht standardisierbar ist (vgl. Ben-Shakhar & Furedy, 1990, S. 10). Außerdem setzt das Vortest-Interview unabdingbar eine direkte persönliche Kommunikation zwischen Untersucher und Pb voraus. Der Untersucher muß nicht nur die Testfragen adäquat formulieren, sondern zusätzlich den Pb auf die anschließende Untersuchung einstimmen und ihn von der Treffsicherheit des Verfahrens überzeugen. Die Methode stellt erhebliche Anforderungen an seine

Kompetenz (Raskin, 1981, S. 17; Raskin & Hare, 1978, S. 134; Undeutsch, 1996, S. 331). Die Testergebnisse sind im hohen Maße vom Gelingen der suggestiven Manipulationen und somit von den psychologischen Fähigkeiten des Untersuchers abhängig (Lykken, 1998, S. 121; Steller, 1987, S. 15). Zwar könnten laut Undeutsch (1996, S. 331) gut ausgebildete „erfahrene Spezialisten“ stets einen adäquaten Testablauf garantieren, allerdings werden die Richtlinien eines entsprechenden Curriculums ebensowenig spezifiziert wie die geforderte Expertise. In aller Regel findet man lediglich Verweise auf die Lehrbedingungen in den USA (z. B. Undeutsch, 1983, S. 406). Aber gerade dort stoßen die erheblichen Qualitätsschwankungen der Schulungsmaßnahmen (vgl. Horvath, 1984, S. 252) und die oft mangelhafte Ausbildung und die fehlenden technischen sowie physiologischen Grundkenntnisse der Untersucher auf erhebliche Kritik (Honts, Raskin & Kircher, 1994, S. 258; Matte, 1996, S. 371). Unter psychodiagnostischen Gesichtspunkten ist es inakzeptabel, daß die Durchführung eines Testverfahrens derart stark von der Kompetenz des Untersuchers abhängt (vgl. Rill & Vossel, 1998, S. 485).

Die **Auswertungsobjektivität** verweist auf die quantitative oder kategoriale Auswertung des registrierten Testverhaltens nach definierten Regeln (Lienert & Raatz, 1994, S. 8). Der klinische Ansatz der psychophysiologischen Aussagebeurteilung mit der global-intuitiven Analyse der polygraphischen Aufzeichnungen impliziert eine geringe Auswertungsobjektivität (Berning, 1992, S. 43). Denn es wird keine Quantifizierung der Reaktionsunterschiede durchgeführt, und der Untersucher darf neben den physiologischen Daten auch zusätzliche subjektive Informationsquellen berücksichtigen (z. B. Verhaltenssymptome von Täuschung und Aufrichtigkeit). Die numerische Auswertung („numerical scoring“, vgl. Abschnitt 2.4.1) hingegen erweckt zunächst den Eindruck einer standardisierten, objektiven Vorgehensweise. Sie stellt aber lediglich ein „Schätzverfahren“ dar (vgl. Steller, 1987, S. 8), das man allenfalls als „semi-objektiv“ bezeichnen kann (Barland & Raskin, 1975a, S. 324f.). Das numerische Ergebnis ist sowohl von den polygraphischen Daten als auch den persönlichen Einschätzungen des Untersuchers abhängig (Honts et al., 1994, S. 254). Nach Furedy (1991, S. 244) handelt es sich um keine Quantifizierung im engeren Sinne. Die Zuordnung der Zahlenwerte (±1, ±2 und ±3) beruht nämlich nicht auf eindeutigen Vorschriften, sondern auf der subjektiven Beurteilung des Untersuchers, ob die Reaktionsunterschiede „leicht“, „deutlich“ oder „stark“ ausgeprägt sind. Zwar konstatiert Steller (1987, S. 9) ohne Quellenverweis, daß die Regeln sehr konservativ seien und man die Punktbeträge von 1, 2 bzw. 3 erst ab Reaktionsverhältnissen von 1:2, 1:3 bzw. 1:4 vergebe. Ob es sich dabei um die gleichen Kriterien handelt, die auch andere Auswerter verwenden (z. B. Barland & Raskin, 1975a, S. 325), bleibt ungewiß. Dementsprechend wird diese Auswertungsregel von O’Toole, Yuille, Patrick und Iacono (1994, S. 257) nur für die Amplituden der elektro-

dermalen Reaktionen bestätigt, während für die kardiovaskulären und respiratorischen Parameter andere Größenverhältnisse gelten sollen.

Es liegen keine verbindlichen Richtlinien vor, welche Kennwerte man zur **Parametrisierung** der Reaktionsstärken heranziehen soll (z. B. Amplitude, Dauer oder andere Verlaufscharakteristika). Lykken (1978, S. 138) weist darauf hin, daß die polygraphischen Kurven äußerst komplex und unspezifisch sind. Die Auswertungsmethode bietet hinreichend viel Ermessensspielraum, welche von zwei Reaktionen stärker ist bzw. ob eine Veränderung in den Aufzeichnungen überhaupt eine distinkte Reaktion darstellt oder nur einen Nacheffekt einer vorherigen Reaktion bzw. ein Artefakt. Bei der numerischen Auswertung können durchaus subjektive Faktoren eine wichtige Rolle spielen. Als Kontrollmaßnahme wird empfohlen, die Aufzeichnungen stets einer „blinden" Reanalyse zu unterziehen (z. B. Berning, 1992, S. 44; Raskin, 1981, S. 17). D. h., ein Polygrapher, der keine Kenntnisse über den Fall und die entsprechende Untersuchung besitzt, wertet die Kurven nochmals aus. Eine solche unabhängige Überprüfung der Ergebnisse kann allerdings die Folgen einer mangelnden Durchführungsobjektivität nicht kompensieren. Darüber hinaus wird dieser Forderung in der Praxis kaum entsprochen (Ben-Shakhar, 1991b, S. 197). Einige Praktiker vertreten sogar die Auffassung, daß nur derjenige Untersucher, der den Test durchgeführt hat, auch in der Lage ist, ihn angemessen auszuwerten (vgl. Barland & Raskin, 1975a, S. 329).

Die **Interpretationsobjektivität** gibt an, inwiefern identische Auswertungsergebnisse auch übereinstimmende Schlußfolgerungen nach sich ziehen (vgl. Amelang & Zielinski, 1997, S. 146). Daß der klinische Ansatz, bei dem die Intuition und der subjektive Eindruck des Untersuchers die Diagnose prägen, eine geringe Interpretationsobjektivität aufweist, ist unmittelbar evident. Im Gegensatz dazu wäre für den numerischen Ansatz eine akzeptable Konkordanz zu erwarten, wenn die Befunde („glaubwürdig", „unglaubwürdig" oder „unentscheidbar") allein auf den Gesamtpunktzahlen und obligatorischen Trennwerten basieren würden. Faktisch gibt es aber auch bei der numerischen Auswertung Probleme mit der Interpretationsobjektivität. Die Grenzen des Indifferenzbereichs („unentscheidbar") sind eigentlich willkürlich und unverbindlich gesetzt (z. B. ±5 oder ±6, vgl. Barland & Raskin, 1975a, S. 325; Podlesny & Raskin, 1978, S. 349; Raskin & Hare, 1978, S. 129). Die Untersucher können ihre diagnostischen Entscheidungsregeln nach Belieben modifizieren (Ben-Shakhar & Furedy, 1990, S. 50). Loftus (1982, S. 384f.) weist auf die Willkürlichkeit der Grenzen und deren möglichen Effekte auf die Trefferquoten hin. Aber auch die Befürworter möglichst einheitlicher numerischer Auswertungsregeln (z. B. Barland & Raskin, 1975a, S. 328) räumen ein, daß eine pauschale Festlegung der Bereichsgrenzen möglicherweise kein Optimum darstellt.

Die o. g. Hypothesen werden durch **empirische Befunde** gestützt. Ben-Shakhar, Lieblich und Bar-Hillel (1982, S. 710) kommen anhand der Analyse von mehreren Feldstudien unter Berücksichtigung entscheidungstheoretischer Erwägungen zu dem Ergebnis, daß symmetrische Trennwerte nicht ideal sind. Nach Patrick und Iacono (1991, S. 236) beruhen auch bei der Verwendung numerischer Auswertungsprozeduren die Entscheidungen über die Glaubwürdigkeit der Pbn nicht ausschließlich auf den physiologischen Daten. Ihre Feldstudie deutet darauf hin, daß der intuitive Gesamteindruck, den sich ein Untersucher auf der Basis seines Hintergrundwissens bildet, den Befund maßgeblich beeinflußt. In den diagnostischen Entscheidungsprozeß fließen komplexe Informationen ein, die sich aus der gesamten Interaktion zwischen Untersucher und Pb ergeben.

Die empirischen Ergebnisse decken sich mit Berichten aus der **Praxis**. So betont etwa Steinke (1987) aufgrund seiner Beobachtungen und Erfahrungen bei der israelischen Polizei „das stark subjektive Moment des Explorierers, der den gesamten Ermittlungsstand kennt und dem neben seiner psychologischen Sachkunde, die oftmals für Glaubwürdigkeitsfragen ausreicht, nur noch die Aufzeichnung einfacher Körperreaktionen ähnlich einem Kardiogramm zur Verfügung steht“ (S. 536). In diesem Zusammenhang äußert er Zweifel, „ob der die Wahrheit erkundende Psychologe in der Tat nur die Kurve interpretiert oder noch andere psychologische Bewertungen zur Hilfe nimmt“ (S. 536). Man kann also kaum davon ausgehen, daß die Untersucher – trotz numerischer Auswertungsverfahren – von nicht-physiologischen Informationsquellen gänzlich unbeeinflußt bleiben (Iacono, 1991, S. 205).

Computergestützte Polygraphen, die eine elektronische Datenaufbereitung sowie Wahrscheinlichkeitsaussagen über die Glaubwürdigkeit der Pbn bieten, sind von den o. g. Problemen der Auswertungs- und Interpretationsobjektivität weniger betroffen. Sie weisen aber andere Probleme auf. Die konkrete Auswertungsprozedur ist unklar und keiner neutralen Analyse zugänglich, da die entsprechenden Algorithmen nicht publiziert werden (Furedy, 1996b, S. 57). Auch bleibt die Frage offen, inwiefern die unterschiedlichen Programme, die auf dem Markt kursieren (vgl. Yankee, 1995, S. 64f.), vergleichbare Ergebnisse liefern. Die zugrundeliegenden Diskriminanzfunktionen wurden teilweise anhand experimenteller Analogstudien entwickelt und kreuzvalidiert, ohne ihre Generalisierbarkeit auf die Feldanwendung sicherzustellen (Honts, 1994, S. 81). Darüber hinaus sei betont, daß selbst bei einer computergestützten Auswertung die Güte der Befunde nicht gewährleistet ist (Furedy & Heslegrave, 1991b, S. 244). Denn die potentiellen Einflüsse der Untersucher auf die Auswahl und Darbietung der Fragen (vgl. Elaad, Ginton & Ben-Shakhar, 1994, S. 290) werden dadurch nicht eliminiert.

Zusammenfassend ist festzuhalten, daß manche kritischen Autoren (z. B. Furedy, 1996a, S. 98) dem KFT aufgrund seiner geringen Objektivität und Standardisierung den Status eines psychologischen Tests in Abrede stellen, sofern man darunter eine objektive und standardisierte Erhebung einer Verhaltensstichprobe versteht (vgl. „objective and standardized measure of a sample of behavior", Anastasi & Urbina, 1997, S. 4). Vielmehr handle es sich um eine komplexe, dynamische Interviewsituation (vgl. Furedy, 1991, S. 244). Die Anforderungen an die Kompetenz des Untersuchers sind enorm und übersteigen bei weitem das übliche Maß, das man an andere psychometrische Verfahren, z. B. moderne Intelligenztests, anlegt (Ben-Shakhar & Furedy, 1990, S. 11). Der Untersucher muß über interrogatives Geschick und eine „gute Menschenkenntnis" verfügen. Zudem bedarf es persuasiver Fähigkeiten, um die Pbn von der Unfehlbarkeit der Methode zu überzeugen, sowie einer umfangreichen psychophysiologischen Erfahrung bei der Auswertung der Aufzeichnungen. Selbst unter der Annahme, daß einzelne Untersucher all diesen Anforderungen gerecht werden und – wie viele Polygraphers von sich selbst behaupten – eine annähernd perfekte Treffsicherheit erzielen, kann man daraus keine Rückschlüsse auf die Validität der Methode als solche ziehen. In diesem Fall würden die hohen Trefferquoten weniger ein spezifisches Charakteristikum des Verfahrens darstellen, sondern wären eher auf die individuellen Fähigkeiten des Untersuchers zurückzuführen.

Die **Verbesserungsmöglichkeiten** des KFT sind im Hinblick auf seine Objektivität gewiß noch nicht ausgeschöpft, sofern es gelingt, die Durchführungs-, Auswertungs- und Interpretationsrichtlinien zu optimieren. Eine Bewertung des Verfahrens muß aber neben seinem Potential auch den Status quo in der Praxis berücksichtigen. Und gerade im anwendungsbezogenen Bereich zeigen sich kaum Anstrengungen, die fundamentalen Kritikpunkte anzuerkennen und nach Lösungen zu suchen. Dabei bleibt auch zweifelhaft, ob die Problematik der Standardisierung und Objektivierung im Kern überhaupt lösbar ist. Letztendlich gerät man unweigerlich in einen Konflikt zwischen der Forderung nach einer modernen Testkonstruktion unter psychometrischen Gesichtspunkten und der Notwendigkeit, die Fragen individuell an den jeweiligen Einzelfall anpassen zu müssen. Selbst neuere Ansätze, die versuchen, einige methodische Mängel des konventionellen KFT zu eliminieren (z. B. der Directed Lie Test, vgl. Abschnitt 2.5), beinhalten eine manipulative Komponente, und die gewünschte Wirkung der Fragen auf den einzelnen Pb kann nicht sichergestellt werden.

Zur Reliabilität des KFT:

Zur Abschätzung der **Meßgenauigkeit** eines psychologischen Testverfahrens gibt es verschiedene Methoden (z. B. Testwiederholungs-, Paralleltest- und Testhalbierungs-

methode bzw. Konsistenzanalyse), die teilweise unterschiedliche Fehlerquellen berücksichtigen (siehe Lienert & Raatz, 1994, S. 175ff.). Bislang liegen nur wenige empirische Arbeiten zur Zuverlässigkeit des KFT vor. Diese beschäftigten sich überwiegend mit der sog. „Inter-Rater-Reliabilität“ (vgl. Berning, 1992, S. 45ff.; Carroll, 1988, S. 20f.). Dabei werden mehrere polygraphische Aufzeichnungen von mindestens zwei unabhängigen Auswertern analysiert und deren Ergebnisse in Relation zueinander gesetzt. Meist geht man so vor, daß mehrere Untersucher eine Stichprobe physiologischer Datensätze nochmals „blind“ auswerten (ohne Kenntnisse vom konkreten Fall bzw. Ablauf der Originaluntersuchung). Anschließend werden die Resultate der Reanalysen untereinander bzw. mit den ursprünglichen Testbefunden verglichen. Die Inter-Rater-Reliabilität ist somit zugleich ein Maß für die interpersonale Auswerterübereinstimmung bzw. Auswertungsobjektivität (vgl. Steller & Dahle, 1999, S. 149; Undeutsch, 1983, S. 408). Da sich jedoch in den Methoden zur Bestimmung der Objektivität auch die Einflüsse einer mangelnden Meßgenauigkeit manifestieren können (Amelang & Zielinski, 1997, S. 143), wird insbesondere im anglo-amerikanischen Schrifttum die Objektivität mitunter als eine Facette der Reliabilität aufgefaßt. Folglich unterscheiden sich die Definitionen von Objektivität im internationalen Vergleich zu einem gewissen Grad (vgl. Lienert & Raatz, 1994, S. 7).

Angesichts der **ausgeprägten Subjektivität** des KFT (s. o.) berichten einige Publikationen (z. B. Barland & Raskin, 1975a, S. 326f.; Dawson, 1980, S. 13; Horvath, 1977, S. 132; Patrick & Iacono, 1991, S. 232; Podlesny & Raskin, 1978, S. 351) erstaunlich hohe prozentuale bzw. korrelative Übereinstimmungskennwerte (vgl. auch Übersicht von Undeutsch, 1979, S. 232f., 1983, S. 408f.). Damit können jedoch die Reliabilität bzw. Objektivität von Auswertung und Interpretation noch nicht als empirisch belegt gelten. Nach Ben-Shakhar und Furedy (1990, S. 11, S. 34) ist trotz mangelnder Objektivität des KFT in kontrollierten Studien mit hohen Inter-Rater-Reliabilitätskoeffizienten zu rechnen. Dies gilt insbesondere dann, wenn die Untersucher eine ähnliche Ausbildung absolviert haben, das gleiche Auswertungsverfahren anwenden und sich dessen bewußt sind, daß ihre Ergebnisse wissenschaftlich überprüft werden. In den meisten Arbeiten erfolgten die Erstauswertungen und die blinden Reanalysen durch Angehörige der gleichen Institution (z. B. Forschungsgruppen, private Polygrapher-Institute, Polizeidienststellen etc.). Die beteiligten Untersucher verfügten über einen weitgehend deckungsgleichen fachlichen Hintergrund. Insofern bestehen Zweifel an der effektiven Unabhängigkeit der Auswerter. Nach Ben-Shakhar und Furedy sind die gefundenen hohen Inter-Rater-Reliabilitäten eventuell auf ein Methodenartefakt zurückzuführen und nicht unmittelbar auf die Praxis generalisierbar, zumal dort ein sehr heterogener Ausbildungs- und Kenntnisstand der Untersucher vorherrscht.

Außerdem resultierte in einigen der genannten Studien (z. B. Barland & Raskin, 1975a) nur dann eine hohe Auswerterübereinstimmung, wenn man alle Fälle außer acht ließ, die von mindestens einem der Untersucher als „unentscheidbar“ eingestuft wurden. Es gingen also nur relativ eindeutig interpretierbare Aufzeichnungen in die Analysen ein. **Zweifelsfälle**, die für die Zuverlässigkeit des Auswertungsverfahrens besonders kritisch sind, blieben unberücksichtigt. Entsprechend lassen sich hohe Inter-Rater-Reliabilitäten nicht konsistent replizieren. Z. B. ließen Kleinmuntz und Szucko (1984, S. 450) in ihrer Feldstudie die Kategorie „unentscheidbar“ nicht zu. Einhundert polygraphische Aufzeichnungen wurden sechs Sekundärauswertern zur Reanalyse vorgelegt. Die Untersucher sollten allein anhand der physiologischen Daten dichotom eine eindeutige Diagnose über die Glaubwürdigkeit der Pbn treffen. Es zeigte sich eine niedrige durchschnittliche Korrelation zwischen ihren Befunden (r = .43) mit einer großen Streuung der Koeffizienten (.24 – .56). Zwar hält Berning (1992, S. 47f.) diese Studie für wenig repräsentativ, zumal auch keine Angaben vorliegen, ob die Untersucher ein einheitliches (numerisches) Auswertungsverfahren verwendeten. Die Ergebnisse deuten aber darauf hin, daß v. a. in der Feldanwendung des KFT nicht grundsätzlich mit einer hohen Auswerterübereinstimmung gerechnet werden darf (vgl. auch Carroll, 1988, S. 21).

Einen anderen Aspekt der Reliabilität bezeichnet man als **interne Konsistenz**. Bei der psychophysiologischen Aussagebeurteilung wird damit überprüft, inwiefern die unterschiedlichen physiologischen Reaktionsparameter Gleiches messen und somit interkorrelieren oder ob sie zumindest partiell voneinander unabhängig sind und unterschiedliche latente Variablen erfassen (vgl. Fiedler, 1999, S. 12). Die Logik des KFT legt hohe Interkorrelationen nahe. Schuldige und Unschuldige sollen konsistent, d. h. auch über mehrere Kanäle hinweg, auf die relevanten Fragen oder Kontrollfragen stärker reagieren. Eine solche Annahme vernachlässigt jedoch das Konzept der Reaktionsspezifität, also die Möglichkeit, daß aufgrund individual- bzw. stimulusspezifischer Reaktionsmuster (Engel, 1960, S. 305f.; Reaktions- bzw. Situationsstereotypien nach Lacey & Lacey, 1958, S. 73; vgl. Übersicht von Stern & Sison, 1990, S. 200ff.) personen- bzw. reizabhängige Diskrepanzen in den unterschiedlichen physiologischen Variablen auftreten können (siehe auch Lykken, 1960, S. 258). Dementsprechend bescheinigt Blinkhorn (1988, S. 32f., S. 39) dem KFT eine geringe interne Konsistenz. In der unveröffentlichten Untersuchung von Barland und Raskin (1976, zitiert nach Blinkhorn, 1988, S. 32f.) zeigten die kardiovaskulären und respiratorischen Maße praktisch keinen Zusammenhang. Nur die Stärke der Hautwiderstandsreaktionen korrelierte allenfalls mäßig mit den beiden anderen Variablen.

Insgesamt erlauben die bisherigen Reliabilitätsschätzungen keine eindeutigen Rückschlüsse auf die Zuverlässigkeit des KFT als Ganzes (Fiedler, 1999, S. 18). Die interne

Konsistenz der unterschiedlichen physiologischen Maße ist als relativ gering einzustufen. Und die Inter-Rater-Reliabilität berücksichtigt im wesentlichen nur Meßfehler, die auf eine mangelnde Zuverlässigkeit der Auswertung und der Interpretation zurückzuführen sind (Ben-Shakhar & Furedy, 1990, S. 34). Die gefundenen hohen Übereinstimmungsquotienten bzw. -koeffizienten deuten lediglich darauf hin, daß verschiedene Auswerter auf der Basis gleicher polygraphischer Aufzeichnungen zu kongruenten Befunden gelangen *können* (vgl. Lykken, 1998, S. 128f.). Die wirklich kritische Frage nach der Meßgenauigkeit besteht aber weniger im Hinblick auf die innere Konsistenz bzw. Auswertung und Interpretation der erhobenen polygraphischen Daten, sondern darin, ob das gesamte Verfahren – einschließlich des Vortest-Interviews (Formulierung der Fragen und suggestive Manipulationen) und der eigentlichen Testdurchführung (Fragendarbietung mit physiologischen Messungen) – zu reliablen Ergebnissen führt. Um dies zu überprüfen, müßte man eine Stichprobe von Pbn wiederholt unter äquivalenten Bedingungen von verschiedenen Untersuchern testen lassen und deren Ergebnisse miteinander vergleichen (im Sinne der Retest- bzw. Paralleltestreliabilität, vgl. Ben-Shakhar & Furedy, 1990, S. 33f.). Allerdings besteht ein gravierender Mangel an systematischen Untersuchungen zu derartigen Reliabilitätsschätzungen (Berning, 1992, S. 52; Blinkhorn, 1988, S. 39). Insofern lassen sich keine definitiven Aussagen über die Meßgenauigkeit des KFT treffen.

Zur Validität des KFT:

Lykken, ein besonders vehementer Kritiker des KFT (z. B. Lykken, 1974, 1978, 1979, 1988, 1991a), beanstandet v. a. die **mangelnde kriterienbezogene Validität** des Verfahrens. Seiner Ansicht nach liegt die Treffsicherheit in der praktischen Anwendung nur unwesentlich über dem Zufallsniveau (vgl. Lykken, 1991b, S. 214). Bei Tätern sei zwar durchaus mit einer hohen Entdeckungsrate zu rechnen, sofern es ihnen nicht gelinge, durch entsprechende Gegenmaßnahmen (s. u.) die polygraphischen Aufzeichnungen unentdeckt zu ihren Gunsten zu sabotieren. Aber auch viele Unschuldige würden fälschlich als „unglaubwürdig“ eingestuft: „Based on my analysis, one would expect that most – but not all – guilty subjects would fail the CQT – and that many innocent subjects would fail it also“ (Lykken, 1998, S. 124). Dieses Validitätsproblem basiere auf den unplausiblen Prämissen des KFT. Lykken bezweifelt die Annahme, man könne die Kontrollfragen generell so formulieren, daß sie für unschuldige Pbn eine größere Bedrohung darstellen würden als die relevanten Fragen. Der Erfolg der suggestiven Manipulationen während des Vortest-Interviews sei nicht kontrollierbar. Auch Unschuldige könnten die Relevanz der tatbezogenen Fragen für das Testergebnis erkennen. Wenn jedoch die eigentliche Testlogik des KFT durchschaut wird, droht ein Validitätsverlust. Lykkens Analyse impliziert somit die Gefahr, daß viele glaubwürdige Pbn die

relevanten Fragen als besonders bedrohlich bewerten und darauf stärker emotional sowie physiologisch reagieren als auf die Kontrollfragen. Diese Hypothese wird durch die psychophysiologische Aussageforschung zum KFT gestützt (vgl. die Abschnitte 2.9.1 und 2.9.2). Eine Reihe von Validitätsstudien zeigt für unschuldige Pbn ein überproportional hohes Risiko für **falsch positive Befunde** (z. B. Barland & Raskin, 1975a, S. 325; Horvath, 1977, S. 131; Kleinmuntz & Szucko, 1984, S. 450). D. h., die Wahrscheinlichkeit falscher Befunde liegt bei Unschuldigen insgesamt höher als bei Schuldigen (vgl. auch die Übersichtsarbeiten von Ben-Shakhar & Furedy, 1990, S. 41ff.; Berning, 1992, S. 115ff.; Steller, 1987, S. 35ff.). Aber auch die Möglichkeit **falsch negativer Befunde** ist nicht zu unterschätzen. Der KFT hat sich als anfällig gegenüber Gegenmaßnahmen erwiesen („countermeasures“, vgl. Honts et al., 1994, S. 252f.). Schuldige können etwa durch geeignete körperliche und mentale Manipulationsversuche (z. B. Muskelkontraktion, schmerzhafte Selbstreizung, Imagination emotionaler Situationen oder Kopfrechnen) ihre Reaktionen auf die Kontrollfragen gezielt erhöhen und dadurch die Sensitivität des Verfahrens reduzieren (zusammenfassend Gudjonsson, 1988, S. 128ff.).

Trotz der geringen Objektivität bzw. fraglichen Reliabilität und des von Lykken zunächst a priori postulierten Mangels an kriterienbezogener Validität ergeben viele empirische Arbeiten zum KFT (auch für Unschuldige) relativ hohe Trefferquoten. Dieses **Paradox** ist auf methodische Eigenarten der verwendeten Forschungsparadigmen zurückzuführen, die in einem späteren Abschnitt (2.9) näher erörtert werden.

Aber selbst wenn die Trefferquoten wie in den meisten publizierten Studien zumindest statistisch signifikant über dem Zufallsniveau liegen, belegen solche Ergebnisse noch nicht die diagnostische Güte des Verfahrens. Beim KFT verläuft nämlich der Entscheidungsprozeß nicht nach wissenschaftlichen Prinzipien (Ben-Shakhar, 1991a, S. 239). Wie bereits erwähnt, beruht die Glaubwürdigkeitsdiagnose nur zu einem geringen Teil auf den physiologischen Aufzeichnungen (vgl. Ben-Shakhar, 1991b, S. 196f.; Ben-Shakhar & Furedy, 1990, S. 28f.; Saxe & Cross, 1991, S. 226). Eine Vielzahl komplexer Daten kann in die Urteilsbildung einfließen (z. B. auch Aktenkenntnisse oder Bewertungen des Probandenverhaltens vor und nach der Testung). Bereits anhand dieser zusätzlichen Informationen ist der Untersucher gegebenenfalls in der Lage, ohne Berücksichtigung der körperlichen Reaktionen Trefferquoten zu erzielen, die signifikant höher liegen als bei reinen Zufallsentscheidungen. Daher sollte man nach Furedy und Heslegrave (1991a, S. 183f., 1991b, S. 233ff.) die Treffsicherheit nicht in Relation zum Zufallsniveau bewerten, sondern unter dem Aspekt, welche Bedeutung den physiologischen Daten bei der Erhöhung der Trefferquoten zukommt. D. h., die Validität des KFT ist dahingehend zu überprüfen, inwiefern die Berücksichtigung der physiologischen

Aufzeichnungen neben den nicht-physiologischen Informationsquellen einen spezifischen zusätzlichen Beitrag zur Treffsicherheit des Verfahrens erbringt (vgl. „specific effect“, Ben-Shakhar & Furedy, 1990, S. 98ff.). Dieser Zugewinn an Validität („**inkrementelle Validität**“), den man anhand der polygraphischen Daten erzielen kann, wurde in der bisherigen psychophysiologischen Aussageforschung weitgehend vernachlässigt (Fiedler, 1999, S. 20f.). Schätzungen der inkrementellen Validität ermöglichen aber eine bessere Evaluation des diagnotischen Nutzens des KFT als die bislang vorliegenden Angaben zu den absoluten Trefferquoten. Denn es ist durchaus möglich, daß die polygraphischen Daten einen nur unwesentlichen Beitrag zur kriterienbezogenen Validität leisten.

Neben der Kritik an der diagnostischen Gültigkeit der Befunde werden auch Probleme in bezug auf die **Konstruktvalidität** thematisiert. Nach Furedy (1991, S. 245) wird der Begriff „Lügendetektion“ („lie detection“, „detection of deception“) gezielt irreführend verwendet. Der KFT erfaßt eher die Angst der Pbn anstatt deren Täuschungsversuche, und die Ursachen dieser Angst sind kaum spezifizierbar (vgl. Saxe, 1991, S. 225f.). Neben der Furcht der Schuldigen vor Entdeckung und Strafe kommen auch andere angstauslösende Faktoren in Betracht, wie z. B. die Befürchtungen unschuldiger Pbn hinsichtlich der negativen Konsequenzen, die sich aus einem irrtümlich positiven Befund ergeben können. Im Einzelfall kann man nicht eindeutig feststellen, ob die erhöhten relevanten Reaktionen durch die Befürchtung eines Täters, überführt zu werden, bedingt sind oder durch die Angst eines Unschuldigen vor einer Falschbezichtigung.

Der KFT bietet keine Möglichkeit der Überprüfung, ob auf die relevanten Fragen wahrheitsgemäß oder wahrheitswidrig geantwortet wurde, da die **Kontrollfragen keine Kontrollbedingung** im wissenschaftlichen Sinne darstellen (Furedy & Heslegrave, 1991a, S. 165; Lykken, 1979, S. 49). Die beiden Fragentypen unterscheiden sich nicht nur im Hinblick auf den Wahrheitsgehalt der Anworten, sondern auch hinsichtlich zusätzlicher Faktoren (z. B. emotionale Signifikanz oder Breite des Zeitrahmens, den sie umfassen; vgl. Furedy, 1996a, S. 101; Furedy, Davis & Gurevich, 1988, S. 684). Und selbst wenn man unterstellen würde, daß die Grundannahmen des Verfahrens zuträfen und der Wahrheitsgehalt der Antworten der ausschlaggebende Faktor für die Reaktionsunterschiede wäre, würde man entgegen der Kategorisierung („deception indicated“ vs. „no deception indicated“) eher die Lüge bei den Unschuldigen entdecken (Furedy, 1986, S. 687). Denn nur bei dieser Gruppe erwartet man eine entsprechende Variation des Wahrheitsgehalts. Schuldige hingegen sollen sowohl auf die relevanten als auch auf die Kontrollfragen lügen. Raskin und Podlesny (1979, S. 55) sowie Raskin und Kircher (1991, S. 216f.) lehnen diese Kritik als wissenschaftlich unbegründet ab. Beim KFT handle es sich nicht um ein psychologisches Experiment, so daß Zweifel an der Bedin-

gungskontrolle unangemessen seien. Die sog. „control questions“ würden keine Kontrollbedingung darstellen, sondern lediglich eine Situation schaffen, auf die Unschuldige in Relation zu den relevanten Fragen mit erhöhter Besorgnis reagieren sollen. Darum sei die mißverständliche Bezeichnung „Kontrollfrage“ durch den angemesseneren Begriff „Vergleichsfrage“ zu ersetzen („comparison question“, Raskin & Kircher, 1991, S. 217). Die potentielle Aussagekraft der Testergebnisse bleibe von diesem Kritikpunkt zunächst unberührt. Tatsächlich ist die Konstruktvalidität weder eine notwendige noch hinreichende Bedingung für hohe Trefferquoten. Sie wird aber im Hinblick auf die theoretischen Grundlagen der psychophysiologischen Aussagebeurteilung relevant, wenn es um die Frage geht, welche psychologischen und physiologischen Prozesse für die erwarteten Reaktionsunterschiede verantwortlich sind.

Ein weiterer Aspekt der Konstruktvalidität wird in der weniger psychophysiologisch als vielmehr rechtspsychologisch orientierten Diskussion hierzulande häufig vernachlässigt. Er betrifft die **Rolle der physiologischen Ableitungen** als Maße für die Erregung des peripheren vegetativen Nervensystems und im weiteren Sinne als Indikatoren für die latenten psychologischen Variablen.

Bei der „Lügendetektion“ bestimmt man die **kardiovaskuläre Aktivität** durch den sog. „relativen Blutdruck“ („cardio“, Podlesny & Raskin, 1977, S. 789). Dazu wird ähnlich wie beim Riva-Rocci-Verfahren eine pneumatische Oberarmmanschette auf einen mittleren Druck zwischen den systolischen und den diastolischen Wert aufgepumpt (vgl. auch Schandry, 1998, S. 159f.). Außerdem erfassen manche Polygraphen zusätzlich photopletysmographisch die Veränderungen des peripheren Blutvolumens bzw. der Pulsvolumenamplitude im Finger. Zum einen ist die Messung des „relativen Blutdrucks“ umständlich und belastend für den Pb, da die Manschette den Blutstrom beeinträchtigt und deswegen nach relativ kurzer Zeit wieder gelockert werden muß. Zum anderen ist sie wenig valide. Selbst Befürworter des KFT (z. B. Matte, 1996, S. 383) räumen ein, daß der „relative Blutdruck“ keine direkte kontinuierliche Blutdruckmessung gestattet. Er spiegelt wohl in komplexer Weise Druck- und Volumenschwankungen im Oberarm wider (vgl. auch Schandry, 1998, S. 159). Wenn der Blutdruck reell ansteigt, hängt die Reaktionsrichtung der Pulsamplitude im „Cardio“-Kanal (Zu- oder Abnahme) davon ab, ob der Manschetteninnendruck oberhalb oder unterhalb eines Wertes liegt, der in etwa dem mittleren arteriellen Blutdruck entspricht (Geddes & Newberg, 1977, S. 199ff.). Ferner ermöglichen die im „Cardio“-Kanal bzw. Photoplethysmogramm abgebildeten Pulse keine genauen Angaben zur phasischen Herzschlagfrequenz. Demnach wird die kardiovaskuläre Aktivität nicht gemäß dem aktuellen Stand der physiologischen Meßtechnik bestimmt, d. h. die Herzschlagfrequenz per Elektrokardiogramm (vgl. Vossel & Zimmer, 1998, S. 67ff.) und die

kontinuierlichen arteriellen Blutdruckschwankungen nach dem Peñaz-Prinzip (z. B. FINAPRES-System, vgl. Rüddel & Curio, 1991).

Viele der handelsüblichen Polygraphen zur „Lügendetektion“ messen die **elektrodermale Aktivität** (**EDA**) über den Hautwiderstand anstatt über die Hautleitfähigkeit, wobei meist auch keine nicht-polarisierbaren Ag/AgCl-Elektroden und keine isotonische Elektrolytpaste verwendet werden (vgl. Matte, 1996, S. 172f.). Letztere Ableitungsform nach dem **Konstantspannungsprinzip** entspricht aber den gegenwärtigen psychophysiologischen Standards (Fowles et al., 1981, S. 238f.; Lykken & Venables, 1971, S. 671). Außerdem korreliert die Hautleitfähigkeit – nicht der Widerstand – linear mit der zugrundeliegenden ekkrinen Schweißdrüsenaktivität und deren Innervation durch das sympathische Nervensystem (Schandry, 1998, S. 187f.; Vossel & Zimmer, 1998, S. 50f.). Darüber hinaus konnten Velden und Vossel (1985, S. 294) zeigen, daß in psychophysiologischen Habituations- und Streßexperimenten die Reaktionsamplituden der Hautleitfähigkeit statt des Hautwiderstands in einer direkten Beziehung zu den latenten psychologischen Variablen (z. B. Stärke der Orientierungs- bzw. Defensivreaktionen und deren Habituation bzw. Sensitivierung) stehen (vgl. auch Velden, 1994, S. 59ff.). Dieser Zusammenhang ist insofern von Belang, als die entsprechenden Konzepte (v. a. Orientierungsreaktion und Habituation) mitunter zur theoretischen Erklärung der psychophysiologischen Aussagebeurteilung herangezogen werden (vgl. Abschnitt 2.10).

Ethische Kritikpunkte:

Eine umfassende Erörterung der moralischen oder juristischen Probleme der psychophysiologischen Aussagebeurteilung würde den Rahmen der vorliegenden Arbeit sprengen. Dennoch sollen im folgenden einige **ethische Bedenken** gegenüber dem KFT Erwähnung finden, zumal sie auch in einem engen Zusammenhang mit methodischen Aspekten stehen.

Die Ausführungen zur kriterienbezogenen Validität haben gezeigt, daß der KFT im erhöhten Maße anfällig ist gegenüber falsch positiven Befunden. Diese Verzerrung der **Fehlerrate zuungunsten Unschuldiger** widerspricht einerseits dem Rechtsgrundsatz „in dubio pro reo“ (Rill & Vossel, 1998, S. 486). Andererseits ergibt sich daraus die **Gefahr falscher Geständnisse**. Ein kontrovers diskutierter Nebeneffekt des KFT liegt darin begründet, daß viele Beschuldigte die Tat eingestehen, wenn ein positiver Befund resultiert (vgl. Lykken, 1991b, S. 217; Salzgeber et al., 1997, S. 219f.). Zum standardmäßigen Ablauf der Begutachtung gehört das Nachtest-Interview, das dem eigentlichen Test folgt, sofern der Pb als „unglaubwürdig“ klassifiziert wurde (vgl. Abschnitt 2.4.1). Insbesondere die nordamerikanischen Ermittlungsbehörden nutzen das Nachtest-Inter-

view als Verhör, das bis zu mehrere Stunden dauern kann und im wesentlichen dazu dient, den Beschuldigten zu einem Geständnis zu bewegen (Furedy, 1996a, S. 99). Man kann annehmen, daß solche Geständnisse zu einem nicht spezifizierbaren Kontingent falsch sind, weil Unschuldige mit einer relativ hohen Wahrscheinlichkeit Fehlklassifikationen erleiden und danach einem starken psychologischen Druck und emotionalem Streß ausgesetzt sind (Furedy, 1993, S. 264; Furedy & Liss, 1986, S. 109ff.; Furedy & Heslegrave, 1991a, S. 158f.; Lykken, 1998, S. 235ff.). Das Auftreten unzutreffender Selbstbezichtigungen wird durch die rechtliche Situation in den USA forciert, da dort die Option des „plea bargaining“ besteht („Aushandeln des Plädoyers“, vgl. Berning, 1992, S. 56). Der Angeklagte kann durch ein (Teil-)Schuldbekenntnis ein geringes Strafmaß und die Einstellung weitergehender Anklagen seitens der Staatsanwaltschaft erzielen (siehe auch Steller, 1987, S. 29). In diesem Zusammenhang ist zu berücksichtigen, daß die Geständnisse im Gegensatz zu den Ergebnissen der psychophysiologischen Aussagebeurteilung für gewöhnlich als Beweismittel zulässig sind (Faigman et al., 1997, S. 563). Folglich können die Befunde polygraphischer Untersuchungen indirekt einen gravierenden Einfluß auf die gerichtlichen Entscheidungen ausüben (Furedy, 1996a, S. 99). Aus der Tatsache, daß die Anwendung des KFT eventuell falsche Geständnisse provoziert, ergeben sich neben den ethischen und rechtlichen Bedenken auch zusätzliche Probleme für die Überprüfung der Validität. Zur Abschätzung der Trefferquoten werden in Felduntersuchungen häufig Geständnisse und eventuell dadurch beeinflußte Gerichts- bzw. Expertenurteile als Kriterium für die objektive Wahrheit („ground truth“, Lykken, 1991b, S. 217) herangezogen. D. h., sie dienen als Anhaltspunkt für die tatsächliche Schuld oder Unschuld des Pb. Und da diese Kriterien nicht unabhängig vom Ergebnis des KFT sind, droht eine Überschätzung der Trefferquoten (zur Problematik von Validitätsuntersuchungen der psychophysiologischen Aussageforschung und der eingeschränkten Eignung von Validitätskriterien bei Feldstudien siehe Abschnitt 2.9).

Neben den ethischen Einwänden, die sich auf die falschen Geständnisse und die hohe Rate irrtümlich positiver Befunde beziehen, übt Furedy (1993, S. 266) weitere Kritik, die er als „**Polygrapher’s Dilemma**“ bezeichnet. Der KFT versetzt den Untersucher unausweichlich in einen Konflikt gegenüber unschuldigen Pbn. Die Formulierung der Kontrollfragen liegt weitgehend in seinem Ermessensspielraum. Fällt deren emotionaler Gehalt zu gering aus, sind schwache Vergleichsreaktionen zu erwarten und das Risiko für falsch positive Befunde steigt. Sind die Kontrollfragen hingegen extrem bedrohlich, sollten Unschuldige zwar tendenziell stärker darauf reagieren und eher korrekt klassifiziert werden. Man muß aber in Kauf nehmen, daß sie psychische Beeinträchtigungen davontragen, da sie durch die Kontrollfragen und die Vortest-Prozedur in starke Bedrängnis gebracht werden. Außerdem läßt man die Pbn in dem Glauben, einer Person,

die das Vergehen der Kontrollfrage begangen habe (z. B. Masturbation), würde man auch das Verbrechen der relevanten Frage zutrauen (z. B. sexueller Kindesmißbrauch). In der Regel erfolgt nach der Untersuchung keine psychologische Nachsorge oder Aufklärung. Aber selbst bei einer moderat ausgeprägten Bedrohlichkeit der Kontrollfragen ist mit negativen Konsequenzen zu rechnen. Die Wahrscheinlichkeit „unentscheidbarer" Ergebnisse wird durch ein ausgeglichenes Reaktionsverhältnis erhöht, so daß – anstelle der intendierten Entlastung – der Verdacht weiterhin bestehen bleibt. Unabhängig vom Testausgang können gravierende Nachteile für unschuldige Pbn resultieren, die jeweils als ethisch bedenklich einzustufen sind. Darüber hinaus muß man berücksichtigen, daß eine Untersuchung mit dem KFT auch für viele Unschuldige einen extremen Stressor darstellt (Lykken, 1988, S. 112). Für eine ausführlichere Diskussion des „Polygrapher's Dilemma" sei auf Honts et al. (1995, S. 202f.) sowie Furedy (1996b, S. 54) verwiesen.

Das eventuelle Vorliegen einer sog. „**testimmanenten Täuschung**" (Achenbach, 1984, S. 352; Berning, 1992, S. 201f.; Delvo, 1981, S. 339) konfligiert unter juristischen Gesichtspunkten potentiell mit dem § 136 a StPO Abs. 1 Satz 1 (Verbot, den Beschuldigten zu täuschen). Der Pb wird während der Vortest-Phase hinsichtlich der Logik (Lügen und starke Reaktionen auf Kontrollfragen führten zu positiven Befunden) und Treffsicherheit (Postulat annähernder Perfektion) des KFT gezielt irregeführt (siehe auch Furedy, 1991, S. 243; Lykken, 1988, S. 112f.). Fiedler (1999) hält die entsprechenden Suggestionen für „ethisch bedenklich" und „mit den ethischen Normen der Psychologie nicht vereinbar" (S. 25). Nach Ansicht von Berning (1992) stellen „diese Maßnahmen ... jedoch bei *gewissenhafter und fachgerechter Anwendung* des Testverfahrens keine Täuschung dar" (S. 186). Unter der Voraussetzung einer gesetzeskonformen Durchführung sei diese auch mit den „ethischen Prinzipien der (deutschen) Psychologen" (S. 187) kompatibel. Der BGH (2000, S. 317f.) hat sich letzterer Auffassung angeschlossen und die Irreführung des Pb lediglich als geringfügig eingestuft. Sofern die Stimulationstests nicht fingiert sind (z. B. mit gezinkten Karten), was nicht zwingend erforderlich ist und zumindest in Deutschland wohl auch kaum praktiziert wird (vgl. Undeutsch & Klein, 1999, S. 68), liegt demnach keine Verletzung des § 136 a StPO vor. Ungeachtet der juristischen und ethischen Argumente tritt in dieser Diskussion wiederum ein zentrales methodisches Manko des KFT zutage, das dessen gesamtes theoretisches Rational in Frage stellt. Unabhängig davon, ob man sie als schwerwiegende Täuschungen oder minimale Irreführungen definiert, das Gelingen der „testimmanenten" Suggestionen ist im Einzelfall nicht sicherzustellen. Die Entscheidung, bei welchen Pbn die Manipulationen erfolgreich waren, bleibt der subjektiven Würdigung des Untersuchers überlassen und entzieht sich einer objektiven Überprüfung (Saxe, 1991, S. 226f.).

2.5 Directed Lie Test (DLT)

Sowohl im strafrechtlichen Kontext (vgl. Honts & Raskin, 1988, S. 57) als auch im beruflichen Umfeld (siehe Beardsley, 1999, S. 12) gewinnt ein **alternatives direktes Verfahren** zur psychophysiologischen Aussagebeurteilung zunehmend an Bedeutung, nämlich der sog. „**Directed Lie Test** (**DLT**)“ (vgl. Honts et al., 1995, S. 205; Lykken, 1998, S. 137; Synonyme: „directed lie control question technique“, Honts & Raskin, 1988, S. 57; „directed-lie control test“, Honts, 1994, S. 79). Diese Modifikation des Kontrollfragentests wird in deutschsprachigen Publikationen auch als „instruierte bzw. gerichtete Lügen-Kontrollfragen-Technik“ (Steller & Dahle, 1997, S. 314) oder „direktiver Lügentest“ (Fiedler, 1999, S. 11) bezeichnet. Sie soll folgende **Kritikpunkte am konventionellen Verfahren** vermeiden (vgl. Honts, 1994, S. 79; Honts & Raskin, 1988, S. 56; Horowitz, Kircher, Honts & Raskin, 1997, S. 109; Raskin et al., 1988, S. 8):

1. Die Durchführung des KFT ist komplex und wenig objektiv. Das Vortest-Interview und die Formulierung der Kontrollfragen sind **nicht standardisierbar**. Sie müssen individuell an den Pb und das jeweilige Delikt angepaßt werden. Die Wirkung der Vortest-Manipulationen hängt im hohen Maße von subjektiven Faktoren des Pb (z. B. Suggestibilität) und des Untersuchers (z. B. psychologische Kompetenz) ab.

2. Die intendierte **wahrheitswidrige Beantwortung** der Kontrollfragen bzw. die Zweifel am Wahrheitsgehalt sind nicht sicherzustellen (darum auch die Bezeichnung „probable lie comparison/control questions“, vgl. Horowitz et al., 1997, S. 108; Reed, 1994). Insbesondere wenn der Pb während des Vortest-Interviews Zugeständnisse macht und die Kontrollfragen jeweils unter Ausschluß der eingestandenen Vergehen revidiert werden, kann eine Situation resultieren, in der er von der Aufrichtigkeit seiner Antworten subjektiv überzeugt ist (Lykken, 1978, S. 139). Dadurch wird die besondere Signifikanz der Kontrollfragen für Unschuldige als eine wesentliche Voraussetzung des Verfahrens gefährdet, und das **Risiko für falsch positive Befunde** steigt.

3. Der Untersucher muß den Pb hinsichtlich der Logik des Verfahrens suggestiv **irreführen** (Konzessionen und starke Reaktionen auf die Kontrollfragen als Indiz für Unglaubwürdigkeit). Die vermeintliche Funktion der Kontrollfragen und ihre Bedeutung für das Testergebnis sind nur schwer zu vermitteln. Wenn die Person die eigentliche Testlogik durchschaut, droht ein Validitätsverlust.

4. Unter eher pragmatischen Gesichtspunkten wird beanstandet, daß manche Pbn die Kontrollfragen als äußerst **indiskret** beurteilen und deren Beantwortung verweigern.

Oder sie bekennen so viele Verfehlungen, daß der Untersucher die Kontrollfragen mehrfach ändern muß, wobei die Effekte der Neuformulierungen auf die körperlichen Reaktionen kaum abzuschätzen sind.

Beim DLT **instruiert** der Untersucher den Pb, bewußt auf die Kontrollfragen zu lügen (vgl. Beispiel von Raskin, 1989, S. 271). Ein wesentlicher Unterschied im Vergleich zum herkömmlichen KFT besteht also darin, daß der Wahrheitsgehalt der Antworten eindeutig feststeht. Die sog. „instruierten Lügen-Kontrollfragen" („directed lie control questions", Honts & Raskin, 1988, S. 57) sind sehr pauschal formuliert und thematisieren kleinere Vergehen, von denen anzunehmen ist, daß sie fast jeder irgendwann bereits begangen hat (z. B.: „Haben Sie schon einmal in Ihrem Leben gelogen?", „Haben Sie jemals gegen eine Regel oder Vorschrift verstoßen?" oder „Haben Sie jemals einen Fehler gemacht?", vgl. Horowitz et al., 1997, S. 111). Der konkrete Wortlaut der Lügen-Kontrollfragen wird ebenfalls im Rahmen eines Vortest-Interviews mit dem Pb besprochen (zur Vorgehensweise siehe Honts, 1994, S. 79; Honts et al., 1995, S. 204). Im Gegensatz zum üblichen KFT führen jedoch die zu erwartenden Zugeständnisse nicht zu einer Umformulierung der Kontrollfragen. Der Pb wird vielmehr angewiesen, während der eigentlichen Testung darauf mit einer Verneinung zu antworten und sich dabei bewußt vor Augen zu führen, daß er lügt. Ferner soll er an konkrete Situationen denken, in denen er die angesprochenen Taten verübt hat und auf seine emotionalen Reaktionen achten. Der Untersucher rechtfertigt diese Instruktion mit dem Argument, er müsse die physiologischen Reaktionen des Pb beim Lügen genauer analysieren, um einen angemessenen Vergleichsstandard für die relevanten Reaktionen zu erhalten. Adäquate körperliche Begleiterscheinungen beim wahrheitswidrigen Beantworten der Lügen-Kontrollfragen würden das Auftreten eindeutiger Ergebnisse begünstigen, wohingegen ein unangemessenes Reaktionsmuster eher zu einem nicht interpretierbaren Ergebnis führe. Über die genaue Bedeutung von „angemessenen" bzw. „unangemessenen Reaktionen" wird der Pb entweder im Unklaren gelassen (vgl. Honts, 1994, S. 79), oder man legt ihm explizit nahe, daß schwache Reaktionen auf die Kontrollfragen einen unentscheidbaren Befund nach sich ziehen können (vgl. Honts & Raskin, 1988, S. 57).

Im Unterschied zum KFT wird der **Stimulationstest** zu Beginn des Vortest-Interviews, d. h. vor der Besprechung der Testfragen, durchgeführt (vgl. Honts & Raskin, 1988, S. 58). Neben der Demonstration der Treffsicherheit erfüllen diese Karten- bzw. Zahlentests beim DLT einen zusätzlichen Zweck. Sie dienen auch zur Rechtfertigung der Instruktion und der Lügen-Kontrollfragen. Die Stimulationstests werden damit begründet, man müsse überprüfen, welche körperlichen Veränderungen der Pb beim Lügen zeige und ob er sich überhaupt für die Untersuchung eigne. Auf diese Weise wird die Testperson irregeführt. Denn unabhängig von den physiologischen Aufzeichnungen

erhält sie die Rückmeldung, sie habe bei den wahrheitswidrigen Antworten deutlich stärker reagiert. Anschließend suggeriert der Untersucher dem Pb, daß er auch während der eigentlichen Untersuchung starke Reaktionen auf die Lügen-Kontrollfragen zeigen müsse, um einen unentscheidbaren oder gar falsch positiven Befund zu vermeiden.

Die **Prämissen des DLT** und seine Auswertung ähneln im Prinzip dem konventionellen Verfahren (vgl. Honts & Raskin, 1988, S. 57; Horowitz et al., 1997, S. 109). Man geht davon aus, daß sich die Pbn besonders intensiv mit jenen Fragen auseinandersetzen, die ihrer Ansicht nach die Chancen, den Test mit einem negativen Befund zu bestehen, am meisten bedrohen. Unschuldige sollen ihre Aufmerksamkeit verstärkt auf die Kontrollfragen richten und diese als besonders bedeutsam bewerten, da sie befürchten müssen, nicht „angemessen“ darauf zu reagieren. Der Signalcharakter dieser Fragen werde durch die Instruktion und die Angst vor einem unentscheidbaren oder gar falsch positiven Testergebnis erhöht. Für die Täter hingegen dürften die relevanten Fragen eine größere Bedrohung darstellen, da sie den konkreten Tatvorwurf wahrheitswidrig verneinen müssen und ihnen gemäß der Instruktion nun die Entdeckung dieser Lüge bevorsteht.

Der DLT bietet aus theoretischer und praktischer Sicht mehrere potentielle **Vorteile** gegenüber dem KFT. Die Testlogik (angemessene bzw. starke Reaktionen auf Kontrollfragen führen zu einem negativen Befund) wird weitgehend korrekt vermittelt (Steller & Dahle, 1997, S. 315). Die Notwendigkeit einer personenspezifischen Formulierung der Kontrollfragen entfällt, wodurch auch die Intimsphäre des Pb besser gewahrt bleibt (Horowitz et al., 1997, S. 109). Außerdem ist keine inhaltliche Parallelität zwischen den relevanten Fragen und Kontrollfragen mehr erforderlich (Honts, 1994, S. 79f). Die gesteigerte Bedeutsamkeit der Lügen-Kontrollfragen für Unschuldige soll v. a. auf die Instruktion zurückzuführen sein und weniger von Frageninhalten, Manipulationen oder subjektiven Faktoren des Untersuchers bzw. Pb abhängen (Honts & Raskin, 1988, S. 57; Horowitz et al., 1997, S. 114). Die Durchführung ist einfacher und besser standardisierbar. Im Prinzip kann man sowohl die Lügen-Kontrollfragen als auch die Instruktion allen zu testenden Pbn in gleicher Weise darbieten. Nach Ansicht seiner Befürworter ist der Grad an Standardisierung derart hoch, daß eine automatische, maschinelle Durchführung in Betracht kommt (Honts, 1994, S. 80). Ferner geht man davon aus, daß die Lügen-Kontrollfragen v. a. für Unschuldige einen adäquateren Vergleichsreiz darstellen als die konventionellen Kontrollfragen. Dadurch soll das Risiko falsch positiver Befunde sinken.

Die wenigen bisher publizierten **Validitätsstudien** zum DLT sprechen zugunsten dieser Annahme. Honts und Raskin (1988) führten eine **Feldstudie** durch, indem sie die Untersuchungen von 25 Kriminalfällen reanalysierten. Als Validitätskriterien dienten Ge-

ständnisse bzw. eine eindeutige Beweislage („incontrovertible physical evidence“, Honts & Raskin, 1988, S. 57). Die verwendeten Befragungstechniken waren ähnlich dem KFT konstruiert. Man hatte lediglich eine der drei üblichen Vergleichsfragen durch eine instruierte Lügen-Kontrollfrage ersetzt und mit den entsprechenden Anweisungen dargeboten. Die numerische Auswertung der Tests erfolgte jeweils sowohl unter Beachtung als auch Vernachlässigung der Lügen-Kontrollfrage. Durch deren Einbeziehung stieg die Spezifität von 80% auf 100%. Die Sensitivität lag jedoch mit 92% unterhalb des Niveaus von 100% der Auswertung, die nur die herkömmlichen Kontrollfragen berücksichtigte (Trefferquoten nach Ausschluß unentscheidbarer Befunde). In einem **Scheinverbrechen-Experiment** verglichen Horowitz et al. (1997) die Relevant-Irrelevant-Technik, den konventionellen KFT und zwei DLT-Versionen miteinander. Eine Version beinhaltete persönliche Lügen-Kontrollfragen (z. B.: „Have you ever told a lie?“), die andere triviale Lügen-Kontrollfragen (z. B.: „Is Hawaii an island?“, vgl. Horowitz et al., 1997, S. 111). Der DLT mit persönlichen Lügen-Kontrollfragen übertraf den KFT sowohl hinsichtlich der Sensitivität (84% vs. 73%) als auch der Spezifität (87% vs. 86%; jeweils unter Vernachlässigung unentscheidbarer Befunde). Allerdings wird die Aussagekraft beider Untersuchungen von Lykken (1998, S. 139f., S. 317) stark angezweifelt. Darüber hinaus erbrachte die **unveröffentlichte Analogstudie** von Barland (1981, zitiert nach OTA, 1983, S. 76) zum Einsatz des DLT im Bereich der Spionageabwehr sowohl für glaubwürdige als auch unglaubwürdige Pbn relativ niedrige Trefferquoten. Dieses Ergebnis stellt die Effizienz der instruierten Lügen-Kontrollfragen-Technik insbesondere zur Personalselektion und regelmäßigen Sicherheitsüberprüfung („preemployment and periodic screenings“) erheblich in Frage (vgl. OTA, 1983, S. 76f.).

Ungeachtet der verbesserten Standardisierung bleiben auch bei der Verwendung von Lügen-Kontrollfragen einige **Probleme** ungelöst (vgl. auch Steller, 1997, S. 102). Auf eine suggestive Vortestmanipulation kann nicht vollständig verzichtet werden. Dabei wird dem Pb der (irreführende) Eindruck vermittelt, die Kontrollfragen würden einen Anhaltspunkt für seine typische Reaktionsweise beim Lügen liefern. Die Erfolge dieser Suggestionen sind aber ebenso wenig überprüfbar wie beim KFT. Denkbar ist, daß trotz der Instruktion viele Unschuldige die relevanten Fragen als bedeutsamer erachten. Damit wäre das Risiko für falsch positive Klassifikationen weiterhin auf einem hohen Niveau anzusiedeln. Andererseits könnten auch Täter die Relevanz der Lügen-Kontrollfragen überbewerten, was mit einem Anstieg falsch negativer Befunde einhergehen würde. Darüber hinaus sind nicht alle potentiellen Lügen-Kontrollfragen in jedem beliebigen Einzelfall applizierbar. Falls beim Pb die subjektive Gewißheit besteht, daß er die jeweiligen Vergehen noch nie begangen hat, resultieren Probleme mit der Testlogik, und die entsprechende Lügen-Kontrollfrage muß durch eine andere ersetzt werden.

Insgesamt hält Lykken (1998, S. 137ff.) die Annahmen des DLT für völlig unplausibel. Er bezweifelt, daß Unschuldige beim instruierten Leugnen trivialer menschlicher Verfehlungen konsistent stärker reagieren als auf Fragen nach einem konkreten Delikt, dessen man sie irrtümlicherweise beschuldigt. Somit stellen auch die Lügen-Kontrollfragen keine angemessene Kontrollbedingung für die relevanten Fragen dar. Außerdem bleibt unklar, inwiefern die per Anweisung induzierten Falschantworten vergleichbar sind mit intentionalen Täuschungen (Iacono, 2000, S. 781). Im Vergleich zum KFT ist jedoch festzustellen, daß der DLT zumindest im Hinblick auf die Objektivität und Standardisierung eher den Erfordernissen eines psychologischen Testverfahrens gerecht wird (vgl. auch Furedy, 1996b, S. 58).

2.6 Truth Control Test (TCT)

Sowohl am KFT als auch am DLT wird kritisiert, daß die verwendeten Vergleichsfragen nicht als adäquate Kontrollbedingung für die relevanten Fragen fungieren können. Einen alternativen Ansatz zur Vermeidung dieses Problems bietet der „**Truth Control Test** (**TCT**)“ (Lykken, 1998, S. 143). Der TCT geht auf den Vorschlag von Reid (1947) zurück, beim KFT zusätzlich sog. „guilt complex“-Fragen einzuführen: „The ‘guilt complex’ question is based upon an entirely fictitious crime of the same type as the actual crime under investigation, but one which is made to appear very realistic to the subject“ (S. 545).

Mit dem TCT wird der Pb im Hinblick auf **zwei ähnliche Vergehen getestet** (für ein konkretes Beispiel siehe Lykken, 1998, S. 145). Dabei handelt es zum einen um ein reales Verbrechen (z. B. Mord an Herrn X) und zum anderen um eine fiktive Straftat (z. B. Mord an Herrn Y). Während des Vortest-Interviews muß der Untersucher dem betreffenden Pb glaubhaft suggerieren, daß das fingierte Verbrechen ebenfalls Gegenstand der Ermittlungen ist und daß der Pb hinsichtlich beider Delikte unter Tatverdacht steht. Analog zum KFT werden relevante Fragen dargeboten, die direkt auf das reale Verbrechen abzielen (z. B.: „Haben Sie Herrn X ermordet?“). Den relevanten Fragen stellt man jedoch nicht die gängigen Kontrollfragen gegenüber (z. B.: „Haben Sie vor Ihrem 20. Lebensjahr jemals einen anderen Menschen absichtlich verletzt?“), sondern inhaltlich parallelisierte Vergleichsfragen, die sich auf das fiktive Verbrechen beziehen (z. B.: „Haben Sie Herrn Y ermordet?“). Diese „guilt complex“-Fragen werden in Abgrenzung von den üblichen Kontrollfragen („lie control“) auch als „truth control“ bzw. „known-truth questions“ bezeichnet (vgl. Lykken, 1998, S. 144f.), da die Verneinung des Tatvorwurfs stets eine wahrheitsgemäße Beantwortung impliziert. Vorausgesetzt die Vortest-Suggestionen gelingen und der Pb ist wirklich davon überzeugt, daß die

erfundenen Anschuldigungen ernst zu nehmen sind, kann man davon ausgehen, daß die „truth control“-Fragen eine Kontrollbedingung im wissenschaftlichen Sinne darstellen (Ben-Shakhar & Furedy, 1990, S. 24; Lykken, 1998, S. 31). Die dadurch ausgelösten Reaktionen spiegeln wider, wie eine Person auf den Vorwurf einer Tat reagiert, die sie nicht begangen hat und die sie folglich glaubhaft abstreitet.

Gemäß den **Annahmen** des Verfahrens sollen jene Pbn, die bezüglich beider Vergehen unschuldig sind, die tatsächliche und die erfundene Anschuldigung als gleichermaßen bedrohlich bewerten. Konträr dazu sollen die Täter durch die relevanten Fragen in stärkere Bedrängnis geraten, da sie das reale Verbrechen wahrheitswidrig abstreiten müssen, während sie die „truth control“-Fragen wahrheitsgemäß verneinen. Dementsprechend werden die Aufzeichnungen ausgewertet. Ein Pb, der auf die beiden Fragentypen in etwa gleich stark bzw. stärker auf die Vergleichsstimuli reagiert, wird als „glaubwürdig“ klassifiziert. Wenn sich deutlich stärkere Reaktionen auf die relevanten Fragen zeigen, resultiert ein positiver Befund.

Laut Lykken (1998, S. 144) existiert **kein angewandtes Verfahren** der forensischen Psychophysiologie, das strikt nach den Prinzipien des TCT konzipiert ist. Darüber hinaus gibt es **weder Feld- noch Analogstudien** zu dieser Testmethode. Auch der Einsatz von „guilt complex“-Fragen im KFT (vgl. dazu Reid & Inbau, 1977, S. 48ff.) ist nur eine grobe Annäherung an den TCT, da die Untersucher keine besonderen Anstrengungen unternehmen, die Pbn von der Ernsthaftigkeit der fingierten Anschuldigung zu überzeugen. Von dieser Kritik sind auch die Analogstudien von Podlesny und Raskin (1978) sowie Bradley, MacLaren und Black (1996) betroffen. Podlesny und Raskin bauten in ihre Version des KFT eine „guilt complex“-Frage ein, und Bradley et al. verwendeten eine Befragungstechnik, die ähnlich dem TCT konzipiert war. Obwohl die Ergebnisse der Untersuchungen im Prinzip die Grundannahmen des Verfahren stützen (vgl. dazu auch Lykken, 1998, S. 146f.), sind sie dennoch wenig aussagekräftig. Es wurde nicht versucht, die Bedeutung der Vergleichsfragen zu unterstreichen bzw. die Pbn ernsthaft davon zu überzeugen, daß sich die Stimuli auf ein zweites (Schein-) Verbrechen bezogen, dessen man sie ebenfalls bezichtigen würde.

Die **Problematik des TCT** ist weniger theoretischer als vielmehr praktischer Natur. Die Schwierigkeit besteht nicht nur darin, die Pbn davon zu überzeugen, das erfundene Verbrechen sei real und sie stünden tatsächlich unter Tatverdacht. Vielmehr müssen die beiden Tatvorwürfe zumindest für Unschuldige äquivalent und *gleichermaßen bedrohlich* sein. In aller Regel wird jedoch bereits durch die Ermittlungsarbeit vor dem eigentlichen Test die Relevanz der echten Anschuldigung in den Vordergrund gestellt. D. h., deren Bedrohungspotential kann mittels eines erfundenen – wenngleich auch inhaltlich

parallelisierten – Verdachtsmoments kaum noch aufgewogen werden (vgl. Ben-Shakhar & Furedy, 1990, S. 138). Lykken (1998, S. 144ff.) konstruiert zwar ein Anwendungsbeispiel für den TCT und vertritt die Ansicht, daß in manchen Kriminalfällen und unter bestimmten Voraussetzungen durchaus die Möglichkeit besteht, die Pbn erfolgreich zu täuschen. Gleichzeitig äußert er ethische Bedenken gegen diese Form der Irreführung. Man muß nämlich sicherstellen, daß der Pb über kein handfestes Alibi für das fiktive Verbrechen verfügt. Darüber hinaus darf man seinem Verteidiger keine Gelegenheit bieten, die Rechtmäßigkeit des Tatvorwurfs zu überprüfen. Wenn es sich außerdem um eine Art von Verbrechen handelt, von dem der Verdächtige normalerweise Kenntnis erlangt haben sollte (z. B. über die Medien), muß der Untersucher Gründe erfinden, warum der Pb nichts davon erfahren hat (vgl. Bradley, MacLaren & Black, 1996, S. 761). Ähnlich wie beim KFT oder DLT hängt auch das Funktionieren des TCT davon ab, daß die Pbn die Testlogik nicht kennen bzw. durchschauen.

Insgesamt stellen die o. g. methodischen Schwierigkeiten und ethischen Einwände die wesentlichen Gründe dar, weshalb sich der TCT in der Praxis nicht durchgesetzt hat (vgl. Lykken, 1998, S. 147). Zwar berichten Ben-Shakhar und Furedy (1990, S. 119, S. 139), daß japanische Polygraphers eine direkte Befragungstechnik favorisieren, die ebenfalls „guilt complex“-Fragen enthält. Aber aufgrund der gravierenden Anwendungsprobleme sei deren Einsatz dort relativ selten, so daß auch von dieser Seite kaum mit systematischer Validitätsforschung zu rechnen ist. Somit bleibt neben der Praktikabilität auch die Treffsicherheit des Verfahrens fraglich.

2.7 Tatwissentest (TWT)

2.7.1 Vorgehensweise

Der „**Tatwissentest** (**TWT**)“ („guilty knowledge test“, Lykken, 1974, S. 726; auch: „concealed knowledge technique“, Undeutsch, 1983, S. 410) ist das wichtigste **indirekte Verfahren** der forensischen Psychophysiologie. Im Gegensatz zu den direkten Verfahren (z. B. Kontrollfragentest) zielt diese Befragungstechnik nicht unmittelbar auf die Glaubwürdigkeit bzw. Täterschaft des Pb ab, sondern auf dessen Tatwissen. D. h., es soll festgestellt werden, inwiefern ein Beschuldigter Kenntnisse über bestimmte Details des Verbrechens besitzt, von denen Tatunbeteiligte nichts wissen können.

Bei vielen Delikten liegen spezifische Tatumstände vor, die neben den Ermittlungsbehörden nur Tatbeteiligte („Schuldige“ im Sinne von Tätern oder Mitwissern) kennen (vgl. Lykken, 1991b, S. 218). Diese Details werden als sog. **relevante Items** in eine

Reihe gleichartiger Alternativen (**irrelevante Items**) eingebettet und in Form von Multiple-Choice-Fragen dargeboten. Die irrelevanten Items stehen in keinem direkten Zusammenhang mit dem untersuchten Verbrechen. Sie thematisieren aber ähnliche Sachverhalte, die zumindest für Tatunbeteiligte („Unschuldige“) gleichermaßen plausibel sein sollen wie die kritischen Items.

Ein TWT besteht aus mehreren **Multiple-Choice-Fragen**, die sich jeweils auf unterschiedliche Tatdetails beziehen (vgl. Tabelle 3). Lykken (1991b, S. 219) schlägt beispielsweise zehn solcher Fragen vor. Die konkrete Anzahl ist im Endeffekt von den Gegebenheiten des untersuchten Falles abhängig. Zu jeder Frage werden diverse Antwortalternativen formuliert (z. B. ein relevantes und fünf irrelevante Items). Bei deren Auswahl ist darauf zu achten, daß ein Tatbeteiligter auf alle Fälle Kenntnis von den zutreffenden Sachverhalten haben sollte. Man darf folglich keine allzu marginalen Einzelheiten abfragen. Gleichzeitig dürfen Unschuldige nichts davon wissen. Die Multiple-Choice-Fragen und Items werden dem Pb in der Regel verbal dargeboten. Alternative Präsentationsmodi sind jedoch ebenfalls denkbar, beispielsweise andere akustische Stimuli (z. B. Stimmproben tatbeteiligter und unbeteiligter Personen) oder eine schriftliche Darbietung der Items (vgl. Farwell & Donchin, 1991, S. 534) bzw. in Form von Bildern (z. B. unterschiedliche Varianten des Tatorts, vgl. Lykken, 1998, S. 295f.). Die Position des relevanten Items wechselt von Frage zu Frage. Da wegen der Neuheit der angesprochenen Thematik auf die erste Alternative einer Sequenz im allgemeinen die stärkste Reaktion zu erwarten ist, steht das relevante Item nie zu Beginn der Multiple-Choice-Frage. Bei der Standardprozedur wird der Pb aufgefordert, auf jedes Item mit einer einfachen Verneinung zu antworten. Eine Durchführung mit Bejahung (Kugelmass, Lieblich & Bergman, 1967, S. 313) bzw. Nachsprechen der Items (vgl. Balloun & Holmes, 1979, S. 318) oder ohne verbale Antworten (z. B. Gustafson & Orne, 1965b, S. 11) ist ebenso realisierbar. Auch beim TWT kann man zu Beginn der Untersuchung ein Vortest-Interview oder Verhör durchführen, um die bestehenden Tatkenntnisse des Pb zu eruieren. Details, die Unschuldigen bekannt sein können (z. B. aus den Medien, vom Hörensagen, durch vorherigen polizeiliche Vernehmungen bzw. den Anwalt), müssen unbedingt von der Befragung ausgeschlossen werden. Darüber hinaus will man anhand des Vortest-Interviews die Erinnerungen der Schuldigen an das Verbrechen reaktivieren. Dabei werden allenfalls die einzelnen Themengebiete des TWT erörtert, nicht aber der eigentliche Wortlaut der Fragen und Items besprochen (Raskin, 1989, S. 278).

Die **Grundannahme** des Verfahrens lautet, daß nur Pbn mit Tatkenntnissen die kritischen Sachverhalte identifizieren können (Lykken, 1974, S. 727f.). Durch das Wiedererkennen erlangen die relevanten Items für diese Personen eine besondere Bedeutung.

Tabelle 3. Beispiel für einen Tatwissentest mit sechs Multiple-Choice-Fragen und jeweils sechs Items (nach Steller, 1987, S. 11)

Nr.	Frage bzw. Item
	Welche Nummer hatte der Raum, in dem der Diebstahl begangen wurde? War es
1.	Raum 321?
2.	Raum 214?
3.	Raum 411? *
4.	Raum 206?
5.	Raum 129?
6.	Raum 217?
	Wo befand sich der Gegenstand, der gestohlen wurde? War er
7.	in dem Schrank?
8.	auf dem Bücherregal?
9.	in dem Aktenbock?
10.	auf der Fensterbank
11.	in der Schublade? *
12.	in dem Ablagekasten?
	Was für ein Gegenstand war es, der gestohlen wurde? War es
13.	ein Armband?
14.	eine Uhr? *
15.	eine Brosche?
16.	ein Ring?
17.	eine Kette?
18.	eine Geldbörse?
	Welche Farbe hatte das Armband der Uhr, die gestohlen wurde? War es
19.	rot?
20.	golden?
21.	braun?
22.	schwarz? *
23.	silbern?
24.	weiß?
	Welcher Buchstabe stand auf dem Briefumschlag, in dem sich die Uhr befand? War es
25.	F?
26.	A?
27.	Z?
28.	M?
29.	K?
30.	E? *
	Welche Farbe hatte der Briefumschlag, in dem die Uhr war? War er
31.	blau?
32.	grün?
33.	gelb? *
34.	grau?
35.	braun?
36.	rot?

Anmerkungen. * = relevante Items; alle Items werden verneint.

Infolge einer erhöhten Aufmerksamkeitszuwendung und gegebenenfalls evozierten Begleitemotionen sollen die physiologischen Reaktionen darauf stärker ausgeprägt sein. Im allgemeinen erwartet man von Tatbeteiligten ein derartiges Reaktionsprofil (erhöhte

relevante Reaktionen). Im Grunde genommen kann aber jeder **regelhafte Reaktionsunterschied** zwischen den beiden Itemtypen bzw. jede deutliche Abweichung von einer Zufallsverteilung der Reaktionsstärken (z. B. auch konsistent schwächere relevante Reaktionen) als Indiz für vorhandenes Tatwissen gelten (Lykken, 1959, S. 385). Und wenn ein solches systematisches Reaktionsmuster auftritt, resultiert ein positiver Befund („guilty knowledge indicated“, vgl. Elaad, 1990, S. 524). Mittels einer eventuellen Nachbefragung und anschließenden Ermittlungen wäre dann zu klären, worauf dieses Testergebnis beruht (z. B. Täter-, Mitwisser- oder Zeugenschaft). Für Personen ohne Tatwissen sollen die tatbezogenen Alternativen keine erhöhte Signifikanz aufweisen. Demzufolge zeigen sie ein unsystematisches Reaktionsmuster und reagieren nur vereinzelt, rein zufällig auf die relevanten Items stärker.

Bei den **physiologischen Messungen** des TWT kann man sich auf eine einzige Variable beschränken (Lykken, 1991b, S. 218). Abgesehen von einigen Experimenten, die alternative Parameter (mit-)erfaßten – wie etwa ereigniskorrelierte hirnelektrische Potentiale (z. B. Farwell & Donchin, 1991), kardiovaskuläre (Bradley & Ainsworth, 1984) bzw. respiratorische Maße (Elaad, Ginton & Jungman, 1992) oder den Pupillendurchmesser (Bradley & Janisse, 1981) – werden meist die Amplituden der elektrischen Hautleitfähigkeits- bzw. Hautwiderstandsreaktionen bestimmt. Lykken (1960, S. 261) hält die elektrodermale Aktivität für besonders geeignet, da sie sich als ein sehr sensitiver Indikator autonomer Erregung erwiesen hat.

Zur **Auswertung** schlug Lykken (1959, S. 386), der den TWT in der hier vorgestellten Form konzipierte, ein **Punktesystem** vor. Andere Forscher haben dieses sog. „Lykken-Scoring“ (nach Steller, 1987, S. 46) übernommen und teils modifiziert (z. B. Balloun & Holmes, 1979, S. 319; Bradley & Warfield, 1984, S. 687; Davidson, 1968, S. 63; Giesen & Rollison, 1980, S. 7). Für jede Multiple-Choice-Frage werden die Reaktionen auf die Items verglichen, wobei die erste Alternative pro Frage jeweils unberücksichtigt bleibt. Ist die Amplitude der elektrodermalen Reaktion auf das relevante Item die höchste innerhalb der Sequenz, vergibt man zwei Punkte. Handelt es sich um die zweitstärkste Reaktion, resultiert ein Punkt. Andernfalls wird die Frage mit null Punkten gewertet. Anschließend addiert man die Werte über alle Fragen. Wenn die Summe mehr als die Hälfte der im ganzen möglichen Gesamtpunktzahl beträgt (z. B. bei 10 Fragen mehr als 10 Punkte), wird auf das Vorliegen von Tatwissen geschlossen (positiver Befund). Niedrigere Werte (z. B. kleiner oder gleich 10 Punkte) führen zu einem negativen Befund. Das ursprüngliche Scoring von Lykken (1959) sieht keine „unentscheidbare“ Kategorie („inconclusive“) vor. Andere Auswertungsprozeduren hingegen lassen durchaus indifferente Ergebnisse zu (z. B. Elaad, 1990, S. 524; Lykken, 1991b, S. 221).

Es gibt Alternativen zum Punktesystem. Beispielsweise kann man zählen, wie oft der Pb bei einer bestimmten Anzahl von Fragen die stärksten Reaktionen auf die relevanten Items zeigt. Und es läßt sich die **Wahrscheinlichkeit** berechnen, daß eine Person ohne Tatwissen ein derartiges Ergebnis erzielt. So beträgt etwa das Risiko, daß ein Unschuldiger bei einer Multiple-Choice-Frage mit fünf auswertbaren Alternativen rein zufällig auf das relevante Item maximal reagiert 1/5. Bei 10 solcher Fragen sinkt sein Risiko für konsistent stärkere relevante Reaktionen gemäß dem Multiplikationssatz der Stochastik bereits auf $(1/5)^{10} = 1/9765625$. Auf der Basis zusätzlicher Annahmen, z. B. hinsichtlich der Wahrscheinlichkeit mit der ein Schuldiger auf das relevante Item einer Frage maximal reagiert, lassen sich Aussagen darüber treffen, mit welcher statistischen Sicherheit ein bestimmtes Testergebnis einen positiven oder negativen Befund indiziert (vgl. Rechenexempel von Lykken, 1991b, S. 221, 1998, S. 289f.).

Beide o. g. Auswertungsmethoden weisen ein gemeinsames Problem auf. Es werden grundsätzlich nur ausgeprägte relevante Reaktionen als Unterscheidungskriterium herangezogen. Wie bereits angedeutet, kann aber jegliche Abweichung von einer zufälligen Verteilung der Reaktionsstärken Tatwissen indizieren. Wenn etwa ein schuldiger Pb in der Lage ist, durch geeignete Gegenmaßnahmen („countermeasures“) seine Vergleichsreaktionen systematisch zu erhöhen, würde er konsistent schwächere Reaktionen auf die relevanten Items zeigen. Ein solches Reaktionsprofil könnte seine Tatbeteiligung ebenfalls offenlegen. Darum hat Lykken (1960, S. 259) ein spezielles Auswertungsverfahren vorgeschlagen, das auf **Rangbildung** basiert (vgl. auch „method of expected ranks“, Lykken, 1998, S. 301f.). Für jede Multiple-Choice-Frage werden die Reaktionen auf die Items ihrer Stärke nach in eine Rangreihe gebracht (unter Vernachlässigung der ersten Alternative). Anschließend zählt man über alle Fragen hinweg, wie oft die jeweilige relevante Reaktion auf der ersten, zweiten, dritten etc. Position liegt. Für einen Pb ohne Tatwissen würde man annehmen, daß die Häufigkeitsverteilung über die Ränge relativ homogen ist, d. h., die relevanten Reaktionen sollten zufallsbedingt jeden Rangplatz in etwa gleich oft einnehmen. Deutliche Abweichungen von dieser Gleichverteilung (normalerweise zugunsten der vorderen Ränge [1 bzw. 2], aber auch konsistent schwache relevante Reaktionen und somit eine Häufung der hinteren Ränge [3, 4 und 5]) gelten als Indiz für vorhandene Tatkenntnisse und eventuell praktizierte Manipulationsversuche. Die Prozedur setzt aber voraus, daß Schuldige mit ihren Gegenmaßnahmen keine homogene Verteilung der Reaktionsstärken erzielen können (Honts, Devitt, Winbush & Kircher, 1996, S. 85). Außerdem ist sie nur bei Tatwissentests mit hinreichend vielen Fragen sinnvoll einsetzbar. Lykken (1960) verwendete in seiner Laborstudie 25 Multiple-Choice-Fragen mit jeweils sechs Alternativen und konnte anhand der Rangauswertung trotz Gegenmaßnahmen alle Pbn korrekt diagnostizieren. Konträr dazu erwies sich die gleiche Auswertungsmethode im Scheinverbre-

chen-Experiment von Honts et al. (1996, S. 89f.) bei einem TWT mit fünf Fragen und sechs Items als relativ ineffektiv zur Entdeckung von Schuldigen und deren Manipulationsversuchen.

Der TWT wird bisweilen gleichgesetzt mit einem anderen indirekten Verfahren (z. B. Holstein, 1990, S. 157). Der sog. „**Peak of Tension Test** (**POT**)“ (vgl. Lykken, 1998, S. 147ff.) ist ähnlich wie der TWT aufgebaut. Doch seine Durchführung ist eher als eine Ergänzung zu den direkten Befragungstechniken gedacht (Reid & Inbau, 1977, S. 55). Er besteht nur aus einer einzigen Multiple-Choice-Frage (Lykken, 1975, S. 711), die mehrfach wiederholt wird (Reid & Inbau, 1977, S. 56). Die Items sind häufig in einer bestimmten Reihenfolge angeordnet (z. B. ansteigendes Kaliber einer Schußwaffe, vgl. Raskin, 1989, S. 275), bzw. der Pb ist über die Abfolge der Items informiert. Die relevante Alternative wird ungefähr in die Mitte der Sequenz positioniert (Matte, 1996, S. 497f.) Man erwartet, daß während der Befragung die körperliche Erregung (z. B. relativer Blutdruck) eines Tatbeteiligten zunächst antizipatorisch ansteigt, dann beim relevanten Item ein Maximum erreicht („peak of tension“) und anschließend wieder absinkt. Neben dieser Variante, bei der auch der Untersucher das relevante Item kennt („known-solution peak of tension test“), gibt es noch den sog. „searching peak of tension test“ (Raskin, 1989, S. 276). Damit will der Untersucher einen ihm unbekannten Sachverhalt herausfinden. Beispielsweise kann er bei einer Unterschlagung dem vermeintlichen Täter verschiedene Geldbeträge in aufsteigender Sequenz darbieten und anhand des „Erregungsgipfels“ Rückschlüsse auf die mutmaßlich entwendete Summe ziehen. Der POT wurde von Keeler (1933, zitiert nach Lykken, 1960, S. 260) vorgeschlagen und stellt gewissermaßen einen frühen Prototypen des TWT dar. Die beiden Verfahren unterscheiden sich jedoch hinsichtlich der Testkonstruktion und -logik.

Lykken (1960, S. 258ff.) grenzt die direkten und indirekten Verfahren der forensischen Psychophysiologie klar voneinander ab. Nur die direkten zählt er zu dem klassischen Ansatz der „Lügendetektion“ („lie detection“) im engeren Sinne, wohingegen der TWT und der POT eine „guilt detection“ gestatten sollen. Diese Feststellung trifft aber den Sachverhalt nur bedingt, da auch mit den indirekten Befragungstechniken **keine unmittelbare Aufdeckung von Schuld oder Täterschaft** möglich ist, sondern allenfalls von deliktrelevanten Kenntnissen. Insofern ist auch die Bezeichnung „guilty knowledge test“ nicht ganz korrekt, „denn ‘schuldig’ kann immer nur ein ‘auf freie, verantwortliche, sittliche Selbstbestimmung’ angelegter Mensch sein, niemals ein bloßes Wissen“ (Undeutsch, 1983, S. 402). Die deutsche Bezeichnung „Tatwissentest“ weist diesen sprachlichen Makel allerdings nicht auf (Steller, 1987, S. 10).

2.7.2 Kritik am Tatwissentest

Die indirekten Verfahren der forensischen Psychophysiologie werden im Vergleich zu den direkten nur selten angewandt. Ein wesentlicher Grund dafür ist ihr **eingeschränkter Einsatzbereich**. Hierbei handelt es sich um eines der Hauptargumente, das die wissenschaftlichen Befürworter des KFT (z. B. Raskin & Kircher, 1991, S. 219) gegen den TWT vorbringen. Auch die professionellen Polygraphers in Nordamerika mit ihrer Berufsvereinigung, der „American Polygraph Association", gehen davon aus, daß die von ihnen favorisierten Befragungstechniken keine derartigen Restriktionen mit sich bringen (vgl. Furedy, 1996a, S. 103). Aus diesem Grund stelle der TWT keine angemessene Alternative zum KFT dar. Nach Ansicht von Raskin (1989, S. 282) bedeutet aber die begrenzte Anwendbarkeit noch keinen grundsätzlichen Mangel an Validität. Unter bestimmten Voraussetzungen sei der TWT insbesondere als Ergänzung zum KFT oder DLT durchaus brauchbar. Kongruente Ergebnisse könnten sich gegenseitig stützen, wohingegen eine mangelnde Übereinstimmung den Beweiswert der Untersuchungen entkräften würde.

Die indirekten Verfahren ermöglichen **keine hinreichende Differenzierung zwischen Tätern und Unschuldigen mit Tatwissen**. Somit setzt eine zweckmäßige Durchführung des TWT voraus, daß unschuldige Pbn möglichst keine oder nur wenig Kenntnis über den Tathergang erlangt haben. Auch Lykken (1998, S. 300f.) propagiert die Anwendung als Hilfsmittel strafrechtlicher Ermittlungen zu einem frühen Zeitpunkt, wenn noch nicht zu viele Informationen an die Öffentlichkeit gedrungen sind. Darüber hinaus ist das Verfahren nur im Rahmen einer konkreten Verbrechensaufklärung sinnvoll einsetzbar (Ben-Shakhar, 1991a, S. 234). Psychophysiologische Aussagebeurteilungen mit allenfalls diffusen Verdachtsmomenten und ohne einen eindeutig definierten Tatbezug, wie sie häufig im beruflichen Umfeld stattfinden, scheiden somit aus. In der Regel steht auch die Glaubwürdigkeitsbegutachtung möglicher Zeugen oder eine Täterschaftsbeurteilung bei Verdacht auf sexuellen Mißbrauch mit dem TWT nicht zur Debatte (Raskin, 1989, S. 281), wobei diese speziellen Anwendungsbereiche auch beim KFT besonders umstritten sind (vgl. Abschnitt 2.4.2).

Weitere Schwierigkeiten bereitet die **Auswahl der Items**. Oft lassen sich nicht hinreichend viele prägnante Einzelheiten finden, die man verwerten kann. Die beiden einzigen dokumentierten Feldstudien zum TWT (Elaad, 1990, S. 522; Elaad et al., 1992, S. 759) deuten darauf hin, daß in Realfällen nur selten mehr als sechs kritische Details für die Testkonstruktion zur Verfügung stehen. Laut Podlesny (1995, zitiert nach Lykken, 1998, S. 305) ist der Nutzen des TWT für die Ermittlungsarbeit äußerst begrenzt. Er analysierte die Akten von 758 Kriminalfällen, für die beim Federal Bureau of

Investigation (FBI) eine psychophysiologische Aussagebeurteilung beantragt worden war. Seinem Ergebnis zufolge boten weniger als 9% der Fälle hinreichend viele relevante Items, um damit einen brauchbaren TWT zu konstruieren. Lykken (1998, S. 305f.) kontert, daß dessen optimale Durchführung eine adäquate Vorbereitung voraussetzt, die beim FBI in der Regel nicht gegeben ist. Wenn die Kriminalbeamten von Beginn der Ermittlungsarbeit an gezielt auf die Realisierung eines TWT hinarbeiten und nach geeigneten Items dafür suchen würden, dann wäre mit einer deutlich höheren Quote potentieller Anwendungen zu rechnen. Dementsprechend kommt der TWT in anderen Ländern wie Israel, Japan oder China, wo die Ermittlungen stärker darauf ausgerichtet sind, relativ häufig als polizeiliches Hilfsmittel zum Einsatz (vgl. Ben-Shakhar & Furedy, 1990, S. 120ff., Matte, 1996, S. 23f., S. 34).

Außerdem ist eine zu große Anzahl an Items nicht unbedingt zweckmäßig. Iacono, Boisvenu und Fleming (1984, S. 295f.) erhielten eine optimale Gesamttrefferquote, wenn nur die ersten fünf Multiple-Choice-Fragen des TWT gewertet wurden. Sobald man noch mehr Fragen berücksichtigte, blieb die Spezifität zwar auf einem hohen Niveau, die Sensitivität sank aber wieder. Mit allen zehn Fragen betrugen die Trefferquoten (unter Vernachlässigung unentscheidbarer Ergebnisse) 100% für unschuldige und 88% für schuldige Pbn. Darüber hinaus zeigte sich für Schuldige eine positive Korrelation zwischen den Testergebnissen (Punktwerte ähnlich dem Lykken-Scoring) und der Anzahl erinnerter Details. Falsch negative Befunde sind also möglicherweise darauf zurückzuführen, daß nicht alle relevanten Items von den Tätern identifiziert werden (Raskin, 1989, 281f.). D. h., man muß **Wahrnehmungs- und Gedächtniseffekte** berücksichtigen. Die Logik des TWT setzt voraus, daß Tatbeteiligte die kritischen Sachverhalte zur Kenntnis genommen und sich eingeprägt haben (Lykken, 1998, S. 288). Es ist kaum auszuschließen, daß sie bestimmte Details nicht wiedererkennen können. Auf der anderen Seite dürfen die kritischen Sachverhalte für Personen ohne Tatwissen keine besondere Bedeutsamkeit aufweisen. Zwar können einzelne relevante Items auch aus nicht deliktbezogenen Gründen für Unschuldige signifikant hervorstechen. Mit zunehmender Fragenanzahl wird aber die Gefahr für das mehrfache Auftreten derartiger Konstellationen vernachlässigbar klein (Steller & Dahle, 1999, S. 146). Unter der Voraussetzung einer adäquaten Testkonstruktion ist die Wahrscheinlichkeit, daß Tatunbeteiligte regelmäßig auf die relevanten Items stärker reagieren, relativ gering. Durch eine Testverlängerung (Erhöhung der Fragen- und Itemzahl bzw. mehrmalige Darbietung; vgl. Elaad & Ben-Shakhar, 1997, S. 594f.; Lykken, 1959, S. 387f.) und eine geeignete Auswertungsprozedur läßt sich die potentielle Spezifität des Verfahrens annähernd bis auf 100% steigern (Lykken, 1998, S. 287). Allerdings ist damit eine **geringere Sensitivität** zu erwarten (Ben-Shakhar & Furedy, 1990, S. 27; Raskin, 1988, S. 102). Diese Annahmen werden durch eine Reihe weiterer Analogstudien (z. B. Lykken, 1959,

S. 386; Podlesny & Raskin, 1978, S. 351; Steller, Haenert & Eiselt, 1987, S. 339; vgl. auch die Übersichtsarbeiten von Ben-Shakhar & Furedy, 1990, S. 50ff.; Berning, 1992, S. 125ff.; Steller, 1987, S. 44ff.) und v. a. durch die wenigen Feldstudien (Elaad, 1990, S. 525; Elaad et al., 1992, S. 761ff.) zum TWT gestützt. Eine solche Fehlerverteilung, wonach Schuldige mit einem gewissen Risiko unentdeckt bleiben, während Unschuldige annähernd perfekt gegen Falschdiagnosen geschützt sind, entspricht aber durchaus dem Rechtsgrundsatz „im Zweifel für den Angeklagten".

Das **Lykken-Scoring** beruht zwar im Gegensatz zu dem numerischen Auswertungsverfahren des KFT auf relativ eindeutig spezifizierten Regeln. Es ist aber ebenfalls nur „**semi-objektiv**" (Steller, 1987, S. 99f.). Da bei der Amplitudenmessung die Fußpunkte und Maxima der elektrodermalen Reaktionen nicht immer eindeutig bestimmbar sind, können Beeinträchtigungen der Auswertungsobjektivität resultieren. Außerdem werden die Trennwerte für die Klassenzuordnungen unter eher pragmatischen Gesichtspunkten relativ willkürlich gesetzt. Während nach Lykken (1959, S. 386) erst bei einem Summenscore von *mehr* als der Hälfte der insgesamt möglichen Punkte ein positiver Befund erfolgt, stellen andere Forscher (z. B. Bradley & Janisse, 1981, S. 310; Giesen & Rollison, 1980, S. 7) bereits *ab* diesem Wert die entsprechende Diagnose. Statt eines numerischen Auswertungsverfahrens kann man aber auch Wahrscheinlichkeitsaussagen treffen. Ein kardinaler Vorzug gegenüber dem KFT besteht darin, daß sich die A-priori-Fehlerwahrscheinlichkeiten statistisch berechnen lassen. So ist bei gegebener Fragen- und Itemmenge das Risiko, daß ein Unschuldiger ohne Tatwissen rein zufällig auf eine bestimmte Anzahl relevanter Items maximal reagiert, durch die Binomialverteilung definiert (vgl. Lykken, 1991b, S. 221).

Ebenso wie der KFT (vgl. Honts et al., 1994, S. 257f.) gilt der TWT als **anfällig gegenüber körperlichen und mentalen Manipulationsversuchen** (Ben-Shakhar & Dolev, 1996, S. 278ff.; Honts et al., 1996, S. 90f.). Eventuell läßt sich aber diese Vulnerabilität durch die Berücksichtigung zusätzlicher physiologischer Variablen, die im Vergleich zu den elektrodermalen Reaktionen weniger willkürlich beeinflußbar sind (z. B. evozierte hirnelektrische Potentiale), vermindern (Iacono, 2000, S. 784f.). Ein weiterer Vorteil ist, daß man entdeckte Gegenmaßnahmen beim TWT auch als Indiz für Tatwissen werten kann (vgl. Abschnitt 2.7.1). Im Gegensatz zu Schuldigen dürften Tatunbeteiligte nicht in der Lage sein, ihre Reaktionen systematisch in Richtung der relevanten oder irrelevanten Items zu manipulieren, da für sie die beiden Itemtypen nicht zu unterscheiden sind. Beim KFT ist es jedoch möglich, daß ein Unschuldiger die Rolle der Kontrollfragen überbewertet (entsprechend der suggerierten Testlogik: starke Kontrollreaktionen als Schuldindiz). Durch eine mutwillige Erhöhung der relevanten Reaktionen kann er

entgegen seiner eigentlichen Intention, den Test zu bestehen, einen falsch positiven Befund provozieren (Lykken, 1998, S. 303).

Die Kritiker des KFT (vgl. Abschnitt 2.4.2) befürworten in der Regel den TWT und benennen dabei folgende **Vorteile**:

Der TWT entspricht eher einem wissenschaftlich fundierten Testverfahren, zumal er die Möglichkeit einer **Standardisierung und Objektivierung** bietet (Furedy & Heslegrave, 1991a, S. 182). Die Formulierung der Fragen und Items ist überwiegend durch die Tatumstände festgelegt und somit weitgehend unabhängig von der Interaktion zwischen Untersucher und Pb. Insgesamt erweist sich der Test als relativ resistent gegenüber potentiellen Untersucher-Effekten (Ben-Shakhar & Furedy, 1990, S. 28ff.). Man kann die Objektivität zusätzlich dadurch erhöhen, daß unterschiedliche Personen zur Konstruktion, Durchführung und Auswertung des TWT eingesetzt werden (anders als beim KFT, der gewöhnlich von einem einzelnen Untersucher geleitet wird). Zudem lassen sich Testleiter-Effekte reduzieren, indem der Untersucher, der den TWT durchführt, selbst die relevanten Items nicht kennt (Steller & Dahle, 1999, S. 145).

Sofern die Alternativen einer Multiple-Choice-Frage für unschuldige Pbn annähernd gleich plausibel und ähnlich emotional erregend sind, stellen die irrelevanten Items eine **angemessene Vergleichs- bzw. Kontrollbedingung** dar (Lykken, 1998, S. 291). Die Gleichwertigkeit der Items läßt sich vor der eigentlichen Untersuchung von Tatverdächtiger dadurch kontrollieren, daß man den TWT an „naiven“ Personen ausprobiert, die definitiv über kein Tatwissen verfügen (Lykken, 1991b, S. 220f.). Unter diesen Voraussetzungen sind Reaktionsunterschiede zwischen den beiden Itemtypen mit einer hohen Wahrscheinlichkeit auf Tatkenntnisse zurückzuführen (Furedy, 1986, S. 687). Alternative Erklärungsmöglichkeiten für das Zustandekommen solcher Reaktionsprofile, die eventuell beim KFT eine wesentliche Rolle spielen (z. B. die Angst des Pb), können eher ausgeschlossen werden (Saxe, 1991, S. 229).

Nach Lykken (1959, S. 386, 1974, S. 732) dürften sich **intra- oder interindividuelle Unterschiede** in der körperlichen Erregung bzw. Reaktivität nur unwesentlich auf die Treffsicherheit auswirken, da nur der Reaktionsvergleich, nicht aber das absolute Niveau zählt. Dies gilt ebenso für andere personenbezogene Faktoren (z. B. Angst vor positivem Befund, Vertrauen in die Methode, Täuschungsmotivation etc.). Bei der Auswertung des TWT kann man Reaktionsstereotypien berücksichtigen (vgl. diesbezügliche Problematik des KFT, Abschnitt 2.4.2). Auch individualspezifische Abweichungen der Reaktionen zwischen relevanten und irrelevanten Items (z. B. qualitative

Unterschiede) gelten als Indiz für Tatwissen. Es ist nur erforderlich, daß der Pb auf die beiden Itemtypen konsistent andersartig reagiert (vgl. Lykken, 1998, S. 304).

Die Durchführung des TWT provoziert **weniger ethische Bedenken** als der KFT (Furedy, 1993, S. 266f.). Ferner kann man das Verfahren nur im Zusammenhang mit einem konkreten Tatvorwurf sinnvoll anwenden (Ben-Shakhar, 1991a, S. 234), wodurch die Gefahr einer kommerziellen, mißbräuchlichen Nutzung im beruflichen Umfeld minimiert wird (Lykken, 1974, S. 736).

Insgesamt ist die Funktionsweise des TWT in geringerem Umfang von theoretischen Zusatzannahmen abhängig (vgl. Steller & Dahle, 1997, S. 315). Seine Prämissen sind einfach und recht schlüssig. Systematische Reaktionsunterschiede zwischen den beiden Itemtypen setzen eine entsprechende Diskriminationsleistung und somit Tatwissen voraus. Welche vermittelnden Prozesse dabei ein Rolle spielen, ist für die diagnostischen Schlußfolgerungen zunächst ohne Belang. Mittels einer geeigneten Itemauswahl ergibt sich die unterschiedliche Signifikanz der tatbezogenen Details für Täter und Tatunbeteiligte von selbst. Beim KFT hingegen muß die differentielle Bedeutsamkeit von relevanten Fragen und Kontrollfragen für Schuldige und Unschuldige während des Vortest-Interviews suggestiv vermittelt werden (Steller, 1987, S. 15). Die Validität der Ergebnisse ist unmittelbar an das Gelingen dieser Manipulationen und die Gültigkeit der zugrundeliegenden Testlogik gekoppelt.

Zusammenfassend kann man festhalten, daß die Probleme des TWT v. a. in der konkreten Anwendung liegen (Einsatzfähigkeit und Testkonstruktion), während seine Konzeption im Gegensatz zum KFT relativ unumstritten ist. Auch der Bundesgerichtshof (vgl. BGH, 2000, S. 327f.) hat in seinem Grundsatzurteil keine fundamentalen Zweifel an seiner diagnostischen Güte geäußert. Eine Untersuchung mit dem TWT führe „jedenfalls im Zeitpunkt der Hauptverhandlung“ (S. 308), wenn also der Angeklagte sowie Teile der Öffentlichkeit bereits über die wesentlichen Tatumstände informiert sind und seine Anwendung ohnehin weitgehend ausscheidet, zu einem Ergebnis ohne Beweiswert. Damit rückt aber die Durchführung während eines frühen Ermittlungsstadiums in den Bereich des Möglichen (vgl. Hamm, 1999, S. 923). Speziell dort kann der TWT als Hilfsmittel genutzt werden, um den Kreis potentiell Verdächtiger einzuengen und die Ermittlungen auf jene Personen zu konzentrieren, die nachweislich über Tatwissen verfügen.

2.8 Guilty Actions Test (GAT)

Der „**Guilty Actions Test** (**GAT**)“ (Bradley & Rettinger, 1992, S. 55) stellt eine Modifikation des Tatwissentests dar. Er basiert auf zwei Kritikpunkten am herkömmlichen Verfahren, wovon der erste theoretische Aspekte anspricht und der zweite auf praktische Probleme abzielt (vgl. Bradley, MacLaren & Carle, 1996, S. 154):

1. Bei der Standardprozedur des TWT (mit Verneinung der Items) und in den dazugehörigen Validitätsstudien liegt eine **Konfundierung von Tatwissen und Täuschung** vor (Bradley & Warfield, 1984, S. 688). Schuldige verfügen über Tatwissen und müssen auf die relevanten Items lügen. Unschuldige hingegen sind nicht in der Lage, zwischen relevanten und irrelevanten Items zu unterscheiden, und verneinen somit alle Items (subjektiv) wahrheitsgemäß. Folglich bleibt unklar, ob das physiologische Reaktionsprofil der Täter allein auf der Identifikation der relevanten Items (Tatwissen) beruht oder ob darüber hinaus auch die wahrheitswidrigen Antworten (Täuschungen) zu stärkeren relevanten Reaktionen führen (Bradley & Rettinger, 1992, S. 55). Der diagnostische Wert des TWT bleibt von dieser Kritik relativ unberührt. Unter der Voraussetzung, daß nur Pbn mit Tatwissen die beiden Itemtypen unterscheiden können, sind die vermittelnden Prozesse, die letztlich zu stärkeren relevanten Reaktionen führen, nebensächlich, solange man aus den Reaktionsmustern valide Schlußfolgerungen über die Kenntnis von Tatdetails ziehen kann (vgl. auch Furedy et al., 1988, S. 684).

2. Der TWT ermöglicht **keine ausreichende Differenzierung zwischen Tätern und Unschuldigen mit Tatwissen**. Das ist insofern problematisch, als aus einem positiven Befund meist nicht nur auf Tatkenntnisse geschlossen wird, sondern indirekt auch auf die Täterschaft (obwohl eine solche Konklusion strenggenommen nicht zulässig ist). Von unschuldigen Verdächtigen, die über Tatwissen verfügen (z. B. Mitwisser oder Zeugen), dieses aber verheimlichen (z. B. um sich selbst oder die wahren Täter zu schützen), erwartet man ein ähnliches Reaktionsprofil wie von Schuldigen (d. h. stärkere relevante Reaktionen). Beide Personengruppen erkennen die kritischen Items und beantworten sie jeweils wahrheitswidrig. Dementsprechend sind sie mittels des TWT kaum voneinander abzugrenzen. Für Unschuldige mit Tatwissen ist eine hohe Rate irrtümlich positiver Befunde zu erwarten (im Sinne von „Täterschaft indiziert“ bzw. „schuldig“). Diese Hypothese wird empirisch gestützt (Bradley, MacLaren & Carle, 1996, S. 157f.). Daraus folgt, daß die Anwendung des TWT nur dann sinnvoll ist, wenn noch keine Detailinformationen über das Verbrechen an die Öffentlichkeit gedrungen sind (vgl. Abschnitt 2.7). Dies hat zwei Gründe (Bradley & Rettinger, 1992, S. 55). Einerseits können auf diesem Wege auch Unschuldige Tatwissen erlangen, so daß sie nicht mehr von Tatbeteiligten zu unterscheiden sind. Andererseits werden die Täter in

die Lage versetzt, ihre Kenntnisse von Tatdetails durch deren Veröffentlichung zu legitimieren.

Unter den beiden o. g. Aspekten erhofft man sich vom GAT folgende **Vorteile** (Bradley, MacLaren & Carle, 1996, S. 155):

1. **Auf theoretischer Ebene** soll die Konfundierung von Tatwissen und Täuschung vermieden werden, um dadurch ein besseres Verständnis der psychologischen Prozesse zu erlangen, die zu Reaktionsunterschieden zwischen den beiden Itemtypen führen.

2. **Auf anwendungsbezogener Ebene** soll der GAT eine bessere Differenzierung zwischen Tätern und Unschuldigen mit Tatwissen ermöglichen, um somit den potentiellen Einsatzbereich zu erweitern und das Risiko falsch positiver Befunde zu reduzieren.

Die methodischen **Modifikationen** sind minimal. Im Vergleich zum TWT wird beim GAT lediglich die Formulierung der Multiple-Choice-Fragen geringfügig verändert (Bradley, MacLaren & Carle, 1996, S. 154). Während der konventionelle TWT nur nach der Kenntnis von Tatdetails fragt (z. B.: „In welchem Raum *wurde* der Diebstahl begangen? War es Raum 203, Raum 102, Raum 305 etc.?“), zielt der GAT sowohl auf das Tatwissen als auch auf die Täterschaft ab (z. B.: „In welchem Raum *haben Sie* den Diebstahl begangen? War es Raum 203, Raum 102, Raum 305 etc.?“). Demzufolge vereinigt der GAT sowohl Merkmale der direkten als auch indirekten Verfahren der forensischen Psychophysiologie. Ansonsten folgt die Durchführung den üblichen Prinzipien (vgl. Abschnitt 2.7.1; Einbetten der relevanten Items in plausible Alternativen, Permutieren der Position der kritischen Items, Verneinung aller Alternativen durch den Pb, Messung der elektrodermalen Reaktionen).

Im Grunde genommen wirkt sich die Umformulierung der Fragen nur auf den **Wahrheitsgehalt der Antworten** von Unschuldigen mit Tatwissen aus. Die Prämissen beider Verfahren (TWT und GAT) gehen davon aus, daß Täter die kritischen Alternativen erkennen und wahrheitswidrig beantworten. Unschuldige ohne Tatwissen können nicht zwischen relevanten und irrelevanten Items unterscheiden und verneinen alle Alternativen aufrichtig. **Unschuldige mit Tatwissen** verhalten sich beim TWT analog zu den Schuldigen (Identifikation und wahrheitswidrige Beantwortung der relevanten Items), wohingegen sie beim GAT die tatbezogenen Details zwar identifizieren können, die Tatbegehung aber wahrheitsgemäß abstreiten. Falls neben dem Tatwissen auch der Wahrheitsgehalt der Antworten einen Effekt auf die Reaktionsstärke ausübt, dürfte der GAT eine gewisse Differenzierung zwischen den drei Gruppen gestatten.

Zum besseren Verständnis wird zunächst auf die bisher veröffentlichten **Analogstudien** zum GAT eingegangen (Bradley, MacLaren & Carle, 1996; Bradley & Rettinger, 1992; Bradley & Warfield, 1984). Die Untersuchungen hatten folgende **Gemeinsamkeiten**: Eine Gruppe von Pbn (Schuldige) wurde instruiert, ein Scheinverbrechen zu begehen. Dabei handelte es sich stets um einen inszenierten Raubmord. Die Szenarios beinhalteten jeweils zehn kritische Details, die man bei der anschließenden psychophysiologischen Aussagebeurteilung (GAT bzw. TWT) als relevante Items verwendete. Bei den entsprechenden Multiple-Choice-Fragen formulierte man zu jedem relevanten Item vier irrelevante Alternativen. Als physiologische Variable wurden die Amplituden der Hautwiderstandsreaktionen bestimmt und mit dem Lykken-Scoring (vgl. Abschnitt 2.7.1) ausgewertet (Bradley & Rettinger erfaßten zusätzlich die Respiration, die hier jedoch unbeachtet bleibt). Lag die Gesamtpunktzahl oberhalb eines Trennwerts von zehn Punkten, resultierte ein positiver Befund. Für das Erzielen eines negativen Befunds stellte man eine finanzielle Belohnung in Aussicht. Sowohl die Punktwerte („Lykken-Scores“) als auch die Treffsicherheit wurden statistisch analysiert.

In allen drei Studien gab es außerdem **Unschuldige mit Tatwissen**, die das Scheinverbrechen zwar nicht begingen, aber in irgendeiner Weise über die kritischen Details in Kenntnis gesetzt wurden. Z. B. durften einige Vpn die Tat als „Zeugen“ beobachten (Bradley, MacLaren & Carle, 1996, S. 156; Bradley & Warfield, 1984, S. 685), oder andere mußten eine schriftliche Zusammenfassung des Szenarios lesen (vgl. Bradley & Rettinger, 1992, S. 56; Bradley & Warfield, 1984, S. 685). Im Experiment von 1984 sollte zusätzlich eine Gruppe Unschuldiger eine Aktivität durchführen, die in keinem unmittelbaren Zusammenhang mit dem Scheinverbrechen stand, aber ebenfalls die kritischen Details beinhaltete. Dadurch erwarb die Gruppe Kenntnis von den relevanten Items, ohne sich über deren Tatbezug bewußt zu sein. Diese Bedingung wird aufgrund ihrer mangelnden „Realitätsnähe“ (vgl. Bradley & Warfield, 1984, S. 688) im folgenden nicht berücksichtigt. Bradley und Rettinger sowie Bradley und Warfield untersuchten außerdem eine Kontrollgruppe **Unschuldiger ohne Tatwissen**. In beiden Studien absolvierten alle Pbn einen GAT (von Bradley & Warfield noch als „guilty knowledge test“ bezeichnet). Bradley, MacLaren und Carle (1996, S. 156) verglichen zudem den GAT und TWT direkt miteinander, in dem sie jeweils die Hälfte der Gruppen mit einem der beiden Verfahren testeten. Unabhängig davon wurde auch der Antwortmodus interindividuell über drei Stufen variiert (Verneinung, Itemwiederholung oder Schweigen).

Für den **GAT** konnten Bradley und seine Mitarbeiter zeigen, daß die schuldigen Pbn im Durchschnitt höhere Lykken-Scores erzielten als alle anderen Gruppen (also auch Zeugen oder schriftlich informierte Unschuldige). Unter Standardbedingungen (Verneinung

der Items) wurden Schuldige fast immer (90% – 100%) korrekt klassifiziert, während nur ein Teil der Unschuldigen mit Tatwissen irrtümlich als „schuldig“ eingestuft wurde. Die entsprechende Rate falsch positiver Befunde schwankte je nach Studie zwischen 25% (Bradley & Warfield, 1984, S. 687) und 50% (Bradley, MacLaren & Carle, 1996, S. 157; Bradley & Rettinger, 1992, S. 58). Für Unschuldige ohne Tatwissen hingegen resultierten erwartungsgemäß stets valide negative Diagnosen. Die Unterschiede zwischen Schuldigen und Unschuldigen mit Tatwissen waren nicht auf Gedächtniseffekte zurückzuführen, d. h., die Gruppen differierten nicht signifikant im Hinblick auf das Erinnern bzw. Wiedererkennen der relevanten Items. Die Ergebnisse deuten darauf hin, daß in vielen Fällen allein die Kenntnis von Tatdetails *nicht* für einen positiven GAT-Befund ausreicht (nur eine Teil der Unschuldigen mit Tatwissen wurde als „schuldig“ eingestuft). Eventuell bedarf es dazu zusätzlich einer wahrheitswidrigen Beantwortung der relevanten Items (wie bei den Schuldigen).

Bradley, MacLaren und Carle (1996) stellten einen **Vergleich zwischen GAT und TWT** an und variierten außerdem den Antwortmodus (Verneinung, Itemwiederholung oder Schweigen). Im TWT erzielten sowohl Schuldige als auch Unschuldige mit Tatwissen (Zeugen) relativ hohe Lykken-Scores, wenn sie mit Nein antworteten, d. h. auf die relevanten Items logen. Im Gegensatz dazu erbrachte der GAT für die Unschuldigen mit Tatwissen unter allen Antwortbedingungen relativ niedrige Punktwerte. Bei Verneinung der Items wurden 90% (GAT) bzw. 80% (TWT) der Schuldigen entdeckt, während 50% (GAT) bzw. nur 10% (TWT) der Unschuldigen valide negative Diagnosen erhielten. Dieses Ergebnis weist auf einen Effekt des Wahrheitsgehalts hin. Personen mit Tatwissen – gleichgültig, ob schuldig oder unschuldig – leugnen die relevanten Items des TWT wahrheitswidrig (vgl. die hohen Raten positiver Befunde). Beim GAT hingegen müssen die informierten Unschuldigen unter keiner Antwortbedingung lügen, also auch nicht bei der Verneinung der relevanten Items (vgl. niedrigere Rate falsch positiver Befunde im Vergleich zum TWT).

Insgesamt ist festzustellen, daß für beide Verfahren die Kenntnis tatrelevanter Details eine notwendige Bedingung für positive Befunde darstellt. In vielen Fällen reicht jedoch Tatwissen allein nicht aus, um mit dem GAT als schuldig eingestuft zu werden. Folglich ermöglicht der GAT eine bessere Differenzierung zwischen Schuldigen und Unschuldigen mit Tatwissen als der TWT.

Dennoch bleiben einige **Probleme** ungelöst:

Die Annahme, daß Lügen auf die kritischen Items die relevanten Reaktionen verstärkt (Bradley & Rettinger, 1992, S. 58f.), läßt sich mit den **bisherigen Studien** nicht ein-

deutig belegen. Bradley und Mitarbeiter variierten zwar den Wahrheitsgehalt interindividuell. Sie analysierten jedoch nicht direkt die Reaktionsamplituden, sondern nur die per Lykken-Scoring ermittelten Punktwerte und Trefferhäufigkeiten. Daraus kann man nur mittelbar schließen, daß bereits die Kenntnis von Tatdetails überzufällig zu einem Reaktionsprofil mit erhöhten relevanten Reaktionen führt und daß Täuschungen die Reaktionsunterschiede zu den irrelevanten Items steigern.

Bradley, MacLaren und Carle (1996, S. 157) klassifizierten anhand des TWT mit Verneinung der Items 80% der Täter und 90% der Unschuldigen mit Tatwissen als „schuldig“. Auf den ersten Blick widerspricht ihr Ergebnis den Resultaten zweier **anderer Experimente**, die auch mit dem TWT eine klare Diskrimination zwischen diesen Gruppen erzielten (Giesen & Rollison, 1980, S. 8f; Stern, Breen, Watanabe & Perry, 1981, S. 680f.). In beiden Experimenten wurden jedoch die Unschuldigen über die relevanten Items informiert, ohne deren Bezug zum Scheinverbrechen zu kennen (z. B. Lesen einer Geschichte, in der die relevanten Items ohne kriminelle Assoziationen eingebettet waren). Diese Bedingung kann man kaum mit jenen Unschuldigen vergleichen, die sich der Tatrelevanz ihrer Kenntnisse bewußt sind (vgl. etwa die Augenzeugen von Bradley, MacLaren & Carle, 1996). Abgesehen davon sollten die Pbn von Giesen und Rollison bzw. Stern et al. auf alle Items schweigen, so daß dort keine Variation des Wahrheitsgehalts vorlag. Auch Bradley et al. fanden beim TWT ohne verbale Antworten doppelt so viele valide positive (60%) wie falsch positive Befunde (30%). Die Hypothese, wonach Tatwissen nicht zwingend zu einem positiven Befund führt und der Wahrheitsgehalt der Antworten eine gewisse Rolle spielt, ist also mit allen drei Studien vereinbar. Insofern kann man annehmen, daß der TWT zumindest bei wahrheitswidriger Verneinung und bekannter Tatrelevanz der kritischen Items nicht zwischen Schuldigen und Unschuldigen mit Tatwissen differenziert.

Die **Differenzierung** gelingt aber auch mit dem GAT **nicht optimal**. 25% bis 50% der Unschuldigen mit Tatwissen wurden in den Analogstudien irrtümlich als „schuldig“ eingestuft. Außerdem konnten Elaad und Ben-Shakhar (1989, S. 450) mittels einer Reanalyse der Daten von Bradley und Warfield zeigen, daß der durchschnittliche Lykken-Score der betreffenden Unschuldigen signifikant über dem per Zufall erwarteten Wert lag. Ein positiver GAT-Befund ist also kein stringenter Täterschaftsnachweis (vgl. Bradley, MacLaren & Carle, 1996, S. 158). Ebenso läßt sich die **Generalisierbarkeit** dieser Ergebnisse auf die Feldanwendung in Frage stellen (Bradley & Warfield, 1984, S. 688). Bei Realfällen ist anzunehmen, daß die relevanten Items aufgrund des konkreten Tatverdachts, der drohenden Konsequenzen (Verurteilung und Strafe) und einer entsprechend gesteigerten Täuschungsmotivation auch für Unschuldige mit Tatwissen

eine größere Signifikanz aufweisen. Daraus könnten stärkere relevante Reaktionen und mit erhöhter Wahrscheinlichkeit falsch positive Diagnosen resultieren.

Zusammenfassend läßt sich konstatieren, daß der potentielle Nutzen des GAT auch von der diagnostischen Zielsetzung abhängt. Falls man von der Tatbegehung absieht und allein tatrelevante Kenntnisse nachweisen will, ist der TWT zu bevorzugen, da mit dem GAT das Tatwissen informierter Unschuldiger eher unentdeckt bleibt. Wenn jedoch auch die Täterschaft von Interesse ist, kann der GAT eine Grobunterscheidung zwischen Schuldigen und Unschuldigen mit Tatwissen ermöglichen. Dabei muß man jedoch berücksichtigen, daß die Differenzierung nicht perfekt erfolgt, sondern anfällig gegenüber falsch positiven Befunden ist. In der Feldanwendung liegt dieses Risiko eventuell noch höher. Sofern aber die Laborergebnisse auf Realfälle übertragbar sind, könnte der GAT über die Diagnose von Tatkenntnissen hinaus einen ersten richtungsweisenden Ansatz zur Täterschaftsermittlung bieten.

2.9 Psychophysiologische Aussageforschung

Dieses Kapitel skizziert die **Validitätsforschung** zur forensischen Psychophysiologie. Es wird jedoch keine umfassende Abhandlung der bisherigen Studien geboten. Dazu sei auf die bereits vorliegenden Übersichtsarbeiten verwiesen (z. B. Ben-Shakhar & Furedy, 1990; Berning, 1992; Raskin, Honts & Kircher, 1997; Steller, 1987). Statt dessen werden hier die Methoden der psychophysiologischen Aussageforschung erörtert, ihre Ergebnisse zusammengefaßt und unter Berücksichtigung der methodenspezifischen Vor- und Nachteile diskutiert. Die Darstellung der empirischen Befundlage ist auf die konventionellen Verfahren beschränkt (KFT und TWT). Die wenigen Validitätsstudien zum DLT bzw. GAT wurden bereits in den jeweiligen Abschnitten geschildert. Die Unterscheidung zwischen **Feld- und Analogstudien** orientiert sich an der Terminologie der OTA (1983, S. 47f., S. 61f.). Davon lassen sich wiederum **Laborexperimente** abgrenzen, in denen psychophysiologische Aussagebeurteilungen ohne direkten Bezug zu einem (Schein-)Verbrechen bzw. nicht in Analogie zu realen Untersuchungen durchgeführt werden (Steller, 1987, S. 38).

In den **folgenden Abschnitten** werden jeweils die *Vorgehensweisen* von Feld- und Analogstudien beschrieben, ihre *Ergebnisse für den KFT bzw. TWT* getrennt zusammengefaßt und schließlich die *Kritik* an den Forschungsansätzen erläutert. Anschließend wird auf die Laborexperimente ohne Verbrechenssimulationen eingegangen und ein Fazit aus der bisherigen psychophysiologischen Aussageforschung gezogen.

2.9.1 Feldstudien

Vorgehensweise von Feldstudien:

Sog. Feldstudien schätzen die kriterienbezogene Validität der Befragungstechniken anhand von **Realfällen**. D. h., die Ergebnisse „in vivo“ – unter nicht experimentell kontrollierten Bedingungen – durchgeführter Tests werden auf ihre Gültigkeit hin überprüft (OTA, 1983, S. 39). Dabei läßt sich die objektive Wahrheit („ground truth“, Podlesny & Raskin, 1977, S. 782) hinsichtlich der Täterschaft bzw. Glaubhaftigkeit oft nicht mit absoluter Sicherheit bestimmen. Deshalb vergleicht man vorwiegend im strafrechtlich relevanten Kontext die Befunde der psychophysiologischen Aussagebeurteilungen mit anderen Glaubwürdigkeitsindikatoren, wie z. B. Geständnisse bzw. Gerichts- oder Expertenurteile. Diese post hoc anfallenden Kriterien können bezogen auf den Einzelfall sowohl Schuld als auch Unschuld indizieren. Ein Geständnis belastet den Pb, wenn er sich selbst bezichtigt, und dient gleichzeitig als Entlastungsnachweis für andere Tatverdächtige. Entsprechendes gilt auch für gerichtliche Verurteilungen. Bei Expertenentscheidungen soll eine Gruppe von mehreren Fachleuten (z. B. Juristen) ohne Kenntnis der polygraphischen Befunde oder Gerichtsurteile nur auf der Basis der Aktenlage Einschätzungen über die Schuld der Pbn abgeben (z. B. Bersh, 1969, S. 400f.). Da die Expertenurteile nicht immer einstimmig ausfallen, wird teilweise auf Mehrheitsvoten zurückgegriffen (z. B. Barland & Raskin, 1976, zitiert nach Berning, 1992, S. 69f.). Darüber hinaus unterzieht man in Feldstudien die polygraphischen Aufzeichnungen meist einer blinden Reanalyse, um zu vermeiden, daß zusätzliche Informationsquellen neben den physiologischen Daten die Testergebnisse beeinflussen. Die Resultate dieser Sekundärauswertung und/oder die Originalbefunde werden dann den jeweiligen Kriterien gegenübergestellt und daraus die Trefferquoten berechnet.

Feldstudien zum KFT:

Hier werden zunächst die **Ergebnisse** einiger Übersichtsarbeiten exemplarisch aufgelistet. Nach Ben-Shakhar und Furedy (1990, S. 49) schwanken die Trefferquoten von KFT-Feldstudien für Schuldige zwischen 76.0% und 94.3% (mit Stichprobengrößen gewichteter Mittelwert: 84.0%). Die Trefferquoten für Unschuldige liegen je nach Studie zwischen 20.0% und 90.5% (gewichteter Mittelwert: 72.2%). Berning (1992, S. 117f.) berichtet Bandbreiten von 89.5% bis 100% valider positiver Befunde (ungewichteter Durchschnitt: 97.2%) und 35.7% bis 94.8% valider negativer Befunde (Durchschnitt: 74.8%). Raskin et al. (1997, S. 575) stellen die Ergebnisse von vier ihrer Ansicht nach „high quality field studies“ zusammen. Deren Trefferquoten betragen 73% bis 100% für Schuldige (gewichteter Durchschnitt: 88%) und 30% bis 83% für Un-

schuldige (gewichteter Durchschnitt: 49%). Bei Steller (1987, S. 36f.) streuen die Angaben zur Sensitivität von 75.0% bis 98.6% und zur Spezifität von 12.5% bis 91%. Demnach kommen die Autoren im Hinblick auf die kriterienbezogene Validität des KFT zu abweichenden Gesamtergebnissen und bisweilen diskrepanten Schlußfolgerungen.

Die **Schwankungen zwischen den Übersichtsarbeiten** lassen sich einerseits auf die Auswahl der Feldstudien zurückführen. Dabei berücksichtigen die Autoren teilweise unterschiedliche Arbeiten. Andererseits treten auch Abweichungen bei der Ergebnisdarstellung einzelner Untersuchungen auf, da verschiedene Teilergebnisse verwertet werden (z. B. numerische vs. global-intuitive Auswertung bzw. Originaluntersuchung vs. blinde Reanalyse). Darüber hinaus sind die Prozentwerte wesentlich davon abhängig, ob man unentscheidbare Fälle als eigenständige Kategorie einkalkuliert (vgl. Ben-Shakhar & Furedy, 1990, S. 52; Raskin et al., 1997, S. 575) oder nur eindeutig positive und negative Befunde zur Bestimmung der Treffsicherheit heranzieht (z. B. Berning, 1992, S. 118[4]). Letztere Vorgehensweise ist in vielen Studien üblich (Lykken, 1988, S. 116). Sie liefert – sofern uninterpretierbare Ergebnisse vorliegen – unweigerlich höhere Trefferquoten, da der Grundwert der Prozentrechnung gesenkt wird. Zur besseren Einschätzung des diagnostischen Nutzens ist allerdings auch der Anteil unentscheidbarer Tests anzugeben, zumal sie in manchen Feldstudien v. a. bei Unschuldigen bis zu ca. einem Drittel der Fälle ausmachen können (vgl. Ben-Shakhar & Furedy, 1990, S. 49).

Bei Betrachtung der Trefferquoten fallen mehrere Dinge auf. Die **Streuung** zwischen den einzelnen Feldstudien ist in allen Übersichtsarbeiten relativ groß. Die Prozentangaben schwanken von annähernd 100% (primär bei Schuldigen) bis hin zu Werten, die erheblich unterhalb des Zufallsniveaus von 50% liegen (speziell bei Unschuldigen). Die Spannweiten der Trefferquoten sind für negative Befunde größer als für positive. In Anlehnung an Patrick und Iacono (1991, S. 235) läßt sich dies darauf zurückführen, daß man im wesentlichen zwei Gruppen von Feldstudien zum KFT unterscheiden kann. In manchen resultieren Spezifitätswerte von um die 90%, während andere die Trefferquoten für unschuldige Pbn nur im Bereich des Zufallsniveaus ansiedeln. In beiden Gruppen findet man eine durchschnittliche Sensitivität von über 80%. Es ist anzunehmen, daß die hohen Trefferquoten einiger Feldstudien auf Methodenartefakte zurückzuführen sind (s. u.). Darüber hinaus sind die vielfältigen Varianten des KFT zu berücksichtigen, die in der Praxis zur Anwendung kommen und somit die Vergleichbarkeit der Studien

[4] Berning (1992, S. 116f.) gibt im Text auch die Trefferquoten unter Einbeziehung der nicht zuordbaren Befunde wieder. Der Einfachheit halber sind hier aber nur die Ergebnisse ihrer tabellarischen Übersicht zusammengefaßt

in Zweifel ziehen. Andere mutmaßliche Gründe für die Streuungen erscheinen weitaus spekulativer. Dem häufig vorgebrachten Einwand, Mängel in den verwendeten Befragungstechniken und in der Kompetenz der Untersucher seien für niedrige Trefferquoten verantwortlich (z. B. Raskin, 1978, S. 145, 1982, S. 346), kann man das Argument von Ben-Shakhar und Furedy (1990, S. 48) entgegnen, daß die Mehrzahl der Studien mit den extremsten Fehlerraten von sehr erfahrenen KFT-Untersuchern mit akademischer Ausbildung durchgeführt wurden.

Im **Durchschnitt** liegen jedoch sowohl die Sensitivität als auch die Spezifität mehr oder weniger deutlich oberhalb der Zufallstrefferquote. Dabei läßt sich ein Trend der Fehlerverteilung zuungunsten Unschuldiger erkennen. Mehr Unschuldige als Schuldige werden falsch klassifiziert. Diese überproportionale **Häufung irrtümlich positiver Befunde** entspricht den Erwartungen, die eine kritische Analyse der Methode nahelegt (vgl. Abschnitt 2.4.2). Nach Raskin (1981, S. 15f.) darf man die einseitige Fehleranfälligkeit des KFT nicht nur im Sinne einer Benachteiligung Unschuldiger interpretieren. Statt dessen schlägt er vor, die Testergebnisse unter dem Gesichtspunkt ihrer diagnostischen **Aussagekraft** zu betrachten. Relativ viele Unschuldige werden fälschlich als unglaubwürdig eingestuft und nur wenige Täter als glaubwürdig. Darum ist es sehr wahrscheinlich, daß es sich bei einem Pb mit einem negativen Testergebnis tatsächlich um einen Unschuldigen handelt. Im Gegensatz dazu besteht ein ausgeprägtes Risiko, daß Pbn mit einem positiven Befund in Wirklichkeit unschuldig sind. In allgemeinen diagnostischen Termini ausgedrückt (Amelang & Zielinski, 1997, S. 363f.; Kallus & Janke, 1988, S. 135), ist der prädiktive Wert negativer Zuordnungen relativ hoch und der Prädiktionswert positiver Klassifizierungen eher niedrig. Aufgrund der Annahme, daß negative Befunde eine hohe **Verläßlichkeit** aufweisen und somit den Tatvorwurf stark in Zweifel ziehen, wird gefolgert, daß sich der KFT besonders gut als forensischer Entlastungsnachweis bzw. als Unschuldsindiz eignet (Steller, 1987, S. 151). Positive Befunde hingegen sind wegen ihrer geringeren Aussagekraft mit Vorsicht zu bewerten.

Der mutmaßliche Nutzen des KFT als **Entlastungsverfahren** muß jedoch angezweifelt werden (vgl. Rill & Vossel, 1998, S. 484). Neben dem Risiko für Unschuldige, durch Falschklassifikationen zusätzlich belastet zu werden, ist auch zu bedenken, daß die o. g. Überlegungen auf dem Postulat einer hohen Trefferquote für Schuldige beruhen. Es besteht aber Grund zu der Annahme, daß die meisten Feldstudien die Sensitivität des KFT überschätzen. Und je weniger valide negative Entscheidungen getroffen werden, desto niedriger liegt die Verläßlichkeit dieser Diagnose. Außerdem ist deren prädiktiver Wert auch von der Grundrate Schuldiger innerhalb der getesteten Population abhängig. Mit steigendem Anteil Schuldiger sinkt die Verläßlichkeit negativer Befunde (Raskin, 1987, S. 399ff., 1988, S. 103ff.).

Feldstudien zum TWT:

Die einzigen bis dato veröffentlichten **Feldstudien zum TWT** wurden in Israel durchgeführt. Im Rahmen ihrer Ermittlungsarbeit setzt die israelische Polizei neben dem KFT oft auch indirekte Verfahren zur psychophysiologischen Aussagebeurteilung ein (Ben-Shakhar & Furedy, 1990, S. 123). Auf die beiden Studien wird etwas ausführlicher eingegangen, da sie in den bisherigen Übersichtsarbeiten nur sporadisch Berücksichtigung fanden.

Elaad (1990) untersuchte 98 durch Geständnisse verifizierte Fälle (48 Schuldige und 50 Unschuldige). Die Stichprobe stammte aus der Grundgesamtheit von Tatwissentests, die die Israel Police Scientific Interrogation Unit im Zeitraum von 1979 bis 1985 durchgeführt hatte. Anhand der Hautwiderstandsreaktionen und einer modifizierten Version des Lykken-Scorings wurden von den 50 unschuldigen Pbn 46 (92%) korrekt negativ diagnostiziert und einer falsch positiv (2%). Unter Vernachlässigung der drei unentscheidbaren Fälle (6%) resultierte für Unschuldige eine Trefferquote von 97.9%. Jeweils 20 schuldige Pbn erzielten einen validen positiven (42%) bzw. falsch negativen Befund (42%), und bei acht Schuldigen (16%) war das Ergebnis nicht interpretierbar. Ohne die Kategorie „unentscheidbar“ betrug die Sensitivität 50%. Mittels einer optimierten Entscheidungsregel stieg die Trefferquote für Schuldige auf 65%, und für Unschuldige sank sie geringfügig auf 94%.

Analog zu der Studie von Elaad (1990) analysierten **Elaad et al. (1992)** eine Stichprobe von 80 Tatverdächtigen, die zwischen 1985 und 1991 in der o. g. Polizeiabteilung einen TWT absolviert hatten. Als Validierungskriterium dienten wiederum Geständnisse. Jeweils die Hälfte der Pbn war kriteriumsbezogen schuldig bzw. unschuldig. Elaad et al. berücksichtigten neben elektrodermalen auch respiratorische Größen. Anhand der Amplituden der Hautwiderstandsreaktionen wurden 97.4% der Unschuldigen und 53.3% der Schuldigen korrekt klassifiziert (jeweils unter Ausschluß nicht interpretierbarer Ergebnisse). Die Auswertung der Atmungsreaktionen erbrachte entsprechende Trefferquoten von 97.2% bzw. 53.1%. Durch die Kombination beider Variablen stieg die Sensitivität auf 75.8% und die Spezifität sank auf 94.1%. Außerdem lag die Rate unentscheidbarer Fälle mit insgesamt 16.25% etwas höher, als wenn nur eine der beiden physiologischen Parameter herangezogen wurde (jeweils 15%).

Lykken (1998, S. 291f.) **kritisiert**, daß die Tests in beiden Feldstudien aus nur ein bis sechs unterschiedlichen Multiple-Choice-Fragen bestanden (im Durchschnitt ca. zwei), die zwei- bis viermal dargeboten wurden. Hätte man mehr tatrelevante Details abfragt, dann wäre mit einer höheren Treffsicherheit zu rechnen gewesen. Im Einklang mit die-

ser Hypothese zeigte sich in der Arbeit von Elaad et al. (1992, S. 762) mit zunehmender Fragenanzahl ein gradueller Anstieg der Quote valider positiver Befunde. Die Zusammenhänge waren jedoch nicht statistisch signifikant. Im Gegensatz dazu fand Elaad (1990, S. 525) eine maximale Sensitivität, wenn lediglich zwei verschiedene Multiple-Choice-Fragen gestellt wurden. Mit zunehmender Fragenmenge sank die Trefferquote für Schuldige wieder ab. Dieser Effekt ist möglicherweise darauf zurückzuführen, daß in der praktischen Anwendung des TWT die Entscheidung, mehr als zwei Fragen darzubieten, vom Ergebnis der ersten beiden abhängt. D. h., falls ein kurzer TWT keine eindeutigen Ergebnisse erbringt, wird er verlängert. Bei jenen Schuldigen, die mit mehr als zwei Fragen getestet wurden, wäre somit auch die Trefferquote relativ niedrig, die sich nur anhand der ersten beiden Fragen ergeben hätte. Die Resultate von Elaad (1990, S. 525) bestätigen diese Annahme und sprechen dafür, daß man die Reduktion der Sensitivität mit steigender Fragenanzahl nicht direkt auf die Testverlängerung zurückführen kann, sondern die betreffenden Pbn insgesamt relativ schlecht mit dem TWT zu entdecken waren.

Darüber hinaus wurde der TWT fast immer im Anschluß an einen KFT durchgeführt (Elaad, 1990, S. 523; Elaad et al., 1992, S. 759). Somit sind **Sequenzeffekte** nicht auszuschließen. Die Analogstudie von O'Toole et al. (1994, S. 260f.) deutete darauf hin, daß eine sequentielle Testung jedes Pb mit zwei unterschiedlichen Verfahren (KFT und danach TWT bzw. umgekehrt) eine geringere Entdeckbarkeit Schuldiger im jeweils zweiten Test nach sich zog. Die Autoren führten diesen Befund auf Habituationseffekte zurück, die ein Schwinden der Reaktionsunterschiede zwischen den relevanten Reizen und den Vergleichsstimuli bewirken könnten. Zwei weitere Scheinverbrechen-Experimente erzielten ähnliche Ergebnisse. Balloun und Holmes (1979, S. 318) boten ihren Pbn die Fragenreihe eines TWT (mit veränderter Abfolge der Items) zweimal nacheinander dar. Bei der Testwiederholung sank die Sensitivität rapide im Vergleich zum ersten Durchgang. Bradley und Janisse (1981, S. 310f.) fanden ebenfalls außergewöhnlich niedrige Trefferquoten für einen TWT, der stets einem KFT folgte. In einem anderen Laborexperiment (Waid et al., 1979, S. 17f.) mußten die Vpn der Experimentalgruppe Codewörter auswendig lernen und wurden diesbezüglich mit mehreren Verfahren getestet. Die Autoren führten die geringe Sensitivität der von ihnen verwendeten indirekten Befragungstechniken (POT und TWT) darauf zurück, daß man sie jeweils erst im Anschluß an direkte Verfahren (zwei KFTs) einsetzte.

Ungeachtet dessen, daß die potentielle Treffsicherheit des TWT eventuell unterschätzt wurde, zeigten die Feldstudien eine **Fehlerverteilung**, die sich weniger stark ausgeprägt auch in Analogstudien abzeichnet (s. u.). Elaad (1990) und Elaad et al. (1992) bestätigten die Erwartung, daß man beim TWT mit deutlich mehr falsch negativen als

falsch positiven Befunden rechnen muß (vgl. Abschnitt 2.7.2). Dieser Trend ist der typischen Fehlerrichtung des KFT genau entgegengesetzt. Er läßt sich dadurch erklären, daß Tatbeteiligte oft nicht alle relevanten Items zur Kenntnis nehmen, sich merken und später beim TWT wiedererkennen (Raskin et al., 1997, S. 579). Eine Reihe von Faktoren kann die Wahrnehmung, Gedächtnisspeicherung und Identifikation relevanter Details beeinträchtigen (vgl. Elaad, 1990, S. 522), wie z. B. die Aufmerksamkeitsausrichtung der Pbn während der Tat, ihr emotionaler und körperlicher Zustand, die physikalischen Gegebenheiten am Tatort, die Komplexität der Ereignisse, die Zeitspanne zwischen Tat und Test, der Erhalt zusätzlicher Informationen über den Tathergang, die Art der Befragung etc. (für eine ausführlichere Diskussion möglicher Einflußfaktoren siehe auch Wells & Loftus, 1984).

Der TWT weist also eine besondere Anfälligkeit für falsch negative Befunde auf, während Unschuldige annähernd perfekt gegen Fehlklassifikationen geschützt sind. Analog zum KFT läßt diese typische Fehlerverteilung auch bestimmte Rückschlüsse auf die **Verläßlichkeit der Diagnosen** zu. Bei angemessener Testdurchführung ist die Wahrscheinlichkeit sehr gering, daß ein positiv Getesteter über keinerlei Tatwissen verfügt. Andererseits kann es vorkommen, daß Schuldige z. B. aufgrund von Aufmerksamkeits- oder Erinnerungsdefiziten bzw. erfolgreichen Manipulationsversuchen irrtümlich als „glaubwürdig“ klassifiziert werden. Demzufolge ist die Aussagekraft positiver Befunde relativ stark ausgeprägt (Raskin, 1989, S. 289). Darum soll sich der TWT in der forensischen Anwendung v. a. als **Belastungsverfahren** und zur Anfechtung der Unschuldsvermutung eignen (vgl. Steller, 1987, S. 152). Im Vergleich dazu ist der prädiktive Wert negativer Befunde als geringer einzustufen, da ein überproportionales Risiko für Falschklassifikationen Schuldiger besteht. Allerdings muß man ähnlich wie beim KFT beachten, daß der Prädiktionswert der Diagnosen auch von der Grundrate Schuldiger in der getesteten Population abhängt.

Kritik an Feldstudien:

Im folgenden werden grundlegende methodische Bedenken diskutiert, ohne detailliert auf einzelne Feldstudien und die Kontroverse, welche davon verläßliche Schätzungen der Treffsicherheit liefern, einzugehen. Die Methodenkritik bezieht sich im wesentlichen auf die Untersuchungen zum KFT, zumal noch wenige Studien zum TWT vorliegen. Prinzipiell muß aber jegliche Form von psychophysiologischer Aussageforschung im Feld die folgenden Aspekte beachten. Probleme ergeben sich dabei v. a. hinsichtlich der **Qualität der Validierungskriterien**, der **Unabhängigkeit zwischen Testbefund und Kriterium** und der **Selektivität der Stichproben** (vgl. Elaad, 1990, S. 522). Auch andere Kritikpunkte erscheinen diskussionswürdig, z. B. die Tatsache, daß viele Feld-

studien zum KFT aus dem beruflichen Umfeld stammen (vgl. Horvath, 1984, S. 239) und ihre Ergebnisse somit nicht direkt auf die forensische Anwendung übertragbar sind. Zudem werden die mangelnde Kompetenz der beteiligten Untersucher, die oft nicht nachvollziehbare Auslese der Fälle und die Einbeziehung von Tests an Opfer bzw. Zeugen moniert (Lykken, 1981, S. 123; Raskin, 1987, S. 396ff.). Einige der Studien ermöglichen auch nur Rückschlüsse auf die Inter-Rater-Reliabilität, nicht aber auf die Validität (Lykken, 1998, S. 129). Grundsätzlich muß man an der bisherigen Feldforschung beanstanden, daß stets bereits vorliegende polygraphische Untersuchungen post hoc reanalysiert wurden, anstatt die Datenerhebung prospektiv zu planen. Diese Vorgehensweise macht die Ergebnisse besonders anfällig für methodische Probleme, z. B. bei der Wahl der Validierungskriterien und Stichproben (s. u.).

Wie bereits erwähnt, kann man die tatsächliche Schuld oder Unschuld der Pbn in Realfällen nicht ohne weiteres eindeutig bestimmen. Sie werden lediglich durch entsprechende Kriterien indiziert (Geständnisse, Gerichts- und Expertenurteile). Allerdings sind diese **Kriterien selbst nicht vollkommen valide** (OTA, 1983, S. 39). Geständnisse können falsch sein (vgl. Gudjonsson, 1999, 205ff.; Kassin, 1997, S. 224ff.). Gerichtsentscheidungen sind ebenfalls unzuverlässig. Einerseits lassen sich Fehlverurteilungen nicht ausschließen. Andererseits zeigt ein Rechtssystem, das nach der Maxime „im Zweifel für den Angeklagten“ entscheidet, die Tendenz, Schuldige freizusprechen bzw. Verfahren aus Mangel an Beweisen einzustellen (Steller, 1987, S. 29f.). Daher greifen manche Feldstudien zum KFT auf Expertenurteile zurück. Deren Gültigkeit ist wiederum umstritten, zumal sie teilweise auf der gleichen Datengrundlage wie Gerichtsbeschlüsse basieren (z. B. Ermittlungsakten).

Ferner sind Geständnisse, Gerichts- und Expertenurteile nicht unabhängig vom Ergebnis der pychophysiologischen Aussagebeurteilung. Man spricht in diesem Zusammenhang von einer **Kriteriumskontamination** (vgl. „contaminated criteria“, Ben-Shakhar & Furedy, 1990, S. 50). Es ist davon auszugehen, daß die Rückmeldung positiver Befunde („unglaubwürdig“) die Geständnisbereitschaft der Pbn erhöht, zumal sie im Anschluß daran einem Nachtest-Interview unterzogen werden. Viele Pbn legen im Laufe dieses Verhörs ein Geständnis ab (vgl. Raskin & Kircher, 1991, S. 221). Sie bilden somit die Gruppe der kriteriumsbezogen Schuldigen. Aufgrund des starken psychologischen Drucks besteht die Gefahr, daß sich Unschuldige fälschlicherweise selbst bezichtigen (vgl. Abschnitt 2.4.2). Ein positives Testergebnis kann also das Auftreten sowohl zutreffender als auch unzutreffender Geständnisse begünstigen. Die **Abhängigkeit der Geständnisse vom Testergebnis** ist auch einer der möglichen Gründe, weshalb die professionellen Untersucher („Polygraphers“) derart stark von der Treffsicherheit ihrer Methode überzeugt sind. Da Geständnisse im Nachtest-Interview oft die einzige Form

von Feedback darstellen, die die Untersucher über die Gültigkeit ihrer Diagnosen erhalten, werden die Polygraphers in der praktischen Arbeit nur selten mit Fehlern konfrontiert. Sie erliegen der „selbsterfüllenden Prophezeiung“, daß Geständnisse nahezu immer die vorherigen Befunde bestätigen (Lykken, 1991b, S. 217ff.; Patrick & Iacono, 1991, S. 229). In Feldstudien mit **Gerichts- und Expertenurteilen** als Kriterien sollten die entsprechenden Entscheidungen ohne Kenntnis vom Ergebnis der psychophysiologischen Aussagebeurteilung getroffen werden, um eine Kontamination zu vermeiden. Doch selbst wenn diese Forderung erfüllt ist, können die Testbefunde die Kriterien indirekt beeinflussen. Beispielsweise indem die Gerichte und Expertengremien die per Nachtest-Interview induzierten Geständnisse bei ihrer Urteilsbildung mitberücksichtigen. Die Entscheidungen der Polygraphen-Untersucher und der Gerichte bzw. Experten sind auch deshalb nicht unabhängig voneinander, weil sie teilweise auf den gleichen Hintergrundinformationen (z. B. Aktenlage) beruhen (Ginton, Daie, Elaad & Ben-Shakhar, 1982, S. 132).

Das Problem **selektiv verzerrter Stichproben** („sampling bias“, Patrick & Iacono, 1991, S. 229) hängt eng mit den Mängeln der verwendeten Kriterien und deren Abhängigkeit von den Testbefunden zusammen. Falls etwa Experteneinschätzungen herangezogen werden (z. B. Bersh, 1969), dann gehen eher eindeutige Fälle in die Analyse ein, bei denen die Experten zu einem Konsens bzw. einer Mehrheitsentscheidung hinsichtlich der Schuld oder Unschuld der Pbn gelangen konnten. Diese Fälle sind jedoch nicht unbedingt repräsentativ für die Grundgesamtheit aller polygraphischen Untersuchungen (vgl. Ben-Shakhar et al., 1982, S. 709; OTA, 1983, S. 39). Man muß davon ausgehen, daß die psychophysiologische Aussagebeurteilung v. a. in Zweifelsfällen angewendet wird, wenn die Beweislage relativ unklar ist (Fiedler, 1999, S. 25; Raskin et al., 1997, S. 571).

In den meisten Feldstudien wurden allerdings **Geständnisse** als Validierungskriterium herangezogen. Das Problem der Selektivität betrifft dabei nicht nur die Frage, ob die geständigen Pbn repräsentativ für alle Getesteten sind (vgl. dazu Lykken, 1988, S. 117). Iacono und Mitarbeiter (Iacono, 1991; Iacono & Patrick, 1987, S. 469f., 1988, S. 218; Patrick & Iacono, 1991) konnten methodenkritisch zeigen und empirisch belegen, daß die gängige Forschungsstrategie zu einer systematischen Elimination von falschen Befunden aus den Stichproben und somit zu einer Überschätzung der Treffsicherheit führen kann. Dazu muß man sich nochmals folgendes vergegenwärtigen (vgl. Iacono, 1991, S. 202): In der praktischen Anwendung des KFT ist es Usus, die Pbn nach einem positiven Befund einem Nachtest-Interview zu unterziehen, das v. a. dazu dient, Geständnisse zu provozieren. Erzielt der Pb einen negativen Befund, erfolgt in der Regel kein Verhör. Statt dessen testet man – sofern vorhanden – andere Verdächtige, bis schließlich einer

davon als „unglaubwürdig“ eingestuft wird. Auf diese Person konzentrieren sich dann die weiteren Ermittlungen. Da die Polygraphers von der Treffsicherheit ihrer Verfahren überzeugt sind, beenden sie im allgemeinen nach einem positiven Befund die Testreihe, sofern nicht Grund zur Annahme besteht, daß mehrere Personen an der Tat beteiligt waren. Vereinfacht kann man also sagen, daß sich die Stichproben in solchen Feldstudien überwiegend aus Pbn zusammensetzen, die entweder aufgrund eines positiven Befundes im Nachtest-Interview ein Schuldbekenntnis ablegten; oder sie erzielten negative Befunde, so daß die Untersucher weitere Verdächtige testeten, bis einer davon schließlich gestand und dadurch die anderen entlastete (vgl. Lykken, 1998, S. 130; Patrick & Iacono, 1991, S. 229). Nur ein Bruchteil der insgesamt durchgeführten Untersuchungen geht in die Studien ein, und zwei Arten von Fehlern werden systematisch aus den Stichproben eliminiert, nämlich falsch positiv getestete Pbn, die keine Geständnisse ablegen, weil sie unschuldig sind, und falsch negativ getestete, die nicht gestehen, weil sie sich keinem Nachtest-Interview unterziehen müssen. Nur unter der seltenen Voraussetzung, daß die Geständnisse unabhängig von der polygraphischen Untersuchung erfolgen und in den Feldstudien berücksichtigt werden, offenbaren sich die entsprechenden Irrtümer. Andernfalls können selbst Fehldiagnosen zur Überschätzung der Trefferquoten beitragen (Iacono, 1991, S. 205). Im Extremfall ist etwa vorstellbar, daß zunächst ein Täter einen irrtümlich negativen Befund erzielt, anschließend ein Unschuldiger falsch positiv getestet wird und unter dem Druck des Nachtest-Interviews die Tat gesteht. Bei der Berechnung der Trefferquoten würde man beide Falschklassifikationen als korrekte Entscheidungen werten, zumal die Fälle nur selten weiterverfolgt werden und Geständniswiderrufe Unschuldiger bzw. spätere Geständnisse der wahren Täter in den entsprechenden Feldstudien kaum Berücksichtigung finden.

Wie kann man sich aber unter diesen Bedingungen erklären, daß die **Trefferquoten nicht immer annähernd 100%** betragen? Auch dafür bietet Iacono (1991, S. 205) einen möglichen Grund. In den meisten Feldstudien werden statt der ursprünglichen Diagnosen blinde Reanalysen der polygraphischen Aufzeichnungen zur Bestimmung der Kriteriumsvalidität herangezogen. Da die Originalbefunde nur teilweise auf den physiologischen Daten beruhen und die Auswertungsobjektivität nicht perfekt ist, resultiert eine partielle Unabhängigkeit mit den Reanalysen. D. h., die Sekundärauswertungen stimmen nicht vollständig mit den ursprünglichen Diagnosen überein, so daß auch ihr Zusammenhang mit den Geständnissen schwächer ist. Entsprechend dieser Annahmen fanden Patrick und Iacono (1991, S. 234f.) deutlich höhere Trefferquoten für die Originalbefunde als für die Reanalysen.

Insgesamt läßt sich feststellen, daß die mitunter ausgeprägte Treffsicherheit des KFT in Feldstudien teils auf Methodenartefakte zurückzuführen ist. Da Schuldbekenntnisse im

Nachtest-Interview nahezu immer den vorherigen positiven Befund bestätigen, wird insbesondere die Trefferquote für Schuldige überschätzt (Iacono, 1991, S. 207). Falschklassifikationen Unschuldiger treten eher zutage, z. B. wenn der betreffende Pb nicht gesteht und die Ermittlungen auf zusätzliche Verdächtige ausgeweitet werden, von denen sich schließlich einer als eigentlicher Täter erweist (Patrick & Iacono, 1991, S. 237). Auf dem Hintergrund dieser Methodenkritik erhält auch das Paradox zwischen den hohen empirischen Validitätsschätzungen und der als gering einzustufenden Objektivität bzw. Reliabilität des KFT eine Erklärung (siehe Abschnitt 2.4.2).

2.9.2 Analogstudien

Vorgehensweise von Analogstudien:

Analogstudien untersuchen die Methoden der psychophysiologischen Aussagebeurteilung unter kontrollierten, experimentellen Bedingungen. Dazu bedienen sie sich sog. „**Scheinverbrechen**“ („mock crimes“, OTA, 1983, S. 61). Für gewöhnlich werden die Vpn randomisiert den Versuchsgruppen zugeordnet (wie etwa bei den Analogstudien der Forschungsgruppe um Raskin; z. B. Barland & Raskin, 1975a; Kircher & Raskin, 1988, Podlesny & Raskin, 1978; Raskin & Hare, 1978). Im einfachsten Fall instruiert man die Experimentalgruppe, ein simuliertes Delikt (z. B. einen fingierten Diebstahl) zu begehen („Schuldige“), während andere Pbn diesbezüglich „unschuldig“ sind (Kontrollgruppe). Das Kriterium („ground truth“) ist somit eindeutig über die Gruppenzugehörigkeit definiert. Üblicherweise wird das Scheinverbrechen im Rahmen eines mehr oder minder komplexen Rollenspiels verübt.

In wenigen Studien wurden Pbn zu **echten Vergehen** animiert, und die Teilnehmer durften selbst entscheiden, ob sie die Tat begehen wollten (Balloun & Holmes, 1979; Ginton et al., 1982; Heckel, Brokaw, Salzberg & Wiggins, 1962). Die Untersuchung von Ginton et al. (1982) gilt als Paradebeispiel einer wirklichkeitsnahen Simulation. Einer Gruppe von Polizisten wurde die Gelegenheit geboten, bei einer inszenierten schriftlichen Eignungsprüfung zu mogeln, ohne daß sie über den Versuchscharakter der Prozedur informiert waren. Die Entscheidung, ihre vermeintlichen Prüfungsergebnisse zu manipulieren, blieb den Teilnehmern überlassen. Eventuelle Täuschungsversuche konnten aber ohne Wissen der Pbn objektiv ermittelt werden. Anschließend beschuldigte man sie alle des Betrugs und bot ihnen an, sich durch einen KFT zu entlasten. Die Polizisten sollten annehmen, daß der weitere Verlauf ihrer Karriere vom Testergebnis abhängen würde. Diese Studie zeichnete sich durch ihre besondere Realitätsnähe (z. B. Freiwilligkeit der Täterschaft, Bedrohlichkeit der Konsequenzen), aber auch durch me-

thodische Probleme (mangelnde Bedingungskontrolle, kleine Stichprobe, relativ hohe Ausfallquote schuldiger Pbn) und ethische bzw. rechtliche Bedenken (Irreführung der Teilnehmer, Versuchungssituation) aus. Sie nimmt damit eine Sonderstellung im Grenzbereich zwischen Feld- und Analogstudien ein.

Die **Täuschungsmotivation** (Steller, 1987, S. 61) wird nur selten über angedrohte oder verhängte Strafen erhöht (z. B. Elektroschocks wie bei Bradley & Janisse, 1981, S. 309; oder Lykken, 1959, S. 385f.). Statt dessen greift man oft auf finanzielle Anreize für negative Befunde („glaubwürdig“) zurück. Der Verlust einer in Aussicht gestellten Belohnung kann aber im weitesten Sinne ebenfalls als Bestrafung angesehen werden (Podlesny & Raskin, 1977, S. 783f.). Gelegentlich wird auch an den Selbstwert appelliert, z. B. mit der fingierten Instruktion, daß nur besonders intelligente und emotional stabile Personen in der Lage seien, einen „Lügendetektortest“ zu bestehen (vgl. O’Toole et al., 1994, S. 256). Die Durchführung der psychophysiologischen Aussagebeurteilung soll möglichst reale Testbedingungen simulieren. Die Untersucher sind generell nicht über die Gruppenzugehörigkeit der Pbn informiert. Die Testergebnisse werden anschließend mit der Kriteriumsvariable verglichen und daraus die Trefferquoten berechnet.

Darüber hinaus gibt es v. a. zur Untersuchung des TWT eine **Sonderform von Scheinverbrechen-Experimenten**, die eine schriftliche (vgl. Stern et al., 1981, S. 679) bzw. audiovisuelle Darbietung (vgl. Iacono et al., 1984, S. 291f.) eines kriminellen Szenarios praktizieren, ohne daß die Pbn aktiv an der eigentlichen Tat beteiligt sind. Solche methodischen Ansätze werden teilweise (z. B. Giesen & Rollison, 1980) ebenfalls zu den Analogstudien gezählt (siehe OTA, 1983, S. 73f.).

Analogstudien zum KFT:

Der **Übersicht** von Ben-Shakhar und Furedy (1990, S. 45f.) zufolge erzielt der KFT in Analogstudien Trefferquoten für Schuldige von 50.0% bis 91.7% (mit den Stichprobengrößen gewichteter Durchschnitt: 80.1%) und für Unschuldige von 38.5% bis 85.1% (gewichteter Durchschnitt: 62.9%). Ebenfalls unter Berücksichtigung unentscheidbarer Fälle berichten Raskin et al. (1997, S. 572) gewichtete Mittelwerte der Sensitivität von 77% (Streubreite: 53% – 100%) und der Spezifität von 84% (75% – 90%). Unter Vernachlässigung nicht zuordbarer Ergebnisse schwankt die Sensitivität bei Berning (1992, S. 122) im Bereich zwischen 81.7% und 100% (ungewichteter Durchschnitt: 94.9%). Die Spannweite der Spezifität beträgt dort 50.0% bis 97.4% (Durchschnitt: 77.0%). Nach Steller (1987, S. 49) liegt die Rate valider positiver Befunde zwischen 60.4% und

100% und die Rate valider negativer Klassifikationen zwischen 15.8% und 90.0% (jeweils unter Beachtung unentscheidbarer Fälle).

Damit werden die wesentlichen Ergebnisse der Feldstudien zum KFT gestützt. Auch in Analogstudien ist die **Streuung** der Trefferquoten v. a. für unschuldige Pbn relativ groß, und es zeigt sich (mit Ausnahme der Mittelwerte von Raskin et al., 1997) eine schiefe **Fehlerverteilung** in Richtung einer erhöhten Rate falsch positiver Befunde. Darüber hinaus weichen die Angaben der Übersichtsarbeiten voneinander ab, was auf eine unterschiedliche Auswahl von Studien und deren Teilergebnissen sowie Diskrepanzen bei der Berücksichtigung unentscheidbarer Fälle zurückzuführen ist.

Analogstudien zum TWT:

Dem Mangel an Feldstudien steht eine relativ breite **Laborforschung zum TWT** gegenüber, die sich aber nur zum Teil aus Scheinverbrechen-Experimenten zusammensetzt (alternative Ansätze werden im Abschnitt 2.9.3 diskutiert). Ben-Shakhar und Furedy (1990, S. 52) zitieren zehn Analogstudien mit Trefferquoten zwischen 61.1% und 100% für Schuldige (gewichteter Mittelwert: 83.9%) und 80.6% bis 100% für Unschuldige (gewichteter Mittelwert: 94.2%). Da die Auswertung des TWT (z. B. Lykken-Scoring) im allgemeinen keine unentscheidbaren Ergebnisse vorsieht, bleibt diese Kategorie bei der Bestimmung der Prozentwerte unberücksichtigt. Raskin et al. (1997, S. 573) errechnen anhand von fünf Studien (gewichtete) Durchschnittswerte von 86% valider positiver Befunde für Schuldige (Streuung: 80% – 92%) und 99% valider negativer Befunde für Unschuldige (Streuung: 90% – 100%). Nach Steller (1987, S. 45) resultieren aus sieben Untersuchungen Bandbreiten der Sensitivität von 59.4% bis 95.0% (ungewichteter Durchschnitt: 80.1%) und der Spezifität von 87.5% bis 100% (Durchschnitt: 96.6%). Berning (1992, S. 128) verweist auf die gleichen Experimente wie Steller und kommt zu weitgehend übereinstimmenden Ergebnissen.

Vergleicht man die Analog- und Feldstudien, so lassen sich die Resultate zur Spezifität des TWT gut in Einklang bringen. Beide Forschungsansätze ergeben hohe Trefferquoten für Unschuldige. Für Schuldige dagegen zeigt sich ein anderes Bild. Zwar findet man auch in Scheinverbrechen-Experimenten eine erhöhte Rate falsch negativer Befunde. Diese reicht aber im Durchschnitt nicht an die entsprechenden Fehlerquoten der beiden Feldstudien (24.2% – 50%) heran. Es gibt Grund zu der Annahme, daß die Sensitivität des TWT durch Analogstudien überschätzt wird (s. u.). Generell wird aber der bereits bei den Feldstudien angesprochene Trend der Fehlerverteilung durch die Analogstudien gestützt. Entsprechend den Erwartungen sind Unschuldige beim TWT annä-

hernd perfekt gegen Falschdiagnosen geschützt. Im Gegensatz dazu besteht ein erhöhtes Risiko, daß Schuldige irrtümlich als „glaubwürdig“ klassifiziert werden.

Kritik an Analogstudien:

Die wesentlichen **Vorteile** von Scheinverbrechen-Experimenten sind unmittelbar evident (siehe OTA, 1983, S. 62; Podlesny & Raskin, 1977, S. 782f., 1978, S. 344). Anders als bei Feldstudien liegt ein definitives, objektives Validierungskriterium vor. Die Schuld oder Unschuld der Pbn ist eindeutig über die Gruppenzuordnung determiniert. Ferner bieten Analogstudien die Möglichkeit der experimentellen Kontrolle und Standardisierung. Man kann mehrere unabhängige und abhängige Variablen systematisch untersuchen, die Pbn randomisiert den Gruppen zuordnen, potentielle Störvariablen sowie Effektvermengungen (Konfundierungen) vermeiden und Kausalzusammenhänge aufdecken. Unter dem Gesichtspunkt der Güte experimenteller Untersuchungen ist somit eine hohe *interne Gültigkeit* („internal validity“, vgl. Campbell & Stanley, 1963, S. 175; Cook, Campbell & Peraccio, 1990, S. 500) erzielbar. D. h., bei angemessener Vorgehensweise lassen sich die Variationen der abhängigen Variablen unmittelbar auf die Variationen der unabhängigen Variablen zurückführen.

Das grundlegende **Problem von Analogstudien** liegt darin, realistische Tatumstände und Testsituationen zu simulieren. Dabei muß mit erheblichen Unterschieden zwischen Labor und Feld gerechnet werden (zusammenfassend OTA, 1983, S. 62). Besonders bedenklich sind die Diskrepanzen in den motivationalen und emotionalen Bedingungen (Patrick & Iacono, 1991, S. 229). Für gewöhnlich dürfen die Teilnehmer an Experimenten nicht freiwillig entscheiden, ob und wie sie die Tat begehen wollen. Statt dessen erhalten sie entsprechende Instruktionen. Die Scheinverbrechen sind in der Regel minder schwerwiegend und mit weniger Streß bzw. emotionaler Erregung assoziiert als echte Delikte. Ausgeprägte Ängste oder Schuldgefühle lassen sich auf diese Weise kaum induzieren (vgl. Lykken, 1978, S. 141). Die Vpn haben keine tiefgreifenden Konsequenzen zu erwarten (meistens nur Gewinn oder Verlust einer finanziellen Belohnung). Reale Tests hingegen können sowohl gravierende positive (z. B. Entlastung und Einstellung des Verfahrens) als auch negative Folgen (z. B. Verurteilungen und Bestrafungen) nach sich ziehen. Weitere potentielle Abweichungen bestehen im Hinblick auf die Testdurchführung (individuell vs. standardisiert) sowie hinsichtlich des Ausbildungsstands der Untersucher (Praktiker vs. Forscher) und der physiologischen Meßgeräte (portable Polygraphen vs. Laborausrüstung). Ferner sind die Pbn von Analogstudien (häufig Studierende) nicht unbedingt repräsentativ für die im Feld getestete Population (z. B. bezüglich Alter, Bildungsgrad, Sozialisation und sozioökonomischem Status).

Die **Generalisierbarkeit** experimentell ermittelter Trefferquoten auf die reale Anwendung ist umstritten. Damit wird ein Problem angesprochen, das allgemein unter den Bezeichnungen *externe Validität* („external validity", Campbell & Stanley, 1963, S. 175; Cook et al., 1990, S. 509) bzw. *ökologische Validität* („ecological validity", z. B. Bronfenbrenner, 1981, S. 45f.; Neisser, 1976, S. 33) in die Diskussion um die Vor- und Nachteile von Labor- und Feldforschung eingegangen ist. In bezug auf die psychophysiologische Aussageforschung lassen sich im wesentlichen zwei Standpunkte unterscheiden (vgl. Honts, 1996, S. 310). Zum einen wird die Auffassung vertreten, daß die Ergebnisse von Analogstudien nicht ohne weiteres auf die Praxis transferiert werden können (z. B. Ben-Shakhar & Furedy, 1990, S. 39; Lykken, 1998, S. 84f.). Zum anderen argumentieren Befürworter der Methode, daß die Kluft zwischen Labor und Feld überbrückbar ist, wenn man repräsentative Probandenstichproben heranzieht, realitätsnahe Testbedingungen schafft und die Täuschungsmotivation durch Anreize erhöht (vgl. Raskin, 1989, S. 263f.; Steller, 1987, S. 51).

Die **Problematik der externen Validität** muß für den KFT und den TWT getrennt analysiert werden. Denn es besteht Grund zu der Annahme, daß beim KFT eher emotionale Prozesse eine wichtige Rolle spielen (z. B. Angst vor einem positiven Befund), wohingegen die Funktionsweise des TWT v. a. von kognitiven Faktoren (z. B. Wiedererkennen und besondere Bedeutsamkeit tatbezogener Details) abhängen soll (Lykken, 1988, S. 122f.; vgl. auch Abschnitt 2.10). Folglich können die experimentellen Schätzungen der Trefferquoten von KFT und TWT in unterschiedliche Richtungen verzerrt sein.

Beim **KFT** liegt die Schwierigkeit primär darin, die emotionalen Bedingungen echter Tests laborexperimentell nachzuahmen. Im Normalfall dürfte ein realer Tatvorwurf schwerwiegender sein. Demnach wären auch die Bedrohlichkeit der relevanten Fragen und die Angst vor den Folgen eines positiven Befundes stärker ausgeprägt. Daraus ergeben sich zwei Implikationen (Lykken, 1974, S. 734): Scheinverbrechen-Experimente tendieren dazu, einerseits die Sensitivität des KFT zu unterschätzen und andererseits seine Spezifität zu überschätzen. In der Feldanwendung kann man annehmen, daß sowohl schuldige als auch unschuldige Pbn im erhöhten Maße auf die relevanten Fragen reagieren und somit eher als „unglaubwürdig" eingestuft werden.

Im Gegensatz dazu ist beim **TWT** von einer Überschätzung der Trefferquoten für Schuldige auszugehen (Raskin et al., 1997, S. 573). In Analogstudien werden oft günstige Voraussetzungen geschaffen, die das Wahrnehmen, Lernen und Wiedererkennen tatbezogener Details durch die Täter fördern (z. B. Lenkung der Aufmerksamkeit auf die kritischen Einzelheiten, kurze Zeitspanne zwischen Tat und Test, sorgfältige Aus-

wahl der Items; vgl. auch Elaad, 1990, S. 522). Realfälle hingegen bieten weniger optimale Bedingungen, so daß die Identifikation relevanter Items erschwert und somit die Trefferquoten für Schuldige möglicherweise gesenkt werden. Allerdings könnte die Überschätzung der Sensitivität geringer ausfallen, als zunächst angenommen. Man muß bedenken, daß wiedererkannte Tatdetails bei realen Verbrechen eine potentiell höhere Signifikanz aufweisen als bei Simulationen. Personen, die ein echtes Vergehen begangen haben, sollten demnach noch stärkere Reaktionsunterschiede zwischen den erkannten relevanten und den irrelevanten Items zeigen, wodurch ihre Entdeckbarkeit steigen dürfte. Ungeachtet dessen ist bei angemessener Testkonstruktion und Itemauswahl weder im Feld noch im Labor mit einer besonders hohen Rate falsch negativer Befunde zu rechnen. Die Wahrscheinlichkeit, daß Pbn ohne Tatwissen rein zufällig ein derartiges systematisches Reaktionsprofil aufweisen, ist äußerst gering. Dementsprechend argumentiert Lykken (1988, S. 123), daß man die Ergebnisse solcher Analogstudien eher auf die Feldanwendung übertragen kann, zumal emotionale Faktoren, die ohnehin im Labor nur schwer zu simulieren sind, beim TWT eine verhältnismäßig untergeordnete Rolle spielen.

Aus den vorangegangenen Erörterungen wird deutlich, daß sich die Frage der externen Validität nicht pauschal nach dem Alles-oder-Nichts-Prinzip beantworten läßt. Sie muß differenzierter betrachtet werden. Die eigentliche Fragestellung lautet nämlich nicht ob, sondern *inwiefern* man die Ergebnisse von Analogstudien auf andere Pbn, Situationen, Treatments, Meßvariablen etc. generalisieren kann. Es geht also um den **Grad der Generalisierbarkeit**, wobei durchaus Besonderheiten der Befragungstechniken und die unterschiedliche Realitätsnähe von Scheinverbrechen-Experimenten in Betracht gezogen werden müssen.

Darüber hinaus weist das Problem der Generalisierbarkeit im Rahmen der psychophysiologischen Aussageforschung noch einige **Ungereimtheiten** auf. Furedy und Heslegrave (1991a, S. 174) betonen, daß trotz der offensichtlichen Diskrepanzen zwischen Labor und Feld die Auswirkungen dieser Unterschiede und die daraus resultierenden Schlußfolgerungen für die externe Validität bislang kaum spezifiziert wurden. Die o. g. Annahmen hinsichtlich der Über- und Unterschätzung von Trefferquoten durch Analogstudien sind zunächst rein spekulativ. Eindeutige empirische Belege dafür fehlen bislang. Weitgehend unstrittig ist jedoch, daß Scheinverbrechen-Experimente nicht allein zur Klärung der diagnostischen Güte ausreichen (Furedy & Heslegrave, 1991a, S. 174; Raskin, 1987, S. 189). Im Grunde genommen können nur adäquate Feldstudien die Treffsicherheit der Tests mit hinreichend hoher externer Validität schätzen (Fiedler, 1999, S. 24) bzw. die Generalisierbarkeit entsprechender Laborbefunde unter Beweis stellen (Honts, 1996, S. 311).

Der **eigentliche Nutzen von Analogstudien** liegt in der Möglichkeit, anhand kontrollierter Experimente die Wirkung systematischer Bedingungsvariationen auf die Reaktionsstärken bzw. Trefferquoten zu überprüfen. Eine solche Laborforschung bietet etwa die Gelegenheit, unterschiedliche Befragungstechniken gegenüberzustellen (vgl. Lykken, 1998, S. 146) oder die Rolle von Variablen zu untersuchen, die nicht direkt mit Schuld oder Schuldgefühlen zusammenhängen („extra-guilt variables“, Iacono & Patrick, 1987, S. 469). Beispielsweise kann man neben der Täterschaft zusätzliche unabhängige Variablen (wie etwa Täuschungsmotivation oder Antwortmodus) bzw. quasi-unabhängige Variablen (z. B. Personenvariablen oder Persönlichkeitsmerkmale) berücksichtigen und alternative abhängige Variablen erfassen (z. B. angemessenere kardiovaskuläre Maße als den „relativen Blutdruck“). Daraus lassen sich Rückschlüsse auf die psychologischen und physiologischen Grundlagen der Verfahren ziehen bzw. die Effekte von Faktoren analysieren, die eventuell einen Einfluß auf die Reaktionsunterschiede ausüben.

Zusammenfassend kann man festhalten, daß die Trefferquoten von Scheinverbrechen-Experimenten nicht ohne weiteres auf die Feldanwendung übertragbar sind. Analogstudien eignen sich aber durchaus dazu, die Grundlagen und Einflußfaktoren der psychophysiologischen Aussagebeurteilung zu untersuchen. Sie leisten somit einen sinnvollen Beitrag zur Theoriebildung und Hypothesenprüfung.

2.9.3 Laborexperimente ohne Verbrechenssimulationen

Während Validitätsstudien zum KFT in der Regel an konkrete (Schein-)Verbrechen gebunden sind, gibt es beim TWT andere Möglichkeiten, ein bestimmtes Wissen zu induzieren, das die Pbn anschließend verheimlichen sollen. In derartigen **Laborexperimenten** wird unter anderem mit bestimmten Informationen gearbeitet (z. B. Codewörter), die die Vpn vor dem Test lernen müssen (vgl. Thackray & Orne, 1968, S. 331; Waid et al., 1979, S. 16f.), oder man fragt nach personenbezogenen bzw. autobiographischen Sachverhalten, wie Name, Geburtsdatum, Wohnort etc. (z. B. Cutrow, Parks, Lucas & Thomas, 1972, S. 580; Elaad & Ben-Shakhar, 1989, S. 444; Lykken, 1960, S. 259). Oft kommen auch Karten- oder Zahlentests zum Einsatz, die den Stimulationstests des KFT ähneln (siehe Gustafson & Orne, 1963, S. 409, 1964, S. 384, 1965a, S. 415, 1965b, S. 11; Kugelmass et al., 1967, S. 313). Dabei sollen die Pbn z. B. eine von mehreren Karten verdeckt ziehen oder eine Zahl auswählen und notieren. Anschließend werden sie in Form von Multiple-Choice-Items nach der Karte bzw. Zahl befragt. Stärkere Reaktionen auf die relevanten Items in Relation zu den Vergleichsreizen werden als Treffer gewertet, und man kann daraus die Entdeckungsraten berechnen.

Solche Studien dienen aber weniger der Untersuchung der kriterienbezogenen Validität, sondern der Analyse jener Prozesse, die der psychophysiologischen Aussagebeurteilung zugrunde liegen. Ihr heuristischer Wert wird im Abschnitt 2.10 veranschaulicht.

Eine andere Form laborexperimenteller Grundlagenforschung, die im weitesten Sinne mit der forensischen Psychophysiologie zusammenhängt, ist das sog. „**Differentiation-of-Deception**“**-Paradigma** („Differentiation-of-Deception Paradigm“, DDP; vgl. Furedy et al., 1988, S. 683). Das DDP fußt auf einem wesentlichen Kritikpunkt an der „Lügendetektion“, der bereits im Kontext der Konstruktvalidität des KFT angesprochen wurde (Abschnitt 2.4.2). Die Verfahren der psychophysiologischen Aussagebeurteilung erfassen nicht „Lügen“ an sich. D. h., die beobachteten Reaktionsunterschiede zwischen den relevanten Fragen und Kontrollfragen bzw. den relevanten und irrelevanten Items sind nicht direkt auf die Variation von Täuschung und Aufrichtigkeit zurückzuführen. Unter dem Gesichtspunkt der experimentellen Kontrolle liegen mehrere potentiell konfundierende Variablen vor. Die Fragen- bzw. Itemtypen unterscheiden sich nicht nur im Wahrheitsgehalt der Antworten, sondern beispielsweise auch hinsichtlich ihrer Bedeutsamkeit, ihres emotionalen Gehalts und der Häufigkeit, mit der sie dargeboten werden. Im Gegensatz dazu bietet das DDP die Möglichkeit, die Phänomene Täuschung und Aufrichtigkeit stringent voneinander abzugrenzen und die daran beteiligten psychologischen und physiologischen Prozesse zu untersuchen. Ein wesentliches Charakteristikum des experimentellen Paradigmas ist die intraindividuelle Variation zweier Bedingungen, die sich im Prinzip nur hinsichtlich des Wahrheitsgehalts unterscheiden. Dazu werden mehrere inhaltlich parallelisierte Fragenpaare dargeboten (z. B. autobiographischer Art: „Wie alt ist Ihre Mutter?“ vs. „Wie alt ist Ihr Vater?“). Die Pbn sollen jeweils eine Frage wahrheitsgemäß und die andere wahrheitswidrig beantworten. Andere potentiell konfundierende Variablen können kontrolliert werden, so daß quantitative Unterschiede in den gemessenen autonomen Erregungsmaßen im wesentlichen auf der Variation von Täuschung und Aufrichtigkeit basieren. Auf diese Weise will man feststellen, ob „Lügen“ und „Wahrheitsagen“ mit unterschiedlich starken körperlichen Reaktionen einhergehen, und wenn ja, welche Mechanismen dafür verantwortlich sind (z. B. kognitive, emotionale und motivationale Prozesse, sympathische vs. parasympathische Regulation). Eine psychophysiologische Differenzierung von Täuschung und Aufrichtigkeit gilt als nachgewiesen, wenn signifikante Reaktionsdifferenzen auftreten. Durch die systematische Manipulation zusätzlicher unabhängiger Variablen läßt sich die Wirkung derjenigen Faktoren untersuchen, die möglicherweise einen moderierenden Einfluß auf diese Reaktionsunterschiede ausüben.

Die Forschungsgruppe um Furedy fand in mehreren Experimenten ein sog. „**elektrodermales DDP-Phänomen**“, d. h. im Durchschnitt höhere Amplituden der Hautleitfä-

higkeitsreaktionen unter der Bedingung Täuschung. Die Mehrzahl der zusätzlichen Bedingungsvariationen, die von Furedy und Mitarbeitern untersucht wurden, zeigten keine signifikanten Interaktionen mit dem Wahrheitsgehalt. Dies deutete darauf hin, daß die entsprechenden Faktoren keinen bedeutenden Einfluß auf die Stärke des DDP-Phänomens ausübten. Es resultierten weder signifikante Geschlechtseffekte (Furedy et al., 1988, S. 686; Furedy, Posner & Vincent, 1991, S. 95; Hemsley, Heslegrave & Furedy, 1980, S. 287) noch relevante Auswirkungen situationsbezogener Probandenvariablen bzw. Faktoren des Untersuchungskontextes auf das DDP-Phänomen (z. B. Annahmen über die Treffsicherheit der „Lügendetektion“ und subjektiv eingeschätzte Bedeutung der Gedächtnisleistung, Furedy et al., 1991, S. 96; Täuschungsmotivation, Vincent & Furedy, 1992, S. 134; Wahlfreiheit hinsichtlich der wahrheitswidrigen Antworten, Furedy, Gigliotti & Ben-Shakhar, 1994, S. 18). Die These, wonach man das DDP-Phänomen auf eine Konfundierung des Wahrheitsgehalts mit der unterschiedlichen Erinnerungsschwierigkeit bzw. Neuheit wahrheitswidriger und wahrheitsgemäßer Antworten zurückführen könnte, fand ebenfalls keine Bestätigung (Vincent & Furedy, 1992, S. 134). Dagegen deutete eine signifikante Interaktion mit der relativen Häufigkeit wahrheitswidriger Trials darauf hin, daß die Reaktionsdifferenz stärker ausfiel, wenn mehr aufrichtige Antworten gegeben wurden als Täuschungen (Furedy et al., 1994, S. 19).

Gödert, Rill und Vossel (2001) gelang eine **Replikation** des elektrodermalen DDP-Phänomens unter relativ schwach ich-involvierenden Bedingungen bei gleichzeitig verbesserter Kontrolle von Gedächtniseffekten (schriftliche Darbietung einfacher, geschlossener Wissensfragen mit einem geringen emotionalen Gehalt bzw. autobiographischen Bezug). Darüber hinaus resultierten Reaktionsunterschiede zwischen Täuschung und Aufrichtigkeit in kardiovaskulären (phasische Herzschlagfrequenz) und subjektiven Variablen (Selbsteinstufungen der Entspannung, Konzentration und Gelassenheit). Die Gruppenfaktoren elektrodermale Labilität (elektrodermal labile vs. stabile Pbn) und Antwortmodus (verbal vs. Tastendruck) zeigten keine signifikanten Effekte auf die jeweilige Stärke der Reaktionsdifferenz.

Dionisio, Granholm, Hillix und Perrine (2001) replizierten das DDP mit **Pupillenreaktionen** als abhängige Variable. Unter der Bedingung Täuschung fanden sie im Durchschnitt eine stärkere Pupillendilatation als unter der Bedingung Aufrichtigkeit. Der Effekt war unabhängig vom inhaltlichen Bezug der Fragen (semantische vs. episodische Gedächtnisinhalte) und wurde von den Autoren im Sinne einer erhöhten mentalen Beanspruchung beim Generieren wahrheitswidriger Antworten interpretiert.

Insgesamt hat sich das DDP-Phänomen als replizierbar und robust gegenüber einer Vielzahl von Bedingungsvariationen erwiesen. Die Implikationen des Paradigmas für die psychophysiologische Aussagebeurteilung sind zwar relativ begrenzt, da das DDP weder individualdiagnostische noch anwendungsbezogene Fragestellungen untersucht. Allerdings kann eine Ausweitung der Forschung auf diesem Gebiet durch die Berücksichtigung zusätzlicher abhängiger Variablen (z. B. zentralnervöser Indikatoren) und unabhängiger Variablen (z. B. Variation der emotionalen Bedeutsamkeit der Fragen) zum besseren Verständnis der Phänomene Täuschung und Aufrichtigkeit sowie ihrer psychophysiologischen Korrelate beitragen (Vossel & Zimmer, 1998, S. 199). Das langfristige Ziel liegt in der Entwicklung und Überprüfung eines geeigneten Erklärungsmodells für die beobachteten Reaktionsunterschiede.

2.9.4 Fazit zur psychophysiologischen Aussageforschung

Trotz der vielfältigen Bemühungen muß man konstatieren, daß die bisherige psychophysiologische Aussageforschung **keine definitiven Schlußfolgerungen** über die Treffsicherheit der Verfahren in der Realanwendung zuläßt (vgl. Ben-Shakhar & Furedy, 1990, S. 32; Iacono & Patrick, 1988, S. 233). Die meisten Feldstudien werden den an sie gerichteten methodischen Anforderungen nicht gerecht, und die Resultate von Analogstudien lassen sich aufgrund mangelnder externer Validität nicht direkt auf die Praxis übertragen. Das Problem der kriterienbezogenen Validität bleibt weiterhin ungelöst. Eine simple Intensivierung der gegenwärtigen Forschung reicht jedoch nicht aus; statt dessen sind alternative Ansätze vonnöten (Iacono, 1991, S. 202, 2000, S. 789).

Strenggenommen können nur **Feldstudien** valide Schätzungen der Treffsicherheit bieten. Die Gültigkeit und Generalisierbarkeit ihrer Ergebnisse sind jedoch nicht von vornherein gewährleistet, sondern an bestimmte Bedingungen geknüpft, die im Gros der bislang vorgelegten Arbeiten nicht erfüllt waren. Darum besteht weiterhin Bedarf an adäquaten Untersuchungen, die v. a. im Hinblick auf die Validierungskriterien und die Auswahl der Fälle besondere Ansprüche erfüllen, um eine Kriteriumskontamination bzw. verzerrte Stichprobenselektion zu vermeiden (für entsprechende Vorschläge siehe Honts, 1996, S. 311f.; Iacono, 1991, S. 206; Lykken, 1988, S. 116; Patrick & Iacono, 1991, S. 237f.; Raskin, 1987, S. 390ff.; Raskin et al., 1997, S. 574). Auf der Ebene der **Laborforschung** müssen verstärkt Anstrengungen unternommen werden, die theoretische Aufarbeitung der forensischen Psychophysiologie voranzutreiben, ihre psychologischen und physiologischen Grundlagen zu eruieren und jene Faktoren zu analysieren, die eventuell einen Einfluß auf die Treffsicherheit ausüben. Ferner sollte man die Verbesserungsmöglichkeiten bestehender Verfahren ausloten und deren inhärente Probleme

durch die Entwicklung neuer Befragungstechniken beheben. Dazu können auch angemessene Analogstudien einen wichtigen Beitrag leisten (vgl. Podlesny & Raskin, 1977, S. 783f.). Außerdem sind die Voraussetzungen für die Übertragbarkeit experimenteller Befunde auf die Feldanwendung zu klären. Da die Methoden der psychophysiologischen Aussageforschung jeweils ihre spezifischen Vor- und Nachteile aufweisen, scheint der **kombinierte Einsatz beider Ansätze** besonders erfolgversprechend (Podlesny & Raskin, 1978, S. 359; Raskin et al., 1997, S. 571). Auf diese Weise können sich Labor- und Feldforschung gegenseitig Anstöße bieten und sich sinnvoll ergänzen: „Laboratory and field studies should be used hand in hand“ (Iacono & Patrick, 1987, S. 469).

2.10 Theorien zur psychophysiologischen Aussagebeurteilung

Die systematische **theoretische Aufarbeitung** der forensischen Psychophysiologie wurde bislang v. a. auf seiten der direkten Verfahren vernachlässigt. Dennoch liegen neben einigen eher naiv-psychologischen Vorstellungen (z. B. „deception responses“, Reid & Inbau, 1977, S. 61ff.; „anticlimax dampening concept“, Backster, 1974, zitiert nach Lykken, 1998, S. 123) auch wissenschaftlich fundiertere Annahmen vor, die zu einer Erklärung der erwarteten Reaktionsunterschiede beitragen. Das folgende Kapitel faßt die wichtigsten Ansätze zusammen. Es sei aber betont, daß die Zielsetzung der eigenen Studie nicht darin besteht, einzelne Theorien gegenüberzustellen und zu überprüfen. Dazu sind die bisherigen Entwürfe noch zu wenig ausgereift. Keiner davon ist in der Lage, die Komplexität der empirischen Befundlage auch nur annähernd umfassend abzudecken. Gleichzeitig gilt es aber festzuhalten, daß sich die psychophysiologische Aussageforschung nicht gänzlich ohne konkrete Modellvorstellungen, abseits jeglicher psychophysiologischer Konzepte entwickelt hat. Von Beginn an gab es in dieser Disziplin mehr oder weniger elaborierte Ideen, weshalb Schuldige und Unschuldige auf bestimmte Reize unterschiedlich reagieren sollen. Ben-Shakhar und Furedy (1990, S. 101) teilen die Ansätze in **zwei große Gruppen** ein: (1) Theorien, die v. a. **motivationale und emotionale Faktoren** in den Vordergrund stellen, und (2) Theorien, die eher auf **kognitiven Prozessen** beruhen. Die folgenden Abschnitte orientieren sich ebenfalls an dieser Kategorisierung, wenngleich eine solch strikte Trennung in vielen Fällen nicht konsequent durchzuhalten ist, denn: „Cognition and emotion are intimately intermingled with each other“ (Lindsay & Norman, 1972, S. 637).

2.10.1 Motivational-emotionale Ansätze

Ein früher Versuch der Systematisierung stammt von Davis (1961, S. 161ff.). Er dokumentierte **drei ältere Auffassungen**:

1. Die **Theorie der konditionierten Reaktion** betont die *Antezedenzien* des Tests (z. B. Tatbegehung). Sie geht davon aus, daß die mit dem Verbrechen zusammenhängenden Stimuli klassisch konditioniert werden. Wenn man die bedingten Reize später in Form von relevanten Fragen bzw. Items präsentiert, wird die emotionale Erregung, die während des Verbrechens vorlag, als konditionierte Reaktion wieder ausgelöst. Im Kontext realer, deliktbezogener Untersuchungen erscheint dieser Ansatz durchaus plausibel. Man kann damit jedoch die hohen Entdeckungsraten in Laborexperimenten mit Karten- bzw. Zahlentests nur schlecht begründen. Im Normalfall ist kaum anzunehmen, daß beim Ziehen einer Karte bzw. beim Auswählen einer Zahl eine besonders starke autonome Erregung vorliegt.

2. Die **Theorie der Angst vor Strafe** beruht auf der Hypothese, daß unglaubwürdige Personen erhöhte relevante Reaktionen zeigen, weil sie Angst vor den negativen *Konsequenzen* haben, die sich aus der Entdeckung ihrer Täuschungsversuche ergeben können (z. B. Verurteilung und Strafe). Wiederum eignet sich diese Theorie insbesondere zur Erklärung von Tests unter Realbedingungen. Experimente, die trotz fehlender negativer Folgen relativ hohe Entdeckungsraten erzielten (z. B. Gustafson & Orne, 1964, S. 385f.; Kugelmass et al., 1967, S. 314) bzw. in denen die Variation der wahrgenommenen Konsequenzen keine signifikanten Effekte auf die Testergebnisse zeigten (z. B. Bradley & Janisse, 1981, S. 312; Kugelmass & Lieblich, 1966, S. 213f.), sprechen gegen die Allgemeingültigkeit des Modells.

3. Gemäß der **Theorie des Konflikts** gerät man beim Lügen in einen Konflikt zwischen zwei simultan ausgelösten, inkompatiblen Verhaltenstendenzen, nämlich entweder die Wahrheit oder die Unwahrheit zu sagen. Dieser Zustand soll mit einer erhöhten emotionalen Erregung einhergehen und stärkere autonome Reaktionen auf die wahrheitswidrig beantworteten Fragen bzw. Items bewirken. Die konflikttheoretischen Annahmen lassen sich gut vereinbaren mit Analogstudien zum TWT und GAT (Bradley, MacLaren & Carle, 1996, S. 157) sowie Kartentest-Experimenten (Elaad & Ben-Shakhar, 1989, S. 448; Furedy & Ben-Shakhar, 1991, S. 167f.; Gustafson & Orne, 1965b, S. 12; Horneman & O'Gorman, 1985, S. 332), in denen der Antwortmodus variiert wurde (z. B. Verneinung, Bejahung, Itemwiederholung oder Schweigen). Dabei resultierten die höchsten Identifikationsraten bzw. Reaktionsunterschiede zwischen relevanten und irrelevanten Items, wenn die Pbn diese verneinen mußten und somit auf die relevanten

Items logen. Die Theorie kann aber nicht erklären, warum auch die Trefferquoten für die alternativen Antwortarten meist oberhalb des Zufallsniveaus lagen (vgl. zudem Janisse & Bradley, 1980, S. 749f., und Kugelmass et al., 1967, S. 314, die gleichermaßen hohe Trefferquoten für Verneinung und andere Antwortmodi fanden). Denn durch die Bejahung werden die relevanten Stimuli wahrheitsgemäß und die Vergleichsstimuli wahrheitswidrig beantwortet, und bei der Itemwiederholung bzw. beim Schweigen liegen keine Variationen des Wahrheitsgehalts und somit keine Konflikte vor. Insofern kann der verbale Akt des Lügens zu einer erhöhten autonomen Erregung beitragen, er ist aber keine notwendige Bedingung für stärkere relevante Reaktionen.

Mit Rekurs auf die Theorie der Angst vor Strafe unterstrichen Gustafson und Orne (1963, 1965a) die **Rolle motivationaler Prozesse**. Demnach soll eine hohe Täuschungsmotivation, d. h. die Absicht, bei der psychophysiologischen Aussagebeurteilung möglichst glaubwürdig und ehrlich zu erscheinen, zu starken relevanten Reaktionen und somit zu einer höheren Entdeckbarkeit führen. Experimente, in denen die Täuschungsmotivation der Pbn variiert wurde, stützten diese Hypothese (Elaad & Ben-Shakhar, 1989, S. 446f., 1997, S. 595f.; Gustafson & Orne, 1963, S. 409f.). Gustafson und Orne (1965a, S. 415ff.) erweiterten ihren Ansatz, indem sie zeigten, daß sowohl die Motivation, unentdeckt zu bleiben, als auch die Motivation, entdeckt zu werden, mit hohen Trefferquoten in Kartentests einhergingen. Nach dem ersten Durchgang gaben sie ihren Vpn fingierte Ergebnisrückmeldungen und führten dann einen zweiten Test durch. Wenn man den Pbn ein Resultat mitteilte, das ihrer motivationalen Ausrichtung entsprach (z. B. negativer Befund, falls die Pbn motiviert waren, unentdeckt zu bleiben), sank die Entdeckungsrate im anschließenden Kartentest wieder. Dies konnte dadurch erklärt werden, daß aufgrund der Zielerreichung im ersten Test die Motivation der Pbn abgeklungen war. Nach einem zielinkongruenten Feedback hingegen blieb die Trefferquote des zweiten Tests auf einem hohen Niveau, was darauf zurückzuführen war, daß die Motivation unter diesen Bedingungen aufrechterhalten wurde (vgl. auch Ben-Shakhar & Furedy, 1990, S. 105). Elaad und Ben-Shakhar (1989, S. 447) versuchten, die Wirkung einer erhöhten Motivation dadurch zu erklären, daß die Aufmerksamkeit gesteigert und infolgedessen das Ignorieren relevanter Stimuli erschwert wird. Allerdings konnten nicht alle Studien entsprechende Motivationseffekte nachweisen (z. B. Davidson, 1968, S. 63; Furedy & Ben-Shakhar, 1991, S. 167f.; Lieblich, Naftali, Shmueli & Kugelmass, 1974, S. 114f.). Und in einigen der Experimente wurden auch unter schwach ich-involvierenden Bedingungen Trefferquoten erzielt, die oberhalb des Zufallsniveaus lagen (Elaad & Ben-Shakhar, 1989, S. 449; Lieblich et al., 1974, 114f.). Dies deutet darauf hin, daß – ähnlich wie das Lügen bei der Konflikttheorie – eine ausgeprägte Motivation zu stärkeren Reaktionsunterschieden und somit zu einer

erhöhten Entdeckbarkeit führen kann. Sie stellt aber wiederum keine unabdingbare Voraussetzung für überzufällige Trefferquoten dar.

Raskin (1979) schlug ein Erklärungsmodell vor, das im Grenzbereich der Trennung zwischen motivational-emotionalen und kognitiven Ansätzen anzusiedeln ist. Analog zu Gustafson und Orne ging auch er davon aus, daß motivationale Aspekte bei der psychophysiologischen Aussagebeurteilung wichtig sind. Darüber hinaus verwies Raskin auf die Unterscheidung zwischen Orientierungsreaktion (OR) und Defensivreaktion (DR) sowie die Effekte der Reizbedeutung auf die Reaktionsstärke (vgl. Überblick von Baltissen & Sartory, 1998; Cook & Turpin, 1997). Seiner Ansicht nach bewirkt die Intention schuldiger und unschuldiger Pbn, einen negativen Befund zu erzielen, daß die verschiedenen Fragen- bzw. Itemtypen eine differentielle Bedeutsamkeit und somit einen unterschiedlichen **Signalwert** („signal value“, Berlyne, 1960, S. 87) erhalten. Je ausgeprägter der Signalwert der Reize, desto stärkere ORs oder sogar DRs werden ausgelöst. Beim **KFT** ist der unterschiedliche Signalwert v. a. auf die Suggestionen und Manipulationen im Vortest-Interview zurückzuführen. Für Unschuldige sollen die Kontrollfragen besonders bedeutsam sein und für Schuldige die relevanten Fragen. Beim **TWT** haben die relevanten Items nur für Personen mit Tatwissen eine erhöhte Signifikanz. Für Tatunbeteiligte hingegen weisen die kritischen und irrelevanten Alternativen einen äquivalenten Signalwert auf. Da außerdem die Stimuli der direkten Verfahren persönlicher, bedrohlicher und somit emotional erregender sein sollen als die der indirekten Befragungstechniken, erwartete Raskin (1979, S. 592) beim KFT vermehrt DRs und beim TWT eher ORs. Als Beleg für seine Annahmen präsentierte er elektrodermale und kardiovaskuläre Daten, die aus Analogstudien (Podlesny & Raskin, 1978, S. 354f.; Raskin & Hare, 1978, S. 133) und realen Felduntersuchungen stammen. Diese würden dafür sprechen, daß Schuldige auf die relevanten Fragen des KFT vorwiegend mit DRs reagieren, wohingegen sie beim TWT lediglich stärkere ORs auf die relevanten Items zeigen.[5] Allerdings sind solche traditionellen Auffassungen – wonach z. B. ein Anstieg der Herzschlagfrequenz (Akzeleration) nach Reizdarbietung als zuverlässiger Indikator einer DR gilt, während eine initiale Abnahme (Dezeleration) eine OR indiziert (Graham & Clifton, 1966, S. 316f.) – inzwischen mit Skepsis zu beurteilen (vgl. Barry & Maltzman, 1985, S. 26f.; zusammenfassend Baltissen & Sartory, 1998, S. 17ff.). Außerdem wurde versäumt, den theoretischen Ansatz von Raskin weiter auszubauen und einer dezidierten Überprüfung zu unterziehen.

[5] Es sei vermerkt, daß diese Interpretation teilweise mit der von Raskin und Hare (1978, S. 135) sowie Podlesny und Raskin (1978, S. 358) konfligiert, die ihre kardiovaskulären Daten nur im Sinne einer stärkeren Aufmerksamkeitsreaktion Schuldiger nach Präsentation der relevanten Fragen deuteten.

2.10.2 Kognitive Ansätze

Konzepte wie **Wissen**, **Aufmerksamkeit** und **Informationsverarbeitung**, die bereits im Rahmen der motivational-emotionalen Theorien eine gewisse Rolle gespielt haben, stehen bei den kognitiven Ansätzen im Vordergrund.

Nach Lykken (1988, S. 122f.) basiert der TWT eher auf kognitiven als auf emotionalen Prozessen. Entscheidend ist die Fähigkeit, die relevanten Items zu identifizieren. Darin unterscheiden sich Personen mit und ohne Tatwissen. Das **Konzept des Tatwissens** geht davon aus, daß allein die Kenntnis tatbezogener Details für Reaktionsunterschiede zwischen den beiden Itemtypen ausreicht (Lykken, 1959, S. 385, 1960, S. 258). Zur Erklärung greift Lykken (1974, S. 728) ebenfalls auf Annahmen zur Orientierungsreaktion zurück: Im Prinzip kann jedes Item des TWT eine OR auslösen. Dies gilt sowohl für schuldige als auch für unschuldige Pbn. Für Personen mit Tatwissen haben aber die relevanten Items eine erhöhte Signifikanz, d. h. einen zusätzlichen Signalwert. Solche besonders bedeutsamen Stimuli lösen stärkere ORs aus, die sich v. a. in der phasischen elektrodermalen Aktivität manifestieren (Hautwiderstands- [SRR] und Hautleitfähigkeitsreaktionen [SCR], die als valideste OR-Indikatoren gelten; Barry, 1984, S. 131). Für Personen ohne Tatwissen sind alle Alternativen äquivalent, d. h., es werden keine systematischen Reaktionsunterschiede zwischen relevanten und irrelevanten Items erwartet. Unter der Voraussetzung, daß der TWT im wesentlichen auf Aufmerksamkeitsprozessen beruht, wären die Reaktionsunterschiede weitgehend unabhängig von emotionalen oder motivationalen Faktoren wie Lügen, Täuschungsmotivation oder Angst vor Bestrafung. Im Gegensatz dazu legt aber eine Reihe von Untersuchungen zum TWT nahe, daß derartige Faktoren auch bei den indirekten Verfahren der psychophysiologischen Aussagebeurteilung zu berücksichtigen sind (vgl. die o. g. Analogstudien und Kartentest-Experimente).

Die von Ben-Shakhar und Mitarbeitern formulierte „**Dichotomisierungstheorie**“ („dichotomization theory“, z. B. Ben-Shakhar, 1977, S. 409; Ben-Shakhar, Lieblich & Kugelmass, 1982, S. 112) beruft sich auf die Habituation der OR (Abnahme der Reaktionsstärke bei wiederholter Darbietung des auslösenden Reizes, vgl. Fröhlich, 2000, S. 214f.). Sie postuliert, daß Pbn mit Tatwissen die Items des TWT nur nach zwei separaten Kategorien kognitiv verarbeiten, nämlich ob es sich um eine relevante oder irrelevante Alternative handelt. Alle anderen Aspekte der Stimuli werden weitgehend vernachlässigt. Im Sinne von Sokolovs Reiz-Vergleichstheorie (z. B. Sokolov, 1963, S. 286ff., 1966, S. 347ff., 1969, S. 673ff.) könnte man auch sagen, daß pro Kategorie jeweils nur ein einziges neuronales Modell gebildet wird. Ferner sollen die durch die Items ausgelösten ORs und deren psychophysiologischen Begleitreaktionen interkate-

gorial unabhängig sein, so daß die Habituation nur innerhalb der Kategorien generalisiert. Es findet also keine Generalisierung der Habituation über die Kategorien hinweg statt. Somit wäre das systematische Reaktionsmuster der Pbn mit Tatwissen (stärkere relevante Reaktionen) darauf zurückzuführen, daß der TWT mehr irrelevante als relevante Items enthält (z. B. im Verhältnis 5:1). Aufgrund der höheren Darbietungshäufigkeit erfolgt bei den neutralen Items eine schnellere Habituation, und die ORs sind insgesamt schwächer als bei den kritischen Details. Pbn ohne Tatwissen können hingegen nicht unterscheiden, ob es sich um eine tatbezogene oder nicht tatbezogene Alternative handelt, so daß die entsprechenden Reaktionen gleichermaßen von der Habituation betroffen sind. Der wesentliche Unterschied gegenüber Lykkens Ansatz besteht darin, daß die Dichotomisierungstheorie nicht von einem besonders ausgeprägten Signalwert bestimmter Reize für Schuldige ausgeht. Allein die unterschiedlichen relativen Häufigkeiten der beiden Itemtypen sollen für die Reaktionsdifferenzen verantwortlich sein.

Einige der aus dem Dichotomisierungsansatz ableitbaren **Hypothesen** konnten experimentell bestätigt werden (vorwiegend Kartentests mit SCR-Amplitude als physiologische Variable). So ist etwa der Reaktionsunterschied zwischen den beiden Itemtypen um so größer, je weniger relevante Items in Relation zu den irrelevanten dargeboten werden (Ben-Shakhar, 1977, S. 412, 1980, S. 530; Ben-Shakhar, Lieblich & Kugelmass, 1975, S. 287, 1982, S. 113; Lieblich, Kugelmass & Ben-Shakhar, 1970, S. 605f.). Falls man das zahlenmäßige Verhältnis der Itemtypen umkehrt und mehr kritische als neutrale präsentiert, dann resultieren stärkere Reaktionen auf die irrelevanten Stimuli. Entgegen den Erwartungen fällt aber der Betrag der Reaktionsdifferenz deutlich kleiner aus als bei Überwiegen der relevanten Items (Ben-Shakhar, 1977, S. 412). Nach der Dichotomisierungstheorie generalisiert die Habituation nur intra-, nicht jedoch interkategorial. Demnach dürfte die Reaktionsstärke auf ein bestimmtes Item primär durch dessen serielle Position innerhalb seiner Kategorie determiniert sein. Wenn es sich z. B. bei der vierten, neunten und sechzehnten Alternative eines TWT um das erste, zweite und dritte relevante Item handelt, dann sollten die dadurch ausgelösten Reaktionen in etwa den ORs auf die ersten drei Stimuli eines Habituationsverlaufs entsprechen. Diese Annahme wurde von Ben-Shakhar (1980, S. 532) empirisch gestützt. Außerdem sind die Entdeckungsraten bzw. Reaktionsunterschiede unabhängig davon, ob man verschiedene irrelevante Items darbietet oder stets das gleiche wiederholt; Hauptsache, die relativen Häufigkeiten sind in beiden Fällen identisch (vgl. Ben-Shakhar, 1977, S. 411f.). Falls ein TWT nur ein einziges relevantes Item beinhaltet, sollte die Reaktion darauf unabhängig von der Anzahl der vorhergehenden irrelevanten Items sein. Diesbezüglich liegen zwei Experimente mit widersprüchlichen Ergebnissen vor. Ben-Shakhar und Lieblich (1982, S. 280) konnten die Hypothese nicht bestätigen. Dort war die Reaktion auf das relevante Item um so stärker, je weniger irrelevante Items vorher präsentiert

wurden. Ben-Shakhar, Asher, Poznansky-Levy, Asherowitz und Lieblich (1989, S. 36) hingegen fanden keinen Unterschied zwischen einer frühen oder späten Darbietung des kritischen Stimulus. Dies versuchten sie dadurch zu erklären, daß es zu einer Generalisierung der Habituation über die Reizkategorien hinweg kommen kann, wenn sich die relevanten und irrelevanten Items relativ ähnlich sind (wie bei Ben-Shakhar & Lieblich, 1982).

Insgesamt eignet sich der Dichotomisierungsansatz zur Interpretation einiger Befunde, die mit den anderen Theorien nicht vereinbar sind, beispielsweise daß bei Überwiegen der relevanten Items stärkere Reaktionen auf die irrelevanten auftreten (vgl. Ben-Shakhar, 1977). Der Reaktionsunterschied fällt aber schwächer aus, als wenn im umgekehrten Verhältnis mehr relevante als irrelevante Stimuli dargeboten werden. Dies deutet darauf hin, daß zusätzliche Faktoren – wie etwa der potentielle Signalwert der relevanten Items – zu einer Reaktionserhöhung beitragen. Außerdem lassen sich ähnlich wie bei Lykkens Ansatz die Effekte emotionaler und motivationaler Variablen nicht erklären. Darum sieht Ben-Shakhar (1977, S. 412f.) in der Dichotomisierungstheorie weniger ein Gegenmodell, sondern eher eine Ergänzung zu den übrigen Erklärungsansätzen. So ist etwa Tatwissen eine notwendige Voraussetzung für die Dichotomisierung. Und dieser Prozeß unterliegt möglicherweise motivationalen und emotionalen Einflüssen, z. B. indem die kategoriale Verarbeitung der Items durch eine hohe Täuschungsmotivation unterstützt wird oder wahrheitswidrige Antworten die Reaktionen erhöhen. Wie diese Faktoren zusammenwirken, ist jedoch bislang ungeklärt (vgl. auch Ben-Shakhar & Furedy, 1990, S. 113).

Die o. g. kognitiven Ansätze weisen ein gemeinsames Manko auf. Beide beziehen sich nur auf den TWT und gestatten keine Analyse der direkten Verfahren. Waid und Mitarbeiter (Waid & Orne, 1981, S. 80; Waid, Orne, Cook & Orne, 1978, S. 728; Waid, Orne & Orne, 1981, S. 230) haben ein Modell vorgeschlagen, das auch auf den KFT anwendbar ist. Die Grundlage bildeten Befunde zur Psychophysiologie von **Aufmerksamkeits- und Gedächtnisprozessen**, wonach inzidentell gelernte und danach wiedererinnerte Wörter während ihrer ursprünglichen Darbietung stärkere elektrodermale Reaktionen auslösten als Wörter, die später nicht reproduziert werden konnten (Corteen, 1969, S. 83; ähnliche Ergebnisse berichteten auch McLean, 1969, S. 59, und Sampson, 1969, S. 222). Demzufolge ist anzunehmen, daß eine selektive Aufmerksamkeitszuwendung auf bestimmte Stimuli eine erhöhte autonome Erregung und eine bessere Behaltensleistung bewirkt. Diese Vorstellungen lassen sich auf die psychophysiologische Aussagebeurteilung übertragen. So beobachteten etwa Waid et al. (1978, S. 732) und Waid, Orne und Orne (1981, S. 227f.), daß diejenigen Fragen oder Items, die mit einer erhöhten elektrodermalen Reaktionsstärke bzw. Reaktionswahrscheinlichkeit einhergingen,

nach den Tests von den Pbn besser erinnert wurden. Und je mehr relevante als Kontrollfragen die schuldigen Pbn nach einem KFT reproduzieren konnten, desto größer waren die Identifikationsraten und somit die elektrodermalen Reaktionsunterschiede zwischen den beiden Fragentypen (Waid, Orne & Orne, 1981, S. 229). Unabhängig davon resultierte auch für die Schuldigen im TWT-Experiment von Iacono et al. (1984, S. 295) eine signifikante positive Korrelation zwischen der Reaktionsdifferenz und der Anzahl der wiedererinnerten relevanten Items.

Die Annahmen der Forschungsgruppe um Waid weisen deutliche Überschneidungen zum Konzept der „**Verarbeitungstiefe**“ auf („depth/levels of processing“, Craik & Lockhart, 1972, S. 675ff.). Laut diesem Gedächtnismodell kann man z. B. sprachliche Stimuli auf mehreren hierarchisch gegliederten Ebenen kodieren. Bei der sog. „flachen“ Verarbeitung werden v. a. physikalische und sensorische Reizmerkmale erfaßt. Auf den mittleren Ebenen steht unter anderem die Erkennung von Mustern und phonemischen Merkmalen im Zentrum der Analyse. Auf der tiefsten Ebene wird die Reizbedeutung berücksichtigt. Ferner soll sich die Behaltensleistung mit zunehmender Verarbeitungstiefe verbessern (zur empirischen Überprüfung siehe Craik & Tulving, 1975). Bezogen auf die psychophysiologische Aussagebeurteilung ist anzunehmen, daß schuldige Pbn die relevanten Fragen bzw. Items elaborierter verarbeiten, aufgrund der besonderen Bedeutsamkeit der tatbezogenen Stimuli für diese Personengruppe. Im Gegensatz dazu sollen Unschuldige die irrelevanten Items oder Kontrollfragen gleichermaßen bzw. tiefer kodieren.

Neben den grundsätzlichen Problemen des Ansatzes von Craik und Lockhart (zusammenfassend Baddeley, 1997, S. 117f.) wirft die Theorie von Waid und Mitarbeitern jedoch zusätzliche **Interpretationsschwierigkeiten** auf. Die Ergebnisse ihrer Studien sind korrelativer Natur und lassen somit keine Kausalschlüsse zu. Die Zusammenhänge zwischen Reizverarbeitung, Reaktionsstärke und Gedächtnisleistung könnten auch durch andere unbestimmte Faktoren (z. B. emotionaler Art, vgl. Waid, Orne & Orne, 1981, S. 230) vermittelt werden. Darum schlugen Ben-Shakhar und Furedy (1990, S. 110) vor, die Aufmerksamkeitsausrichtung systematisch zu manipulieren, um deren Auswirkungen auf die Reaktionsstärke zu untersuchen. Schumacher (1993, S. 112f.) variierte die Aufmerksamkeit der Pbn während eines TWT. Es resultierten jedoch keine signifikanten Unterschiede zwischen den Lykken-Scores von Schuldigen, die ihre Konzentration auf die Items richten oder davon ablenken sollten bzw. keine diesbezüglichen Instruktionen erhielten. Die drei Gruppen unterschieden sich jedoch erwartungsgemäß in der Anzahl der wiedererinnerten Items. D. h., der Zusammenhang zwischen Aufmerksamkeit und Entdeckbarkeit konnte nicht repliziert werden. Die Frage, weshalb diese Diskrepanzen zu Waid et al. auftraten, wurde von Schumacher (1993, S. 133)

offen gelassen. Somit bleibt unklar, ob seine Ergebnisse auf Schwächen der Theorie oder auf Probleme der Studie zurückzuführen sind. Dabei ist auch zu berücksichtigen, daß in Schumachers Untersuchung die Sensitivität des TWT in allen drei Experimentalbedingungen außergewöhnlich niedrig lag.

2.10.3 Fazit zu den Theorien

Keiner der geschilderten theoretischen Ansätze kann für sich allein genommen eine umfassende Gültigkeit beanspruchen. Sie werfen vielmehr nur Schlaglichter auf einzelne Aspekte der psychophysiologischen Aussagebeurteilung. Ein Hauptproblem der motivational-emotionalen Konzeptionen besteht darin, daß man auch unter entsprechend schwach involvierenden Bedingungen hohe Trefferquoten erzielen kann. Solche Befunde sind eher mit kognitiv orientierten Modellvorstellungen vereinbar, die jedoch die nachgewiesenen Einflüsse emotionaler bzw. motivationaler Variablen nicht ohne weiteres erklären können. Es besteht Grund zur Annahme, daß die beobachteten Reaktionsunterschiede von einer Vielzahl verschiedener Faktoren abhängen, wie etwa Aufmerksamkeit, Signalwert der Reize, Antwortart, Täuschungsmotivation oder Konsequenzen einer Entdeckung (vgl. Orne, Thackray & Paskewitz, 1972, S. 762). Dabei schließen sich die diversen Theorien nicht unbedingt gegenseitig aus. Sie betonen nur unterschiedliche Mechanismen, die durchaus parallel wirken bzw. interagieren können. Komplexere, **integrative Ansätze** sind gefordert, um ein besseres Verständnis der Bedingungsfaktoren zu erlangen. Entsprechende Entwürfe wurden vereinzelt vorgelegt (z. B. Elaad & Ben-Shakhar, 1989, S. 443; Steller, 1987, S. 132ff.), ohne daß sich jedoch weitere Bemühungen abzeichnen, sie einer konsequenten empirischen Überprüfung zu unterziehen.

3. Problemstellung, Generierung der Forschungsfragen und Ableitung von Hypothesen

3.1 Scheinverbrechen-Experiment zum DLT und GAT

3.1.1 Standardisierte Testdurchführung

Die vorliegende Arbeit will einen Beitrag zur Klärung der psychophysiologischen Grundlagen der neueren Befragungstechniken **DLT und GAT** leisten. Ein besonderes Augenmerk liegt dabei auf der **Objektivität** der Tests. Damit soll ein wesentlicher Kritikpunkt an den üblichen Vorgehensweisen vermieden werden. Denn in der praktischen Anwendung aber auch in den meisten Feld- und Analogstudien erfolgt die Durchführung der psychophysiologischen Aussagebeurteilung in einer relativ unstandardisierten Form.

Vor allem die **direkten Verfahren** erfordern eine individuelle Anpassung der Untersuchung an den jeweiligen Pb (z. B. Vortest-Interview und Formulierung der Vergleichsfragen). So gestattet etwa der **KFT** keine Vereinheitlichung der Testbedingungen (vgl. Abschnitt 2.4.2). D. h., die Standardisierbarkeit, die als ein wesentliches Kriterium für die Qualität eines psychologischen Testverfahrens gilt, ist beim KFT nicht gegeben (Furedy, 1996a, S. 98). Da außerdem die Notwendigkeit einer unmittelbaren Interaktion zwischen Untersucher und Pb besteht, kann der Testleiter (absichtlich oder unwillkürlich) Einfluß auf den Verlauf und das Ergebnis der Untersuchung nehmen, wodurch deren Objektivität erheblich beeinträchtigt wird. Die Vermeidung solcher Probleme ist mittlerweile ein wichtiges Anliegen der psychophysiologischen Aussageforschung. Ein Beispiel dafür bietet der **DLT** (Abschnitt 2.5), der nach Ansicht seiner Befürworter sogar derart standardisierbar ist, daß man ihn vollautomatisch, maschinell durchführen könnte (vgl. Honts et al., 1995, S. 205). Obwohl dieser Vorschlag durchaus praktikabel erscheint, hat man ihn in den bisher veröffentlichten Validitätsstudien nicht umgesetzt. Dort wurde entweder eine konventionelle „face-to-face“-Untersuchung realisiert (vgl. Honts & Raskin, 1988, S. 58), oder der Untersucher befand sich zwar außerhalb des Testraums, befragte die Pbn aber dennoch persönlich über eine Wechselsprechanlage (Horowitz et al., 1997, S. 111).

Bei den **indirekten Verfahren** liegt zwar eine potentiell höhere Standardisierung vor (siehe auch Abschnitt 2.7.2), dennoch wird in der Feldanwendung (Elaad, 1990, S. 523; Elaad et al., 1992, S. 759) und in vielen Analogstudien (z. B. O’Toole et al., 1994, S. 256f.) weiterhin eine typische Interviewsituation mit einer individuellen, mündlichen Darbietung der Items praktiziert. In einigen Experimenten zum **TWT** bzw. **GAT** kam

auch eine technisch gesteuerte Reizpräsentation zum Einsatz, beispielsweise per Tonband (z. B. Bradley & Warfield, 1984, S. 686; Schumacher, 1993, S. 79; Steller et al., 1987, S. 337) oder per Bildschirm (vgl. Ben-Shakhar & Dolev, 1996, S. 275; Elaad & Ben-Shakhar, 1997, S. 588). Dadurch konnte man die Testbedingungen für alle Versuchsteilnehmer äquivalent gestalten. Aber selbst in den Analogstudien zu den neueren Befragungstechniken ist eine solche Vorgehensweise nicht üblich. So wurden etwa im Experiment zum TWT und GAT von Bradley, MacLaren und Carle (1996, S. 156) die Fragen und Items vom Untersucher vorgelesen.

Das **eigene Experiment** nutzt einige der noch unausgeschöpften Möglichkeiten zur Verbesserung der Durchführungsobjektivität von DLT und GAT. Die Interaktionen zwischen Testleiter und Pb sollen zusätzlich reduziert und potentielle Untersuchereffekte weitgehend ausgeschlossen werden. Alle Instruktionen erfolgen in standardisierter Form. Die Pbn befinden sich während der Untersuchung allein in der Meßkabine. Die Testfragen bzw. -items werden computerisiert sowohl verbal als auch schriftlich dargeboten. Die Computersteuerung ermöglicht außerdem ein genaueres Timing der Stimulusdarbietung, und durch die multimodale Präsentation wird eine Ablenkung der Aufmerksamkeit vom Reizmaterial erschwert. Diese Vorgehensweise zielt jedoch nicht darauf ab, standardisierte und konventionelle Testprozeduren gegenüberzustellen. Vielmehr soll grundsätzlich die Objektivität als eine wesentliche Voraussetzung für aussagekräftige Ergebnisse gesichert werden (vgl. Abschnitt 2.4.2).

Da die Applikation sowohl des DLT als auch des GAT einen konkreten Tatbezug erfordert, wird hier ebenfalls auf ein **Scheinverbrechen-Experiment** zurückgegriffen. Wie jedoch die folgenden Ausführungen zeigen, unterscheidet sich diese Untersuchung in ihrer Zielsetzung und Methodik von herkömmlichen Analogstudien.

3.1.2 Direkte Untersuchung der Effekte unterschiedlicher Fragen- bzw. Itemtypen auf die physiologischen Reaktionen

Bislang wurden nur **wenige Studien zu den beiden neueren Befragungstechniken** publiziert. Deren Ergebnisse deuten darauf hin, daß der stärker standardisierte DLT ähnliche bzw. sogar höhere Trefferquoten erzielt als der KFT (vgl. Abschnitt 2.5) und daß der GAT im Vergleich zum TWT eine bessere Differenzierung zwischen Schuldigen und Unschuldigen mit Tatwissen ermöglicht (vgl. Abschnitt 2.8). Die betreffenden Feld- und Analogstudien haben sich primär mit der diagnostischen Güte der Verfahren beschäftigt. Dabei standen die Ergebnisse der meist numerischen Auswertungsverfahren und die Gültigkeit der daraus resultierenden Individualdiagnosen im Vordergrund.

Neben grundsätzlichen Bedenken gegenüber derartigen Schätzungen der Treffsicherheit (Abschnitt 2.9) ist ein weiterer Kritikpunkt darin zu sehen, daß die entsprechenden Untersuchungen nur mittelbare Schlußfolgerungen über die Reaktionen auf bestimmte Stimuli und diesbezügliche Unterschiede zulassen. Hierbei handelt es sich um ein fundamentales **Problem der bisherigen psychophysiologischen Aussageforschung**. Abgesehen von einigen Ausnahmen (z. B. Barland & Raskin, 1975a, S. 325ff.; Dawson, 1980, S. 11ff.; Podlesny & Raskin, 1978, S. 353ff.; Raskin & Hare, 1978, S. 132f.) wurden auch bei den meisten Untersuchungen zum KFT und TWT überwiegend Aspekte der kriterienbezogenen Validität analysiert. Davon sind sowohl Feld- als auch Analogstudien betroffen, obwohl bereits Podlesny und Raskin (1977, S. 788) darauf hingewiesen haben, daß eine direkte, objektive Quantifizierung der Reaktionen auf die unterschiedlichen Fragen bzw. Items zu den Mindestforderungen gehört, die man an Analogstudien stellen muß. Dagegen handelt es sich bei den üblicherweise erhobenen Punktwerten und Trefferquoten lediglich um eine indirekte Operationalisierung der Reaktionsunterschiede. Dies gilt insbesondere für die semi-objektiven Scorings der Kontrollfragentechniken, die nach Furedy (1991, S. 244) keine Quantifizierung im eigentlichen Sinne verkörpern, da sie eine relativ willkürliche Auswahl der herangezogenen Reaktionskennwerte erlauben (Abschnitt 2.4.2).

Die fast ausschließliche Berücksichtigung numerischer Auswertungen bringt mehrere **Nachteile** mit sich. Informationen über die absolute Stärke der Reaktionen auf die unterschiedlichen Fragen- oder Itemtypen gehen verloren. Folglich besteht keine Möglichkeit, die Effekte der inter- bzw. intraindividuellen Bedingungsvariationen auf die Reaktionsparameter zu untersuchen und statistisch zu analysieren. Darüber hinaus kann man keine genaueren Angaben zum Betrag der Reaktionsunterschiede machen. Dadurch wird auch die Vergleichbarkeit der Ergebnisse zwischen verschiedenen Studien beeinträchtigt. Aufgrund der kaum nachvollziehbaren Kombination und Gewichtung der Daten aus mehreren Reaktionssystemen bleibt der relative Beitrag einzelner physiologischer Variablen zu den Punktwerten und Befunden unklar (Ben-Shakhar & Furedy, 1990, S. 85). Die Frage, inwiefern eine bestimmte physiologische Variable eine signifikante Differenzierung zwischen den unterschiedlichen Stimuli der Befragungstechniken bzw. zwischen Schuldigen und Unschuldigen ermöglicht, ist jedoch unter sowohl diagnostischen als auch theoretischen Gesichtspunkten von Interesse (vgl. Podlesny & Raskin, 1977, S. 788). Beim TWT und GAT sind diese Probleme weniger gravierend, aber dennoch präsent. Zwar beschränkt man sich in den betreffenden Studien meist auf elektrodermale Reaktionen, und die Regeln des Lykken-Scorings sind eindeutiger und objektiver als die numerischen Auswertungsverfahren der direkten Befragungstechniken. Allerdings lassen auch die Lykken-Scores keine eindeutigen Rückschlüsse auf die Reaktionsstärken zu. Denn der Punktwert pro Multiple-Choice-Frage gibt nur die Rang-

position der Reaktion auf das relevante Item (stärkste bzw. zweitstärkste) in Relation zu den irrelevanten Items wieder (vgl. Abschnitt 2.7.1), nicht jedoch die absoluten Reaktionsstärken bzw. quantitativen Unterschiede zwischen den beiden Itemtypen.

Die **vorliegende Arbeit** weicht von den bisherigen Studien zum DLT und GAT ab. Hier sollen die reizbedingten körperlichen Veränderungen objektiv quantifiziert und die Effekte der unterschiedlichen Fragen- und Itemtypen direkt überprüft werden. Die Untersuchung erfolgt zwar ebenfalls im Rahmen eines Scheinverbrechen-Experiments. Wichtig ist jedoch, daß die Durchführung fingierter Delikte nicht das vorrangige Ziel verfolgt, die kriterienbezogene Validität der psychophysiologischen Aussagebeurteilung unter simulierten Tatbedingungen zu bestimmen. Insofern stellt dieses Experiment keine Analogstudie im klassischen Sinne dar. Es stehen weder Trefferquoten im Vordergrund, noch werden die Tests in Anlehnung an Realbedingungen durchgeführt. Die Implementierung von Scheinverbrechen dient primär zur Variation der Fragen- bzw. Itemtypen, d. h. zur Herstellung des konkreten Tatbezugs der relevanten Stimuli. Zu diesem Zweck wird auf einen simulierten Schmuckdiebstahl zurückgegriffen, zumal es sich dabei um ein gängiges Szenario von Analogstudien handelt (z. B. Ben-Shakhar & Dolev, 1996, S. 275; Horowitz et al., 1997, S. 110; Kircher & Raskin, 1988, S. 293; Podlesny & Raskin, 1978, S. 347f.).

3.1.3 Intraindividuelle Variation von Täuschung und Aufrichtigkeit bei den relevanten Fragen bzw. Items

Im Prinzip beinhalten sowohl der DLT als auch der GAT **drei Fragen- bzw. Itemtypen**:

Der **DLT** umfaßt neben irrelevanten und tatbezogenen Fragen auch Lügen-Kontrollfragen, die die Pbn instruiert wahrheitswidrig beantworten sollen. Es wird angenommen, daß Schuldige auf die relevanten Fragen und Unschuldige auf die Lügen-Kontrollfragen stärker reagieren. Die körperlichen Veränderungen nach Präsentation der irrelevanten Stimuli gehen nicht in die Auswertung ein.

Der **GAT** stellt Multiple-Choice-Fragen, die direkt auf die Täterschaft abzielen. Neben den kritischen **Tatdetails** werden zu jeder Frage mehrere (ähnlich plausible) irrelevante Items dargeboten. Man erwartet, daß Pbn mit Tatwissen stärker auf die relevanten als auf die irrelevanten Items reagieren, während unwissende Personen diesbezüglich keine systematischen Unterschiede zeigen dürften. Darüber hinaus sollen Täter ausgeprägtere Reaktionsdifferenzen aufweisen als Unschuldige mit Tatwissen. Das jeweils erste irre-

levante Item pro Multiple-Choice-Block bleibt bei der Auswertung unberücksichtigt, da es unmittelbar im Anschluß an die Frage dargeboten wird und obendrein die Neuheit der angesprochenen Thematik zu einer Reaktionssteigerung führen kann.

Aufgrund der gängigen Methodik von Validitätsstudien, Reaktionsvergleiche nur auf der Basis von Punktwerten und Trefferquoten vorzunehmen, kann man die körperlichen Veränderungen nach den verschiedenen Fragen- und Itemtypen nicht direkt zueinander ins Verhältnis setzen. Ein solcher Vergleich ist jedoch v. a. im Hinblick auf die tatbezogenen Stimuli interessant, da sich die **Reaktionsunterschiede zwischen Schuldigen und Unschuldigen** im wesentlichen bei den **relevanten Fragen bzw. Items** manifestieren sollen. In einer der wenigen Analogstudien mit objektiver Quantifizierung der physiologischen Aufzeichnungen konnten beispielsweise Podlesny und Raskin (1978, S. 353ff.) zeigen, daß Schuldige und Unschuldige v. a. auf die tatbezogenen Reize von **KFT und TWT** unterschiedlich stark reagierten, während die Differenzen bei den Kontrollfragen bzw. irrelevanten Items geringer ausfielen.

Lykken (1998, S. 119ff.) versuchte, die Prämissen des KFT aus dessen Anwendung, d. h. seiner Durchführung und Auswertung, zu erschließen. Seiner Ansicht nach gehen im Prinzip alle Verfahren der „Lügendetektion“ (und somit auch der **DLT**) zumindest implizit davon aus, daß ein Pb auf eine relevante Frage intensiver reagiert, falls er die Tat begangen hat und darauf wahrheitswidrig antwortet, als wenn er hinsichtlich der Tat unschuldig ist und sie glaubhaft abstreitet:

> A given subject will respond more strongly to a relevant question if he answers it deceptively than if his denial is truthful. That is, if his response would be R_I if he is innocent and R_G if he is guilty, then R_G will be larger than R_I ($R_G > R_I$). (Lykken, 1998, S. 120)

Auch beim **GAT** soll die Differenzierung zwischen Tätern und Unschuldigen mit Tatwissen auf die Variation des Wahrheitsgehalts bei den relevanten Items zurückzuführen sein (vgl. Abschnitt 2.8). D. h., es wird postuliert, daß ein Pb stärkere Reaktionen auf die Frage nach tatbezogenen Details zeigt, wenn er sie wahrheitswidrig statt wahrheitsgemäß beantwortet.

Das eigentliche **Dilemma** dieser Annahmen liegt jedoch darin begründet, daß sie sich anhand der bisherigen Forschungsansätze nicht empirisch überprüfen lassen. Denn neben einer direkten Untersuchung der Effekte auf die Reaktionsstärken legen sie eine intraindividuelle Variation der Täterschaft und somit von Täuschung und Aufrichtigkeit bei den relevanten Stimuli nahe. Beide Annahmen gehen nämlich davon aus, daß sich

die Reaktionsunterschiede zwischen den Fragen bzw. Items, die sich auf eine begangene vs. nicht begangene Tat beziehen und demzufolge wahrheitswidrig vs. wahrheitsgemäß verneint werden, auch *innerhalb* der Personen manifestieren und nicht nur im Vergleich *zwischen* Schuldigen und Unschuldigen. Trotz triftiger Gründe für eine intraindividuelle Analyse wurde sie in keiner der bislang vorliegenden Feld- oder Analogstudien konsequent umgesetzt. Und das, obwohl einer solchen Methodik allgemein im Kontext psychophysiologischer Fragestellungen gewisse Vorteile zugesprochen werden (Fahrenberg, 1983, S. 97). Insofern handelt es sich hierbei sowohl auf seiten der konventionellen Verfahren (KFT und TWT) als auch speziell in bezug auf die neueren Befragungstechniken (DLT und GAT) um ein Desiderat der psychophysiologischen Aussageforschung.

Eine wesentliche Fragestellung des vorliegenden Experiments lautet somit: Reagieren Pbn beim DLT bzw. GAT auf die relevanten Fragen oder Items nach einer Tat, die sie verübt haben und deshalb wahrheitswidrig abstreiten, stärker als auf entsprechende Vergleichsreize nach einer nicht begangenen Tat, die sie wahrheitsgemäß verneinen?

Zur näheren Untersuchung dieses Problems bietet es sich an, die **Prinzipien des Truth Control Tests** (TCT, vgl. Abschnitt 2.6) auf den DLT und den GAT zu übertragen. Beim TCT werden die Pbn hinsichtlich zweier äquivalenter Delikte getestet. Eines von beiden ist real und das andere fiktiv. Bereits Bradley, MacLaren und Black (1996) haben ein Scheinverbrechen-Experiment zu einer Befragungstechnik durchgeführt, die ähnlich dem TCT konzipiert war. Die physiologischen Aufzeichnungen wurden zwar wiederum nur numerisch ausgewertet. Die Ergebnisse deuteten jedoch darauf hin, daß die Schuldigen erwartungsgemäß auf die Fragen nach dem von ihnen begangenen Vergehen stärker reagierten als auf Stimuli, die eine ähnliche, aber nicht durchgeführte Handlung thematisierten. Allerdings wurde dort versäumt, die Pbn von der Brisanz der erfundenen Anschuldigung zu überzeugen, so daß die Reaktionsunterschiede auch auf die potentiell unterschiedliche Tatrelevanz der kritischen Fragen zurückzuführen waren. Um sicherzustellen, daß die Vergleichsfragen des TCT als eine angemessene Kontrollbedingung fungieren können, müssen die Pbn der Ansicht sein, daß sie wirklich hinsichtlich zweier Vergehen unter Tatverdacht stehen (Lykken, 1998, S. 144).

Deshalb wird in dieser Untersuchung folgende **Vorgehensweise** gewählt: Im Gegensatz zu den herkömmlichen Analogstudien, die einen Vergleich schuldiger vs. unschuldiger Pbn vorsehen, variiert man hier die Täterschaft innerhalb der Personen („within subjects“). Das Experiment beinhaltet zwei Scheinverbrechen, die äquivalent gestaltet sind (simulierter Diebstahl eines Fingerrings vs. einer Halskette). Wichtig ist, daß die Pbn nur eines davon durchführen. Bei ihrer Tat werden sie aber mit den relevanten Details

beider Scheinverbrechen konfrontiert. Auf diese beziehen sich die Fragen der anschließenden psychophysiologischen Aussagebeurteilung. Jeweils die Hälfte der Vpn wird instruiert, entweder einen Ring oder eine Kette aus dem Schreibtisch eines Büroraums zu „stehlen“. Unter der Tatbedingung „Ring“ sollen die Pbn nach Betreten des Zimmers eine Schreibtischschublade öffnen. Darin liegt unter anderem die Halskette. Die Pbn dürfen die Gegenstände nur inspizieren, aber nicht berühren. Anschließend sollen sie eine zweite Schublade öffnen und den darin befindlichen Ring entwenden. Bei den Vpn der Tatbedingung „Kette“ ist der Ablauf entsprechend umgekehrt (erst Schublade mit Ring öffnen und betrachten, dann Kette an sich nehmen). Somit ist jeder Pb hinsichtlich eines Diebstahls „schuldig“ und hinsichtlich des anderen „unschuldig“. Da v. a. der Vergleich der Reaktionen innerhalb der Personen interessiert, wird auf eine Gruppe „vollständig unschuldiger“ Pbn verzichtet.

Anschließend müssen alle Versuchsteilnehmer eine **psychophysiologische Aussagebeurteilung** absolvieren. Dabei kommen zwei Befragungstechniken zum Einsatz, die zwischen den Personen („between subjects“) variiert werden. D. h., jeder Pb wird nur mit einem der beiden Verfahren getestet, zumal mehrere Studien darauf hindeuten, daß eine sequentielle Testung mit verschiedenen Befragungstechniken negative Auswirkungen auf die Treffsicherheit des zuletzt eingesetzten Verfahrens haben kann (vgl. Abschnitt 2.9.1). Die hier verwendeten Befragungstechniken sind in Anlehnung an den DLT und den GAT konzipiert. Allerdings thematisieren sie sowohl den Ring- als auch den Kettendiebstahl. Gemäß dem TCT werden die Pbn mit Paaren von relevanten Fragen bzw. Items konfrontiert, die inhaltlich parallelisiert sind und auf die Tatbegehung bzw. die kritischen Details beider Scheinverbrechen abzielen (z. B. DLT: „Haben Sie den Ring gestohlen?“ vs. „Haben Sie die Kette gestohlen?“; GAT: „Welches Schmuckstück haben Sie gestohlen? War es ... ein Ring? ... eine Kette?“; vgl. Tabelle 4). Weitere relevante Stimuli thematisieren unter anderem die Numerierung der Schubladen, Merkmale der Schmuckgegenstände, Büroutensilien im Schreibtisch oder die Gefäße, worin sich die Schmuckstücke befanden. Somit werden mehrere Fragen bzw. Items dargeboten, die sich auf die zwei Scheinverbrechen beziehen (Relevant-Ring vs. Relevant-Kette). Die Pbn sollen beide Tatvorwürfe abstreiten. Und in Abhängigkeit davon, welches Scheinverbrechen sie begangen haben, beantworten sie jeweils eine Hälfte der Fragen bzw. Items wahrheitswidrig (Relevant-Täuschung) und die andere Hälfte wahrheitsgemäß (Relevant-Aufrichtigkeit). Durch die Variation der Tatbedingung (Ring vs. Kette) zwischen den Personen ist sichergestellt, daß – insgesamt gesehen – jeder Stimulus gleich häufig unter den Bedingungen Täuschung und Aufrichtigkeit auftritt. Außerdem läßt sich durch die besondere Inszenierung der Scheinverbrechen ein wesentliches Problem des TCT umgehen. Da alle Pbn bei der Tatbegehung mit den kritischen Details beider Delikte konfrontiert werden, bereitet es relativ wenig Schwierigkeiten, sie anschließend

davon zu überzeugen, daß sie sowohl hinsichtlich des Ring- als auch des Kettendiebstahls unter Tatverdacht stehen.

Tabelle 4. Beispiele für die relevanten Fragen des DLT bzw. Items des GAT und intraindividuelle Variation von Täuschung vs. Aufrichtigkeit in Abhängigkeit von der Tatbedingung

	Fragen- bzw. Itemtyp	
	Relevant-Ring	Relevant-Kette
Testart	Beispiele	
DLT	Haben Sie den Ring gestohlen?	Haben Sie die Kette gestohlen?
GAT War es ...	ein Ring?	eine Kette?
Tatbedingung	Variation des Wahrheitsgehalts	
Ring	Täuschung	Aufrichtigkeit
Kette	Aufrichtigkeit	Täuschung

Anmerkung. Die Fragen bzw. Items werden jeweils verneint.

Diese Methodik zeigt **konzeptionelle Ähnlichkeiten zum „Differentiation-of-Deception"-Paradigma** (DDP, vgl. Abschnitt 2.9.3). Auch beim DDP werden inhaltlich parallelisierte Fragenpaare dargeboten. Darüber hinaus variiert man die Bedingungen Täuschung vs. Aufrichtigkeit intraindividuell und kontrolliert sie interindividuell so, daß auf jede Frage in jeweils 50% der Darbietungen wahrheitsgemäße und wahrheitswidrige Antworten erfolgen. Allerdings werden im DDP möglichst emotional neutrale Inhalte angesprochen (z. B. einfache autobiographische Sachverhalte bzw. Allgemeinwissen), während die tatbezogenen Fragen der vorliegenden Untersuchung durchaus eine gewisse emotionale Signifikanz aufweisen sollen.

3.1.4 Intraindividuelle Variation von Täuschung und Aufrichtigkeit bei den Kontrollfragen des DLT

Beim DLT gilt die wahrheitswidrige Beantwortung der **instruierten Lügen-Kontrollfragen** als eine wesentliche Voraussetzung für valide Testbefunde (vgl. Abschnitt 2.5). Den Pbn wird suggeriert, daß die damit einhergehenden körperlichen Veränderungen ihre typischen Reaktionsmuster beim Lügen offenlegen würden. Diese könne man dann als Referenz für die Reaktionen auf die relevanten Fragen heranziehen. Unklar ist jedoch, inwiefern unaufrichtige Antworten tatsächlich zu einer Reaktionserhöhung beitragen, oder ob die angestrebte Signifikanzsteigerung der Lügen-Kontrollfragen v. a. auf die Vortest-Suggestionen zurückzuführen ist. Dabei muß man auch bedenken, daß eine instruierte Antwort, deren Unwahrheit sowohl für den Untersucher als auch den Pb

außer Zweifel steht, sich möglicherweise fundamental von einer intentionalen Lüge unterscheidet.

Berücksichtigt man jedoch die bisherigen Ergebnisse zum **DDP** (zusammenfassend Gödert et al., 2001, S. 62f.), so besteht Grund zur Annahme, daß selbst dann erhöhte Reaktionen auftreten, wenn man die Pbn zu einer Täuschung instruiert. Andererseits ist zu beachten, daß die Implikationen des DDP für die psychophysiologische Aussagebeurteilung sehr begrenzt sind (vgl. Abschnitt 2.9.3). Außerdem gibt es **empirische Befunde zum KFT**, die gegen entsprechende Effekte bei den Kontrollfragen sprechen. Honts, Raskin und Kircher (1992, S. 267ff.) reanalysierten polygraphische Aufzeichnungen aus Feld- und Analogstudien, in denen mindestens eine Kontrollfrage während den Tests irrtümlich bejaht und somit gemäß der Logik des Verfahrens vermutlich wahrheitsgemäß beantwortet wurde. Die numerischen Auswertungen erbrachten keine signifikanten Unterschiede zu verneinten und damit potentiell wahrheitswidrig beantworteten Kontrollfragen. Diese Studie weist darauf hin, daß zumindest beim KFT die Antworten auf die Kontrollfragen und deren Wahrheitsgehalt von sekundärer Bedeutung sind.

Um zu überprüfen, ob der **Wahrheitsgehalt bei den Kontrollfragen des DLT** eine Rolle spielt, werden auch hier die Bedingungen Täuschung vs. Aufrichtigkeit intraindividuell variiert. Das Reizmaterial besteht aus mehreren inhaltlich parallelisierten Fragenpaaren, die sich auf kleinere Normverstöße beziehen. Die beiden Fragen eines Paares sind entgegengesetzt formuliert (z. B.: „Sind Sie manchmal unehrlich?“ vs. „Sind Sie immer ehrlich?“). Der Wahrheitsgehalt wird in Anlehnung an die Instruktionen des DLT manipuliert. Man weist die Pbn an, alle Vergleichsfragen zu verneinen, wodurch jeweils ein Paarling wahrheitswidrig (Lügen-Kontrollfrage) und der andere wahrheitsgemäß beantwortet wird (letzterer wird im folgenden der Einfachheit halber als „[instruierte] Wahrheit-Kontrollfrage“ bezeichnet, diese ist jedoch nicht mit der „truth control question“ des TCT zu verwechseln). Die Pbn sollen sich bei jeder Frage bewußt vor Augen führen, daß es sich um eine wahrheitsgemäße bzw. wahrheitswidrige Antwort handelt. Außerdem werden sie angewiesen, an konkrete Situationen zu denken, in denen sie die Normverstöße begangen haben. Analog zum DLT rechtfertigt man dieses Vorgehen mit der Begründung, man müsse die physiologischen Reaktionsmuster des Pb beim Lügen und Wahrheitsagen genauer analysieren, um angemessene Vergleichsstandards für die relevanten Reaktionen zu erhalten. Adäquate körperliche Begleiterscheinungen beim wahrheitsgemäßen und wahrheitswidrigen Beantworten der Kontrollfragen würden das Auftreten eindeutiger Ergebnisse begünstigen.

Der Wahrheitsgehalt der Antworten auf die **irrelevanten Fragen bzw. Items** wird nicht variiert. Dies liegt einerseits daran, daß die irrelevanten Fragen für die Auswertung des **DLT** ohne Belang sind. In der vorliegenden Untersuchung besteht ihre Rolle vielmehr darin, im Sinne von „Pufferfragen“ die Aufmerksamkeit der Pbn auf die Reizdarbietung aufrechtzuerhalten, da sie die einzigen Stimuli sind, die bejaht werden. Dadurch will man die Möglichkeit vereiteln, daß manche Pbn nur noch stereotyp mit Verneinungen reagieren, ohne sich den Inhalt der Fragen zu vergegenwärtigen. Andererseits werden bei der Standardprozedur des **GAT** die irrelevanten Items sowohl von Schuldigen als auch von Unschuldigen wahrheitsgemäß verneint. Darum hält man auch in der vorliegenden Studie den Wahrheitsgehalt der entsprechenden Antworten konstant.

3.2 Physiologische Variablen

Im Rahmen der psychophysiologischen Aussagebeurteilung werden (v. a. bei Untersuchungen mit direkten Befragungstechniken) für gewöhnlich **mehrere körperliche Erregungsmaße** erhoben (vgl. Abschnitt 2.3). Dabei handelt es sich in der Regel um elektrodermale, kardiovaskuläre und respiratorische Größen. Wenn man die Effizienz dieser Variablen vergleicht, erbringen elektrodermale Reaktionsparameter meist die beste Diskrimination zwischen den verschiedenen Fragen- oder Itemtypen bzw. Schuldigen und Unschuldigen.

3.2.1 Hautleitfähigkeitsreaktionen (SCRs)

Eine Reihe von Studien sowohl zu den konventionellen Befragungstechniken (KFT und TWT) als auch zu den neueren Verfahren (DLT und GAT) deutet auf eine **Überlegenheit von Hautleitfähigkeits- bzw. Hautwiderstandsreaktionen** gegenüber kardiovaskulären und respiratorischen Maßen hin (z. B. Bradley & Rettinger, 1992, S. 58; Elaad & Ben-Shakhar, 1989, S. 449; Horowitz et al., 1997, S. 113; O’Toole et al., 1994, S. 259; Patrick & Iacono, 1989, S. 351f.; für einen Überblick über ältere Studien vgl. auch Ben-Shakhar & Furedy, 1990, S. 91f.; Podlesny & Raskin, 1977, S. 791f.). Die Empirie widerspricht somit der traditionellen Auffassung vieler professioneller Polygraphen-Untersucher (z. B. Reid & Inbau, 1966, S. 220f.), wonach die EDA in der Praxis wenig zweckmäßig sei, da sie sogar *zu* sensitiv auf emotionale Erregungszustände reagiere (vgl. auch Podlesny & Raskin, 1977, S. 792). Inzwischen ist jedoch ihr besonderer Nutzen weitgehend anerkannt. So werden etwa bei computergestützten Auswertungen polygraphischer Aufzeichnungen (vgl. Abschnitt 2.4.1) elektrodermale Kennwerte am stärksten gewichtet (Matte, 1996, S. 423f.). Demnach kann man die EDA

durchaus als „Conditio sine qua non“ der gegenwärtigen psychophysiologischen Aussageforschung bezeichnen.

In der **vorliegenden Arbeit** wird die Amplitude der Hautleitfähigkeitsreaktion (SCR) als abhängige Variable herangezogen. Neben der o. g. diagnostischen Relevanz der EDA gibt es weitere Gründe für die Berücksichtigung der SCR. Einerseits gilt sie als validester Indikator der Orientierungsreaktion (OR), deren theoretische Bedeutung für die psychophysiologische Aussagebeurteilung bereits erörtert wurde (Abschnitt 2.10). Andererseits spiegelt die Hautleitfähigkeit die zugrundeliegende Innervation der ekkrinen Schweißdrüsen durch den Sympathikus wider (Boucsein, 1988, S. 23ff.; Vossel & Zimmer, 1998, S. 50), so daß eventuelle Reaktionsunterschiede Rückschlüsse auf die Beteiligung des sympathischen Teils des peripheren vegetativen Nervensystems zulassen.

3.2.2 Herzschlagfrequenz (HR)

Abgesehen von dem herausragenden Stellenwert elektrodermaler Reaktionen für die psychophysiologische Aussagebeurteilung belegen einige Analogstudien, daß auch andere körperliche Variablen eine signifikante Differenzierung zwischen Schuldigen und Unschuldigen erbringen. Dies gilt auch für den „relativen Blutdruck“ (siehe z. B. Barland & Raskin, 1975a, S. 326f.; Dawson, 1980, S. 13f.; Kircher & Raskin, 1988, S. 297f.; Patrick & Iacono, 1989, S. 351). Die Eignung dieses sog. „Cardio“-Kanals als Indikator für die Aktivität des Herz-Kreislauf-Systems bleibt jedoch äußerst fragwürdig (vgl. Abschnitt 2.4.2). Seine Messung ist weder reliabel noch valide (Schandry, 1998, S. 160); und sie erfüllt nicht die Anforderungen, die man an ein belastungsfreies psychophysiologisches Verfahren stellt (Furedy, 1986, S. 684). Außerdem deutet das Kartentest-Experiment von Kugelmass und Lieblich (1966, S. 214f.) darauf hin, daß die zusätzliche Erhebung des „relativen Blutdrucks“ die elektrodermalen Reaktionsunterschiede zwischen den kritischen und neutralen Items verringern und somit die Entdeckungsraten beeinträchtigen kann. Aus diesen Gründen wird hier auf phasische Veränderungen der **Herzschlagfrequenz** („heart rate“, HR) als alternative kardiovaskuläre Variable zurückgegriffen.

Neben der in der Praxis gebräuchlichen (vgl. Matte, 1996, S. 364f.), aber relativ unzuverlässigen Abschätzung der Pulsfrequenz anhand des „Cardio“-Kanals wurden in einigen wissenschaftlichen Untersuchungen die **HR-Reaktionen per Elektrokardiogramm** (EKG) bestimmt.

In einer Analogstudie zum **KFT** erfaßten Raskin und Hare (1978, S. 133) außer den üblichen physiologischen Variablen auch die HR. Schuldige und unschuldige Pbn zeigten sowohl auf die relevanten Fragen als auch auf die Kontrollfragen zunächst eine Akzeleration in den gemittelten Reaktionsverläufen. Deren Maximum lag im Durchschnitt ungefähr im Zeitintervall der Antwortgabe. Nach diesem Anstieg reagierten die Schuldigen mit einer ausgeprägten Dezeleration auf die relevanten Fragen. Bei den Kontrollfragen hingegen kehrte die HR lediglich wieder auf das Prästimulus-Baseline-Niveau zurück. Unschuldige zeigten weder auf die relevanten noch auf die Kontrollfragen eine deutliche Dezeleration. Es sei jedoch erwähnt, daß die Zuverlässigkeit und Aussagekraft dieser Befunde von Lykken (1978, S. 140) unter methodischen Gesichtspunkten erheblich in Frage gestellt wurden. Dennoch konnten Podlesny und Raskin (1978, S. 354f.) die KFT-Ergebnisse von Raskin und Hare replizieren. Sie fanden jedoch keine entsprechenden Effekte für den TWT. Die HR-Dezeleration wurde von Raskin und Hare (1978, S. 135) sowie Podlesny und Raskin (1978, S. 358) zunächst noch im Sinne einer Aufmerksamkeitsreaktion interpretiert, die auf eine Informationsverarbeitung externer Stimulation abziele. Später hingegen deutete Raskin (1979, S. 601) die kardiovaskulären Daten als Anzeichen einer Defensivreaktion Schuldiger auf die relevanten Fragen.

Ebenso wie Podlesny und Raskin konnten Balloun und Holmes (1979, S. 320) keine differentiellen HR-Reaktionen Schuldiger und Unschuldiger in einem Scheinverbrechen-Experiment zum **TWT** nachweisen. Im Gegensatz dazu erbrachte die Kartentest-Studie von Cutrow et al. (1972, S. 584f.) auch für die HR eine überzufällige Treffsicherheit. Allerdings variierte dort die Reaktionsrichtung zwischen den Pbn. Einige zeigten eine stärkere Akzeleration, andere eine stärkerer Dezeleration auf die relevanten Items. Letzteren Befund interpretierten die Autoren im Sinne einer stärkeren OR (zur Problematik dieser Annahme siehe Abschnitt 2.10.1).

Darüber hinaus wurden auch Untersuchungen durchgeführt, die die Herzschlagfrequenz über die **Pulsfrequenz** („pulse rate“, PR) operationalisierten. In einem Kartentest-Experiment bestimmten Kugelmass und Lieblich (1966, S. 213f.) die stärksten Veränderungen der PR (unabhängig von der Richtung) innerhalb eines Zeitfensters von 5 Sekunden nach Präsentation der Items. Die damit erzielten Entdeckungsraten lagen jedoch nicht über dem Zufallsniveau. Bradley und Janisse (1981, S. 310) erfaßten in ihrer Analogstudie zum KFT und TWT neben Hautwiderstands- und Pupillenreaktionen auch die phasische PR (dort als HR bezeichnet). Als Reaktionsparameter wurde die maximale Dezeleration binnen 15 Sekunden nach Stimulusdarbietung herangezogen. In beiden Testverfahren differenzierte die PR signifikant zwischen schuldigen und unschuldigen Pbn. Die genauen Reaktionsverläufe blieben jedoch unklar, da man lediglich eine numerische Auswertung der Daten vornahm.

Aufgrund der umstrittenen Befunde zum KFT (vgl. auch die Kontroverse zwischen Lykken, 1978, S. 140, und Raskin, 1978, S. 144) und der o. g. inkonsistenten Ergebnisse zum TWT sollen auch im **vorliegenden Experiment** die HR-Reaktionen erfaßt werden. Ferner hat die Untersuchung dieser Variable im Zusammenhang mit dem DLT und dem GAT den Charakter einer Pilotstudie. Zur sinnvollen Verrechnung über Trials und Personen hinweg und um detailliertere Aussagen über die phasischen Veränderungen machen zu können, werden die Daten **echtzeitskaliert** (vgl. Velden, 1999; Velden & Graham, 1988; Velden & Wölk, 1987). Die Auswertung folgt damit der Empfehlung von Podlesny und Raskin (1977, S. 797). Die Autoren schlugen vor, v. a. im Rahmen explorativer Studien zur psychophysiologischen Aussagebeurteilung die kardiovaskulären Reaktionen auf Basis ihrer zeitabhängigen Verläufe über mehrere Sekunden hinweg zu analysieren.

Die Berücksichtigung der HR ist auch unter **physiologischen Gesichtspunkten** interessant. Im Gegensatz zur Hautleitfähigkeit, die einen Indikator der Sympathikusaktivität darstellt, spiegelt die HR **sowohl sympathische als auch parasympathische Einflüsse** auf das Herz wider. Vereinfacht lassen sich die HR-bezogenen Wirkweisen des Sympathikus und des Parasympathikus als antagonistisch charakterisieren (Birbaumer & Schmidt, 1999, S. 178f.). Während die Aktivität des Herz-Parasympathikus (Vagus) eine Senkung der spontanen HR bewirkt, führt die sympathische Aktivierung zu einem HR-Anstieg (vgl. z. B. Jänig, 1987, S. 239ff.; Schandry, 1998, S. 127ff.). Tierexperimentellen Befunden zufolge beträgt die Latenz von der Reizung des Sinusnerven (pressorezeptorische Afferenz des Vagus) bis zum ersten verlängerten Herzschlagintervall zwischen 200 und 500 Millisekunden; von der Stimulation der Pressorezeptoren bis zur Verringerung der HR vergehen 600 bis 800 Millisekunden (Koepchen, 1982, S. 69). Demgegenüber beläuft sich die Dauer von der Stimulation der sympathischen Herznerven bis zum Anstieg der HR auf 1 bis 3 Sekunden (Levy & Martin, 1979, S. 586). Folglich sind schnelle Änderungen der Herzschlagfrequenz auf Änderungen der vagalen Innervation zurückzuführen (Koepchen, 1982, S. 67; vgl. auch Velden, Karemaker, Wölk & Schneider, 1990, S. 86).

3.2.3 Unterscheidung zwischen den Reaktionen auf die Fragen bzw. Items und den Reaktionen auf den imperativen Reiz der Antwortgabe

Anhand eines Scheinverbrechen-Experiments zum KFT variierte Dawson (1980, S. 9) den **Antwortzeitpunkt** intraindividuell. In mehreren Testdurchgängen sollten die Pbn entweder unmittelbar nach jeder Frage („Immediate Answer“) oder mit einer zeitlichen Verzögerung von 8 Sekunden („Delayed Answer“) antworten. Das Ende der Warte-

periode wurde per Lichtsignal angezeigt. Während bei unmittelbarer Beantwortung die physiologischen Reaktionen auf die Frage und Antwort jeweils ineinander übergingen und somit konfundiert waren, konnte man bei verzögerter Antwortgabe die beiden Reaktionen voneinander trennen. Der Hautwiderstand, der „relative Blutdruck“ und die Atembewegungen wurden numerisch ausgewertet. In der Bedingung „Immediate Answer“ wurden dazu allein die körperlichen Veränderungen nach den unmittelbar beantworteten Fragen herangezogen. In der Bedingung „Delayed Answer“ hingegen analysierte man die Reaktionen auf die Fragen und die Reaktionen auf die verzögerten Antworten getrennt voneinander. Die daraus resultierenden numerischen Scores deuteten darauf hin, daß abgesehen von der „Immediate Answer“-Bedingung nur die *Reaktionen auf die Fragen* der „Delayed Answer“-Bedingung eine zuverlässige Differenzierung zwischen Schuldigen und Unschuldigen ermöglichten. In den Reaktionen auf die *verzögerten Antworten* der „Delayed Answer“-Bedingung zeigten die beiden Gruppen keine signifikant unterschiedlichen Punktwerte. Eine objektive, quantitative Analyse der elektrodermalen und kardiovaskulären Daten stützte diesen Befund. Die beim KFT erwartete Interaktion zwischen Tatbedingung (schuldig vs. unschuldig) und Fragentyp (Relevant vs. Kontroll) manifestierte sich lediglich unter der „Immediate Answer“-Bedingung und in den Reaktionen auf die Fragen der „Delayed Answer“-Bedingung. D. h., weder Schuldige noch Unschuldige zeigten signifikante Reaktionsunterschiede auf die verzögerten Antworten nach den relevanten Fragen und Kontrollfragen. Diese Ergebnisse sind durchaus kompatibel mit der Feststellung, daß bei der psychophysiologischen Aussagebeurteilung keine verbalen Antworten bzw. Lügen nötig sind, sondern allein die Reaktionen auf die Fragen bzw. Items überzufällige Trefferquoten erbringen können (vgl. Abschnitt 2.10.1).

Furedy et al. (1988, S. 684, 1991, S. 92) griffen diesen Ansatz im Rahmen ihres „Differentiation-of-Deception“-Paradigmas (**DDP**) auf. Sie gingen davon aus, man könne unter der Bedingung „Delayed Answer“ **zwei hypothetische Prozesse** voneinander trennen, nämlich **die Intention vs. die eigentliche Durchführung der Täuschung**. Der Vorsatz, zu täuschen bzw. aufrichtig zu sein, äußere sich in den Reaktionen auf die Fragen. Dagegen wurden die körperlichen Begleiterscheinungen der verzögerten Antworten als Korrelate der entsprechenden Handlungen (Lügen vs. Wahrheitsagen) aufgefaßt. Bei unmittelbarer Beantwortung („Immediate Answer“) sollen die Effekte von Absicht und Handlung konfundieren und sich überlagert auf die Reaktionen auswirken. Ähnlich wie Dawson (1980) variierten Furedy und Mitarbeiter den Antwortzeitpunkt als Meßwiederholungsfaktor. Die Antworten wurden entweder direkt oder erst nach einer zehnsekündigen Pause gegeben. Das elektrodermale DDP-Phänomen (stärkere SCR-Magnitude unter der Bedingung Täuschung) trat konsistent bei jenen Reaktionen auf, die im Latenzzeitfenster von 1 bis 5 Sekunden nach Beginn der Fragendarbietung erfaßt

wurden, gleichgültig ob die Vpn unmittelbar oder verzögert antworteten. Darüber hinaus resultierten keine signifikanten Unterschiede zwischen diesen beiden Meßmethoden. Bei den SCRs, die man als Reaktionen auf die verzögerten Antworten wertete, war die Befundlage weniger einheitlich. Während Furedy et al. (1991, S. 95f.) auch unter dieser Bedingung das DDP-Phänomen nachweisen konnten, erbrachte die Untersuchung von Furedy et al. (1988, S. 686f.) im äquivalenten Fall keinen signifikanten Unterschied zwischen Täuschung und Aufrichtigkeit. Die Autoren interpretierten diese Befunde dahingehend, daß das elektrodermale DDP-Phänomen eher auf die Absicht – im Sinne einer Handlungsvorbereitung – zurückzuführen sei und weniger auf die Durchführung der Täuschung. Als eine alternative physiologische Erklärung zogen sie aber auch das Konzept der „Reaktionsinterferenz“ heran („response interference“, Furedy et al., 1988, S. 687). Demnach werde die Stärke einer SCR in Abhängigkeit der Intensität einer zuvor erfolgten Reaktion gedämpft. Diese Annahmen wurden dadurch gestützt, daß die SCRs auf die verzögerten Antworten am schwächsten ausfielen. Furedy und Ben-Shakhar (1991, S. 169) erzielten im Rahmen eines **Kartentest-Experiments** mit Variation des Antwortzeitpunktes ähnliche Ergebnisse. Unter der „Delayed Answer“-Bedingung waren die SCR-Amplituden auf die mit 8 Sekunden Latenz gegebenen Antworten besonders klein und die entsprechenden Reaktionsunterschiede zwischen den relevanten und irrelevanten Items sehr gering. Die Entdeckungsraten lagen deutlich niedriger als bei den SCRs auf die Fragen, aber dennoch signifikant oberhalb des Zufallsniveaus.

Dionisio et al. (2001) fanden in ihrer **DDP-Studie mit der Pupillenweite** vergleichbare Resultate. Dort erfolgte die Beantwortung im Anschluß an eine Wartezeit von 4 Sekunden nach Ende der jeweiligen Frage. Das DDP-Phänomen (größere Pupillendilatation unter der Bedingung Täuschung) konnte sowohl im Verzögerungsintervall als auch in der Antwortphase – nicht jedoch während der akustischen Fragendarbietung – nachgewiesen werden. Der mittlere Reaktionsunterschied zwischen Täuschung und Aufrichtigkeit war am Ende der Warteperiode, also noch vor der Antwortgabe am stärksten ausgeprägt.

In der **Replikation des DDP** von Gödert et al. (2001) wurden die Fragen für die Dauer von 7 Sekunden schriftlich dargeboten und der erforderliche Wahrheitsgehalt durch Signale angezeigt. Die Ausblendung von Frage und Signal auf dem Bildschirm fungierte als imperativer Reiz für die Antwortgabe, so daß auch hier eine Verzögerung der Antwort vorlag. Bei den **elektrodermalen Reaktionen** auf die **Einblendung** der Fragen bzw. Signale resultierten deutlich höhere SCR-Magnituden unter der Bedingung Täuschung. Hinsichtlich der Reaktionen auf die **Ausblendung** waren die Befunde weniger eindeutig (vgl. auch Rill, 1997, S. 106f.). Die Analyse der entsprechenden SCR-Amplituden, die nach der Formel von Venables und Christie (1980, S. 17) zur Annähe-

rung an die Normalverteilung logarithmiert worden waren, erbrachte zwar ebenfalls ein elektrodermales DDP-Phänomen. Die Reaktionsdifferenz zwischen Täuschung und Aufrichtigkeit war jedoch relativ gering. Außerdem manifestierte sich bei den nicht logarithmierten Amplitudenwerten kein signifikanter Haupteffekt des Wahrheitsgehalts. Für die **phasische HR** zeigte sich ebenfalls eine Differenzierung zwischen Täuschung und Aufrichtigkeit. Diese betraf eine akzelerative Komponente der HR-Reaktion im Bereich von 1 bis 7 Sekunden nach Ausblendung von Frage und Signal (geringerer HR-Anstieg unter der Bedingung Täuschung). Auch Gödert et al. (2001, S. 70ff.) versuchten, ihre Daten im Sinne der konzeptionellen Unterscheidung zwischen Täuschungsabsicht und -handlung zu erklären. Diese Interpretation wurde aber durch den Umstand erschwert, daß das elektrodermale DDP-Phänomen primär als Reaktion auf die Fragen bzw. Signale beobachtet wurde, während das kardiovaskuläre DDP-Phänomen erst nach Ausblendung der Stimuli und somit im Zeitraum der Antwortgabe auftrat.

Ähnlich der Vorgehensweise von Gödert et al. werden in der **vorliegenden Studie** die Fragen bzw. Items schriftlich präsentiert und außerdem zusätzlich auditiv dargeboten. Die zeitlich verzögerte Ausblendung auf dem Bildschirm dient wiederum als imperativer Stimulus für die Antwortgabe. Damit will man auch hier die Reaktionen auf die Darbietung und Ausblendung der Fragen bzw. Items trennen, um daraus eventuelle Schlußfolgerungen über die Rolle unterschiedlicher psychophysiologischer Prozesse (z. B. Täuschungsabsicht und -handlung bzw. Reizverarbeitung, Reaktionsvorbereitung und Antwortgabe) beim DLT bzw. GAT ziehen zu können.

3.3 Subjektive Einschätzungen der Reaktionsstärke und der Bedeutsamkeit der Fragen bzw. Items

Im Rahmen psychophysiologischer Problemstellungen wird der **Selbsteinschätzung körperlicher Veränderungen** eine wichtige Funktion beigemessen. Diese zielt insbesondere auf die Bestimmung inter- und intraindividueller Zusammenhänge oder Divergenzen zwischen dem subjektiven Erleben und den physiologischen Variablen ab (Fahrenberg, Walschburger, Foerster, Myrtek & Müller, 1979, S. 46). Solche Aspekte sind auch für die psychophysiologische Aussageforschung von Interesse. Wie bereits im Abschnitt 2.10 erläutert, werden zur Erklärung der differentiellen körperlichen Reaktionen beim KFT bzw. TWT sowohl emotional-motivationale als auch kognitive Ansätze herangezogen. Es ist durchaus anzunehmen, daß die gemessenen Veränderungen der physiologischen Erregung, die mit unterschiedlichen kognitiven Prozessen oder affektiven Zuständen einhergehen, auch der Introspektion zugänglich sind. Das integrative Systemmodell der psychophysiologischen Aussagebeurteilung von Steller (1987, S. 137)

setzt sogar voraus, daß die Selbstwahrnehmung körperlicher Ausdruckserscheinungen und die Interozeption autonomer Erregung entscheidend zur Entstehung differentieller physiologischer Reaktionen beitragen. Darüber hinaus wird postuliert, daß die subjektiv eingeschätzte Bedeutsamkeit der Stimuli und deren **wahrgenommene Relevanz für das Testergebnis** wichtige Bedingungsfaktoren der Reaktionsstärke darstellen.

In der **Analogstudie** zum KFT und TWT von Bradley und Janisse (1981, S. 310) sollten die Pbn nach den Tests die Fragen bzw. Items in eine Rangreihe bringen, je nachdem, wie stark sie ihrer Meinung nach darauf reagiert hatten. Die Ergebnisse dieser Ratings bestätigten die Annahmen des KFT. Im Durchschnitt berichteten Unschuldige intensivere Reaktionen auf die Kontrollfragen. Schuldige gaben an, stärker auf die relevanten Fragen reagiert zu haben. Die Resultate für den TWT wurden nicht dokumentiert. Horowitz et al. (1997, S. 111) verglichen unterschiedliche direkte Befragungstechniken miteinander. Im Anschluß an die Untersuchungen mußten die Vpn die Wichtigkeit der Fragen für den Testbefund und die Stärke der dadurch ausgelösten körperlichen Reaktionen einschätzen. Unabhängig von der Befragungstechnik bewerteten Schuldige und Unschuldige die relevanten Fragen als bedeutsamer für das Testergebnis als die jeweiligen Vergleichsfragen. Die neutralen Stimuli wurden am wenigsten wichtig eingestuft. Ferner berichteten Unschuldige stärkere körperliche Veränderungen auf die Vergleichsfragen als auf die relevanten Fragen. Schuldige schilderten ein inverses Reaktionsmuster. Auf den ersten Blick zeigten die Reaktionsratings eine erstaunliche Übereinstimmung mit den physiologischen Reaktionsprofilen. Dabei ist aber auch zu bedenken, daß Bradley und Janisse (1981, S. 313) nur kleine und nicht signifikante Korrelationen zwischen den subjektiven und objektiven Daten fanden. Ferner muß man wie bei den meisten Selbsteinschätzungsverfahren mit typischen Ergebnisverfälschungen rechnen (Petermann & Noack, 1984, S. 452), z. B. dem Antwortstil der sozialen Erwünschtheit (siehe Heidenreich, 1984, S. 415f.). Demnach ist davon auszugehen, daß die Pbn die vermeintlich an sie gestellten Erwartungen antizipieren und zu differentiellen Beurteilungen der Fragen tendieren, da sie ein solches Reaktionsmuster als sozial erwünscht im Sinne der Meßintention erachten (vgl. auch Bradley & Janisse, 1981, S. 313).

Die **eigene Untersuchung** erfaßt neben den körperlichen Reaktionen zusätzlich erlebnisdeskriptive abhängige Variablen, um festzustellen, ob sich potentielle Unterschiede zwischen den Fragen bzw. Items auch auf der Ebene des subjektiven Erlebens widerspiegeln. In Anlehnung an die Vorgehensweise von Horowitz et al. (1997) werden die Reaktionseinschätzungen für die einzelnen Fragen bzw. Items nach den Tests retrospektiv erhoben. Außerdem bewerten die Pbn, wie wichtig ihrer Ansicht nach die Fragen bzw. Items für die Beurteilung ihrer Glaubwürdigkeit sind. Damit erhält man

auch eine zusätzliche Kontrollmöglichkeit, inwiefern die speziell vom DLT intendierten Bedeutsamkeitsmanipulationen für die unterschiedlichen Fragentypen gelingen.

3.4 Elektrodermale Labilität

Die sog. „**Elektrodermale Labilität**“ (**EL**) ist als ein relativ stabiles psychophysiologisches Personenmerkmal definiert (Vossel, 1990, S. 14f.). Die Operationalisierung erfolgt in der Regel über die Häufigkeit sog. elektrodermaler Spontanfluktuationen während einer reizfreien Ruhephase (siehe auch Boucsein, 1988, S. 357ff.). Als Spontanfluktuationen (NSRs, „non-specific responses“; vgl. Venables & Christie, 1980, S. 9) gelten nicht-reizbezogene phasische Veränderungen der Hautleitfähigkeit, die ein bestimmtes Amplitudenkriterium erfüllen (Schandry, 1998, S. 199ff.). In vielen Untersuchungen wurde ein entsprechender Kriterienwert von 0.02 Mikro-Siemens (µS) herangezogen (z. B. Siddle, O’Gorman & Wood, 1979, S. 522; Siddle, Remington, Kuiack & Haines, 1983, S. 138; Vossel & Rossmann, 1984, S. 98; Zimmer, Vossel & Fröhlich, 1990, S. 252). Eine Person bzw. Personengruppe, die eine relativ hohe Anzahl NSRs pro Zeiteinheit aufweist, bezeichnet man als „elektrodermal labil“; Personen mit einer geringen Häufigkeit von NSRs hingegen als „elektrodermal stabil“ (Vossel, 1990, S. 29f.).

Die **Relevanz der EL** für die Validität der psychophysiologischen Aussagebeurteilung wurde im Rahmen von Experimenten zum KFT und TWT untersucht. Es resultierten relativ uneinheitliche Befunde. So fanden etwa Waid und Orne (1980, S. 4ff.; vgl. auch Waid & Orne, 1981, S. 88f.) sowohl für den KFT als auch für den TWT signifikante Zusammenhänge zwischen den Entdeckungsraten und der EL. Die Ergebnisse deuteten darauf hin, daß elektrodermal stabile, schuldige Pbn eher falsch negative Befunde („glaubwürdig“) erzielten, wohingegen elektrodermal labile Unschuldige häufiger irrtümlich positiv („unglaubwürdig“) klassifiziert wurden. In der Untersuchung von Waid, Wilson und Orne (1981, S. 1122) zeigten sich entsprechende Effekte beim KFT, nicht jedoch beim TWT. Steller und Mitarbeiter (vgl. Steller, 1987, S. 74) fanden ebenfalls keine moderierenden Einflüsse der EL auf die Validität des TWT. Im Gegensatz dazu resultierte im TWT-Experiment von Schumacher (1993, S. 108) für schuldige Pbn, die keine Instruktion bezüglich ihrer Aufmerksamkeit auf die Items erhielten (vgl. Abschnitt 2.10.2), eine signifikante positive Korrelation zwischen der NSR-Häufigkeit und den Lykken-Scores. D. h., labile Schuldige wurden dort eher entdeckt als stabile. Eine recht extensive Laborstudie zur Bedeutung interindividueller Unterschiede in der elektrodermalen Aktivität (EDA) für die psychophysiologische Aussagebeurteilung stammt von Horneman und O’Gorman (1987). Die Autoren interpretierten ihre Ergebnisse zum

KFT und TWT als kompatibel mit denen von Waid und Mitarbeiten, da die EL jeweils mit der Entdeckbarkeit Schuldiger korrelierte.

Die insgesamt **heterogene Befundlage** ist möglicherweise auf verschiedenartige und teils methodisch unzulängliche experimentelle Vorgehensweisen zurückzuführen. Z. B. wurde die EL mitunter nicht als quasi-unabhängige Variable in die Versuchsplanung einbezogen, sondern post hoc berücksichtigt und erhoben (Steller, 1987, S. 74; zur Kritik an einer solchen Methodik siehe auch Vossel, 1990, S. 15).

Auch **Literaturübersichten** zum Einfluß der EL auf die Validität von KFT und TWT vermitteln ein relativ inkonsistentes Bild. Während Steller (1987, S. 74) die Ergebnisse als „wenig aussagekräftig“ einstuft, kommt Berning (1992) zu dem Schluß, daß wir „derzeit, aufgrund der defizitären Erkenntnislage, von erhöhten Fehlerquoten sowohl bei elektrodermal labilen als auch bei elektrodermal stabilen Personen in der jeweils unterschiedlichen Richtung ausgehen müssen“ (S. 147). Damit ist gemeint, daß v. a. beim KFT für elektrodermal Labile eine erhöhte Rate falsch positiver Befunde und für elektrodermal Stabile ein größeres Risiko für falsch negative Befunde erwartet wird. Mit Verweis auf die Untersuchungen von Waid und Mitarbeitern sowie Horneman und O’Gorman (s. o.) gehen Ben-Shakhar und Furedy (1990, S. 78) ebenfalls davon aus, daß die EL ein Korrelat der Treffsicherheit darstellt. Als Fazit fordert Berning (1992) sogar, „die elektrodermale Labilität ... bei *jeder* lügendetektorischen Untersuchung insbesondere mit dem Kontrollfragentest mitzuerfassen“ (S. 147).

Das **eigene Experiment** soll einen zusätzlichen Beitrag zur Klärung der Frage leisten, inwiefern die EL bei der psychophysiologischen Aussagebeurteilung von Belang ist. Darüber hinaus wird der Einfluß dieser Variable erstmals im Zusammenhang mit dem DLT und dem GAT untersucht. Falls die EL bei diesen Verfahren eine Rolle spielt, sollten unterschiedliche Reaktionsprofile für elektrodermal Labile und Stabile auftreten, d. h. insbesondere eine Interaktion zwischen der EL und den verschiedenen Fragen- bzw. Itemtypen.

Wenn außerdem, wie im vorliegenden Fall, Hautleitfähigkeitsreaktionen als abhängige Variable erfaßt werden, ist die Berücksichtigung der EL auch unter **methodischen Gesichtspunkten** sinnvoll. Elektrodermal labile und stabile Personen unterscheiden sich hinsichtlich der Stärke ihrer SCRs auf gleiche Reize und in bezug auf die Habituationsverläufe nach wiederholter Reizdarbietung. In zahlreichen Untersuchungen hat man Zusammenhänge zwischen der Habituationsgeschwindigkeit der SCRs und der Häufigkeit von Spontanfluktuationen gefunden (z. B. Coles, Gale & Kline, 1971, S. 60f.; Dickinson & Smith, 1973, S. 410; Katkin & McCubbin, 1969, S. 57f.). Insbe-

sondere bei bedeutungsvollen Stimuli habituieren elektrodermal Labile langsamer als Stabile (zusammenfassend Vossel, 1990, S. 227). Nach Fahrenberg (1969) kann die Vernachlässigung derartiger interindividueller Unterschiede im Rahmen psychophysiologischer Untersuchungen sowohl eine erhöhte Fehlervarianz als auch prinzipielle Fehler und somit eine Verfälschung der experimentellen Befunde nach sich ziehen.

Die EL ist jedoch nicht nur im Zusammenhang mit elektrodermalen Reaktionen von Interesse. Verschiedene Untersuchungen legen nahe, daß sich Labile und Stabile auch **in anderen physiologischen Variablen unterscheiden**, wie z. B. der phasischen HR. Schell, Dawson und Filion (1988, S. 624) fanden z. B. für Labile eine stärkere HR-Dezeleration als Reaktion auf Töne mit und ohne Signalcharakter. Im Wahlreaktionszeit-Experiment von Zimmer et al. (1990, S. 255) resultierte eine Interaktion zwischen der EL und der Ruhe-HR (hoch vs. niedrig). Stabile mit hoher Ruhe-HR zeigten eine ausgeprägtere antizipatorische HR-Dezeleration als Labile mit hoher Ruhe-HR. Gödert et al. (2001, S. 67ff.) beobachteten im „Differentiation-of-Deception"-Paradigma (DDP, vgl. Abschnitt 2.9.3) wiederum eine stärkere Dezeleration für Labile. Es traten aber keine Interaktionen mit dem Wahrheitsgehalt auf. D. h., die EL übte keine signifikanten Effekte auf die Reaktionsunterschiede zwischen Täuschung und Aufrichtigkeit aus. Besonders bemerkenswert ist allerdings die bereits angesprochene Studie von Waid, Wilson und Orne (1981, S. 1122). Dort konnte der Zusammenhang zwischen EL und Entdeckbarkeit im KFT nicht nur für die SCRs, sondern ebenso für kardiovaskuläre und respiratorische Maße nachgewiesen werden. Unabhängig von Schuld oder Unschuld zeigten elektrodermal Labile im Vergleich zu Stabilen in der Hautleitfähigkeit, im Anstieg des relativen Blutdrucks und in der Reduktion der Atemtiefe häufiger stärkere Reaktionen auf die relevanten Fragen als auf die entsprechenden Kontrollfragen.

3.5 Forschungsfragen und Hypothesen

Im folgenden werden die wesentlichen Forschungsfragen und Hypothesen nochmals zusammenfassend aufgelistet. Die Darstellung ist in **zwei Teilabschnitte** untergliedert: Der erste beschreibt die Annahmen zu den Effekten der unterschiedlichen Fragen- bzw. Itemtypen, der zweite die Hypothesen für die zusätzlich berücksichtigten Variablen.

3.5.1 Forschungsfragen hinsichtlich der Reaktionsunterschiede zwischen den Fragen- bzw. Itemtypen

Forschungsfrage 1: Lassen sich in einem Scheinverbrechen-Experiment mit standardisierter Durchführung und objektiver Auswertung der psychophysiologischen Aussagebeurteilung körperliche Reaktionsunterschiede zwischen den verschiedenen Fragen- bzw. Itemtypen des DLT bzw. GAT nachweisen?

Die bisherigen Analogstudien deuten darauf hin, daß die erwarteten Effekte der Fragen bzw. Items beim DLT bzw. GAT auch unter simulierten Tat- und Testbedingungen zu beobachten sind. Darüber hinaus dürfte eine erhöhte Standardisierung (z. B. vollautomatische Stimuluspräsentation) nicht zwangsläufig die Funktionsweise der Verfahren beeinträchtigen (vgl. auch Bradley & Warfield, 1984, S. 686; Honts et al., 1995, S. 205). In bezug auf die Richtung der Reaktionsunterschiede müssen aber die Besonderheiten der vorliegenden Studie, d. h. speziell die intraindividuellen Variationen der Täterschaft und des Wahrheitsgehalts der DLT-Kontrollfragen, berücksichtigt werden.

Forschungsfrage 2: Reagieren Pbn auf die Fragen bzw. Items nach einer Tat, die sie durchgeführt haben und deshalb wahrheitswidrig abstreiten, stärker als auf äquivalente Fragen bzw. Items nach einer nicht begangenen Tat, die sie wahrheitsgemäß verneinen?

Die Annahmen und empirischen Befunde der direkten und indirekten Verfahren der psychophysiologischen Aussagebeurteilung (vgl. Abschnitt 3.1.3) legen nahe, daß die Pbn intensivere Reaktionen auf die Fragen oder Items nach der von ihnen verübten Tat (Relevant-Täuschung) zeigen als auf entsprechende wahrheitsgemäß verneinte Stimuli (Relevant-Aufrichtigkeit).

Forschungsfrage 3: Hat die Variation des Wahrheitsgehalts bei den Kontrollfragen des DLT einen Effekt auf die Reaktionsstärke?

Diesbezüglich läßt sich nur schwer eine Unterschiedshypothese ableiten, zumal bislang keine entsprechenden systematischen Manipulationen vorgenommen wurden. Ferner erlauben die widersprüchlichen Schlußfolgerungen, die man aus den Ergebnissen zum DDP bzw. KFT ziehen kann (vgl. Abschnitt 3.1.4), keine eindeutigen Vorhersagen. Falls jedoch der Wahrheitsgehalt eine Rolle spielt, wäre zumindest gemäß dem DDP davon auszugehen, daß die instruierten Lügen-Kontrollfragen intensivere Reaktionen evozieren als die Wahrheit-Kontrollfragen.

Die bislang geäußerten Annahmen beziehen sich auf spezifische Problemstellungen der vorliegenden Studie. Darüber hinaus lassen sich **zusätzliche Hypothesen** hinsichtlich der zu erwartenden Reaktionsunterschiede aufstellen, die sich aus den Prämissen der Befragungstechniken ergeben.

Nach Lykken (1998) lautet die wichtigste und gleichzeitig besonders problematische Grundannahme aller Kontrollfragentechniken und somit auch des **DLT**, daß die Reaktionen einer Person auf die **Kontrollfragen** zwischen den Reaktionen auf die relevanten Fragen nach einer begangenen vs. nicht begangenen Tat anzusiedeln sind:

> A subject's arousal response elicited by the control question (R_C) will be smaller than his response would be to the relevant question if he is guilty (R_G) but larger than his response would be if he is innocent (R_I); that is $R_G > R_C > R_I$ for all subjects. (Lykken, 1998, S. 122)

Es sei vermerkt, daß die Überprüfung dieser Hypothese wiederum einen intraindividuellen Vergleich impliziert, zumal sie die bereits im Abschnitt 3.1.3 geäußerte Prämisse ($R_G > R_I$) beinhaltet. Übertragen auf die vorliegende Untersuchung würde man daraus folgern, daß die Reaktionen auf die Lügen-Kontrollfragen zwischen den Reaktionen auf die wahrheitsgemäß bzw. wahrheitswidrig beantworteten tatbezogenen Fragen (Relevant-Aufrichtigkeit bzw. Relevant-Täuschung) liegen sollten.

Für die **irrelevanten Fragen des DLT** lassen sich kaum gerichtete Hypothesen formulieren, da sie in der Regel nicht in die Auswertungen eingehen. Dennoch postulierte Lykken (1998, S. 121), daß Kontrollfragen eine größere körperliche Erregung auslösen können als irrelevante Fragen. Außerdem deutete die Analogstudie von Horowitz et al. (1997, S. 112f.) darauf hin, daß sowohl Schuldige als auch Unschuldige auf die wahrheitsgemäß beantworteten irrelevanten Fragen der Relevant-Irrelevant-Technik besonders schwach reagierten, woraus hohe Raten valider und falsch positiver Befunde resultierten. Unter Verwendung konventioneller bzw. instruierter Lügen-Kontrollfragen stiegen die Trefferquoten für Unschuldige deutlich an. Demzufolge wäre mit stärkeren

Reaktionen auf die Kontrollfragen als auf die irrelevanten Stimuli zu rechnen. Die Übertragbarkeit dieser Ergebnisse auf die vorliegende Untersuchung ist jedoch zweifelhaft, da sie auf Gruppenvergleichen und numerischen Scores beruhten.

Für die **irrelevanten Items des GAT**, die bei der Auswertung berücksichtigt werden, sind schwächere Reaktionen zu erwarten als auf die wahrheitsgemäß beantworteten tatbezogenen Items (Relevant-Aufrichtigkeit). Diese Hypothese läßt sich einerseits daraus ableiten, daß der GAT zwar signifikant zwischen Schuldigen und Unschuldigen mit Tatwissen differenziert, letztere aber dennoch mit einem höheren Risiko falsch positiv klassifiziert werden als Unschuldige ohne entsprechende Kenntnisse (vgl. Abschnitt 2.8). Außerdem konnten Elaad und Ben-Shakhar (1989, S. 450) anhand einer Reanalyse der Daten von Bradley und Warfield (1984) zeigen, daß Tatwissen eine hinreichende Bedingung für positive GAT-Befunde darstellte. Der durchschnittliche numerische Punktwert der Unschuldigen mit Tatwissen lag signifikant über dem Zufallsniveau. D. h., ungeachtet der wahrheitsgemäßen Beantwortung dürften die relevanten Items des GAT stärkere Reaktionen evozieren als die irrelevanten, sofern ihr Tatbezug erkannt wird.

Beim Lykken-Scoring des TWT bzw. GAT werden die jeweils **ersten irrelevanten Alternativen pro Multiple-Choice-Frage** nicht berücksichtigt, da sie aufgrund der Neuheit der angesprochenen Thematik eine besonders hohe körperliche Erregung auslösen sollen. Diese Stimuli können also mit sehr intensiven Reaktionen einhergehen, die möglicherweise sogar das Niveau der wahrheitswidrig verneinten tatbezogenen Items (Relevant-Täuschung) übertreffen. Ähnlich wie bei den irrelevanten Fragen des DLT kann man diese Annahme jedoch auf keine fundierte Datenbasis stützen.

3.5.2 Forschungsfragen hinsichtlich der zusätzlich berücksichtigten Variablen

Forschungsfrage 4: Manifestieren sich die erwarteten Reaktionsunterschiede zwischen den Fragen- bzw. Itemtypen sowohl in elektrodermalen als auch in kardiovaskulären Variablen?

In der Mehrzahl der bisherigen Untersuchungen hat sich die **Hautleitfähigkeitsreaktion** (SCR) als effektivste körperliche Variable der psychophysiologischen Aussagebeurteilung erwiesen (Abschnitt 3.2.1). Daher ist zu erwarten, daß sich eventuelle Reaktionsdifferenzen v. a. in der **SCR-Amplitude** zeigen.

Auf seiten der **kardiovaskulären Maße** ist die Datenbasis weniger eindeutig. Orientiert man sich an den Studien von Raskin und Hare (1978), Podlesny und Raskin (1978) sowie Bradley und Janisse (1981), dann sollten die Unterschiede zwischen den einzelnen Fragen- bzw. Itemtypen v. a. in einer dezelerativen Komponente der **Herzschlagfrequenz** (HR) auftreten. Die relevante Reaktionsrichtung bestünde somit in einer verschieden starken HR-Dezeleration. Diese Annahme ist aber noch nicht hinreichend abgesichert. Andere Experimente zur psychophysiologischen Aussagebeurteilung konnten die entsprechenden Ergebnisse nicht replizieren (vgl. Abschnitt 3.2.2). Außerdem fanden Gödert et al. (2001, S. 67ff.) ein kardiovaskuläres DDP-Phänomen in einer akzelerativen Komponente mit einem schwächeren HR-Anstieg unter der Bedingung Täuschung (vgl. Abschnitt 3.2.3). Es sei aber nochmals betont, daß man einen derartigen Befund nicht direkt auf die vorliegende Untersuchung übertragen kann. Darum werden hier die Verläufe der phasischen HR jeweils über mehrere Sekunden nach Stimulusdarbietung echtzeitskaliert und in Abhängigkeit der Fragen- bzw. Itemtypen analysiert.

Forschungsfrage 5: Treten die erwarteten Reaktionsunterschiede gleichermaßen auf die Präsentation der Fragen bzw. Items als auch nach Darbietung des imperativen Reizes der Antwortgabe auf?

Angesichts der bisherigen Experimente, die den Antwortzeitpunkt verzögerten (Abschnitt 3.2.3), kann man zumindest für die **SCRs** postulieren, daß sich die Reaktionsunterschiede zwischen den Fragen- bzw. Itemtypen bereits im Anschluß an deren Präsentation zeigen. Die Reaktionsunterschiede nach Ausblendung (imperativer Reiz der Antwortgabe) dürften entweder geringer ausfallen oder sogar vollständig ausbleiben. Bei der phasischen **HR** sind keine eindeutigen Hypothesen möglich. Allein die DDP-Studie von Gödert et al. (2001, S. 67ff.) erbrachte eine bislang unbestätigte Differenzierung zwischen Täuschung und Aufrichtigkeit in einer akzelerativen HR-Komponente im Anschluß an die Ausblendung der Fragen.

Forschungsfrage 6: Spiegeln sich die physiologischen Reaktionsdifferenzen auch im subjektiven Erleben der Pbn wider, und wird die Relevanz der Fragen- bzw. Itemtypen für das Testergebnis unterschiedlich bewertet?

Die Analogstudien von Bradley und Janisse (1981) sowie Horowitz et al. (1997) weisen darauf hin, daß die Reaktionsunterschiede introspektiv wahrgenommen werden können und die Pbn die diversen Stimulustypen als unterschiedlich bedeutsam für den Ausgang der psychophysiologischen Aussagebeurteilung erachten.

Forschungsfrage 7: Unterscheiden sich elektrodermal labile und stabile Pbn in der Stärke der Reaktionsdifferenzen zwischen den Fragen- bzw. Itemtypen des DLT bzw. GAT?

Bezüglich dieser Problemstellung ist die vorliegende Studie explorativer Natur, da die EL noch nicht im Zusammenhang mit den beiden neueren Befragungstechniken untersucht wurde. Auch die im Abschnitt 3.4 berichteten inkonsistenten Befunde zum KFT und TWT bieten wenig Anhaltspunkte. Unter der Annahme, daß die EL beim DLT bzw. GAT von Bedeutung ist, wäre eine Wechselwirkung mit den Fragen- bzw. Itemtypen zu erwarten. Falls darüber hinaus der sich abzeichnende Trend einer entgegengesetzten Fehleranfälligkeit (Labile eher falsch positiv, Stabile eher falsch negativ) übertragbar ist, sollten Labile stärkere Reaktionsunterschiede speziell zwischen den wahrheitsgemäß verneinten tatbezogenen Stimuli (Relevant-Aufrichtigkeit) und den Vergleichsreizen (Lügen-Kontrollfragen bzw. irrelevante Items) zeigen. Im Gegensatz dazu dürften Stabile eher eine Nivellierung des Reaktionsprofils aufweisen, insbesondere mit geringeren Differenzen zwischen Relevant-Täuschung und den entsprechenden Vergleichsreizen.

4. Methode

4.1 Versuchsplanung

4.1.1 Versuchsgruppen

Die Untersuchung umfaßte **acht experimentelle Gruppen**, die sich aus der Kombination der jeweils zweistufigen Faktoren Tatbedingung, Testart und Elektrodermale Labilität ergaben (vgl. Tabelle 5).

Tabelle 5. Versuchsgruppen

	Tatbedingung			
	Ring		Kette	
	Elektrodermale Labilität		Elektrodermale Labilität	
Testart	Stabil	Labil	Stabil	Labil
DLT	*n* = 10	*n* = 10	*n* = 10	*n* = 10
GAT	*n* = 10	*n* = 10	*n* = 10	*n* = 10

Die **Tatbedingung** (Ring vs. Kette) wurde interindividuell variiert, indem man jeweils die Hälfte der Probanden (Pbn) instruierte, im Rahmen eines Scheinverbrechens einen Fingerring oder eine Halskette zu entwenden. Die Versuchspersonen (Vpn) wurden zu Beginn des Experiments pseudorandomisiert den Gruppen zugeordnet. Die beiden simulierten Diebstähle waren äquivalent gestaltet. Die Pbn sollten jeweils nur einen davon durchführen. Bei der Tatbegehung wurden sie jedoch auch mit den kritischen Details des jeweiligen anderen Scheinverbrechens konfrontiert.

Nach dem Scheinverbrechen sollten die Vpn eine psychophysiologische Aussagebeurteilung absolvieren. Diese erfolgte anhand zweier unterschiedlicher Testverfahren, die in Anlehnung an den Directed Lie Test und den Guilty Actions Test konstruiert waren. Die **Testart** (DLT vs. GAT) variierte unabhängig von der Tatbedingung zwischen den Personen. Jeweils 50% der Pbn wurden mit einem der beiden Verfahren getestet. Bei der statistischen Auswertung der physiologischen und subjektiven Reaktionen auf die Stimuli ging die Testart nicht als unabhängige Variable in die Analysen ein, d. h., diese wurden für den DLT und GAT getrennt durchgeführt, ohne daß man die beiden Testbedingungen direkt miteinander verglich.

Um den Einfluß der **Elektrodermalen Labilität** (EL) untersuchen zu können, wurde diese als Gruppenfaktor mit den Ausprägungen Stabil und Labil in den Versuchsplan aufgenommen. Die Kennwertbildung erfolgte nach der Häufigkeitsmethode (Vossel,

1990, S. 67ff.). Vor der psychophysiologischen Aussagebeurteilung bestimmte man für jede Vp innerhalb einer fünfminütigen reizfreien Ruhemessung die Anzahl der aufgetretenen elektrodermalen Spontanfluktuationen (NSRs), die ein Amplitudenkriterium von mindestens 0.02 µS erfüllten (vgl. Abschnitt 3.4). Die Pbn wurden anhand einer Mediandichotomisierung (Mediansplit) in elektrodermal Stabile und Labile eingeteilt. Die Datensätze jener Vpn, deren NSR-Anzahl exakt dem Median entsprach, gingen nicht in die weitere Auswertung ein (für eine deskriptive Statistik der NSR-Anzahl vgl. Anhang B: Tabelle 28).

Für jede der insgesamt acht Versuchsgruppen wurde eine **Zellbesetzung von $n = 10$** Pbn veranschlagt, woraus eine **Gesamtstichprobe von $N = 80$** Personen resultierte.

4.1.2 Meßwiederholungsfaktoren

Der DLT und der GAT setzten sich aus mehreren **Fragen- bzw. Itemtypen** zusammen, die bei den betreffenden Pbn jeweils intraindividuell variiert wurden.

Der **DLT** bestand aus **fünf Fragentypen** (vgl. Tabelle 6). Die tatbezogenen Fragenpaare waren inhaltlich parallelisiert und zielten auf das begangene (Relevant-Täuschung) vs. nicht begangene Scheinverbrechen (Relevant-Aufrichtigkeit) ab. Die ebenfalls parallelisierten Kontrollfragenpaare wurden instruiert wahrheitswidrig oder wahrheitsgemäß verneint (Lügen-Kontroll, Wahrheit-Kontroll) und die neutralen Fragen (Irrelevant) wahrheitsgemäß bejaht. Für jeden der fünf Typen gab es drei verschiedene Fragen, die in zwei Durchgängen wiederholt dargeboten wurden, so daß pro Bedingung sechs Trials vorlagen. Der DLT umfaßte somit insgesamt 30 Trials.

Tabelle 6. Fragentypen des DLT

	Fragentyp				
	Relevant-Täuschung	Relevant-Aufrichtigkeit	Lügen-Kontroll	Wahrheit-Kontroll	Irrelevant
Trials	1, 2, ... , 6	1, 2, ... , 6	1, 2, ... , 6	1, 2, ... , 6	1, 2, ... , 6

Der **GAT** bestand aus sechs Multiple-Choice-Fragen mit jeweils sechs Alternativen, die sich zu **vier Itemtypen** gruppieren ließen (vgl. Tabelle 7). Die tatbezogenen Items (Relevant-Täuschung vs. Relevant-Aufrichtigkeit) thematisierten Details des begangenen vs. nicht begangenen Scheinverbrechens. Diese relevanten Alternativen waren stets eingebettet in vier irrelevante Items und nahmen nie die erste Position unmittelbar nach der Frage ein. Die irrelevanten Items wurden gemäß ihrer Reihenfolge innerhalb der

Multiple-Choice-Fragen durchnumeriert (Irrelevant-1 bis Irrelevant-4), d. h., bei einem irrelevanten Item mit der Nummer 1, 2, 3 oder 4 handelte es sich jeweils um die erste, zweite, dritte oder vierte irrelevante Alternative einer Multiple-Choice-Frage. Die Pbn verneinten diese nicht tatbezogenen Items wahrheitsgemäß. Da beim GAT die Reaktionen auf die ersten Alternativen (Irrelevant-1) nicht in die Auswertung eingehen, bildeten diese eine separate Bedingung. Demgegenüber wurden die Items Irrelevant-2 bis Irrelevant-4, die beim GAT als Vergleichsbedingungen für die tatbezogenen Alternativen dienen, zum Itemtyp Irrelevant-Vergleich zusammengefaßt. Pro Multiple-Choice-Frage gab es sechs Trials, wobei die eigentliche Frage und das erste irrelevante Item (Irrelevant-1) stets unmittelbar nacheinander, d. h. in dem gleichen Trial dargeboten wurden. Insgesamt resultierten daraus für den GAT 36 Trials.

Tabelle 7. Itemtypen des GAT

	Itemtyp					
	Relevant-Täuschung	Relevant-Aufrichtigkeit	Irrelevant-Vergleich			Irrelevant-1
			Irrelevant-2	Irrelevant-3	Irrelevant-4	
Trials	1, 2, ... , 6	1, 2, ... , 6	1, 2, ... , 6	1, 2, ... , 6	1, 2, ... , 6	1, 2, ... , 6

Sowohl beim DLT als auch beim GAT wurden die Fragen bzw. Items, die sich auf den Ring- oder Kettendiebstahl bezogen, in Abhängigkeit von der Tatbedingung den beiden Bedingungen Relevant-Täuschung vs. -Aufrichtigkeit zugewiesen (siehe Tabelle 8; vgl. auch Abschnitt 4.4.3), d. h., man konnte die relevanten Stimuli sowohl unter dem Gesichtspunkt ihrer Beantwortung (Wahrheitsgehalt: Täuschung vs. Aufrichtigkeit) als auch im Hinblick auf ihren Tatbezug (Schmuckstück: Ring vs. Kette) analysieren. Im ersten Fall sind unabhängig von der Tat stärkere Reaktionen unter der Bedingung Relevant-Täuschung als unter Relevant-Aufrichtigkeit zu erwarten. Im zweiten Fall ist mit einer Interaktion zwischen der Tatbedingung und dem Tatbezug der Fragen bzw. Items zu rechnen. Demnach dürften die Pbn der Tatbedingung Ring stärker auf die Relevant-Ring-Stimuli reagieren, während die Pbn der Bedingung Kette stärker auf die Fragen bzw. Items nach dem Kettendiebstahl reagieren sollten.

Tabelle 8. Intraindividuelle Variation von Täuschung vs. Aufrichtigkeit bei den relevanten Fragen des DLT bzw. Items des GAT in Abhängigkeit von der Tatbedingung

	Fragen- bzw. Itemtyp	
Tatbedingung	Relevant-Täuschung	Relevant-Aufrichtigkeit
Ring	Relevant-Ring	Relevant-Kette
Kette	Relevant-Kette	Relevant-Ring

4.1.3 Abhängige Variablen

Als abhängige Variablen wurden sowohl **physiologische Kennwerte** als auch **subjektive Daten** erfaßt. Die gemessenen physiologischen Größen waren die Hautleitfähigkeit, das Elektrokardiogramm (EKG), die periphere Durchblutung am Finger und die Atmung (vgl. Abschnitt 4.5). Bei den erlebnisdeskriptiven abhängigen Variablen handelte es sich um retrospektiv erhobene Einschätzungen der Reaktionsstärke und der Bedeutsamkeit der unterschiedlichen Fragen bzw. Items (vgl. Abschnitt 4.6).

4.1.4 Zusätzliche Kontrollvariablen

Im Anschluß an die subjektiven Ratings wurden schriftliche **Gedächtnistests** durchgeführt. Ausschlaggebend dafür waren zunächst Überlegungen zum **GAT**. Eventuell auftretende Reaktionsunterschiede zwischen den wahrheitsgemäß und wahrheitswidrig beantworteten relevanten Items konnten auch durch Gedächtniseffekte bedingt sein. Eine Mutmaßung bestand darin, daß die Pbn die Details des von ihnen begangenen Scheinverbrechens besser erinnerten bzw. wiedererkannten. Möglicherweise wurden die relevanten Items der nicht begangenen Tat eher vergessen und waren somit schlechter von den irrelevanten Items zu unterscheiden. Dieses Problem haben bereits Bradley, MacLaren und Carle (1996, S. 154) im Zusammenhang mit der Testung von informierten Unschuldigen angesprochen. Sofern unschuldige Pbn bereits vor der psychophysiologischen Aussagebeurteilung Kenntnis von den kritischen Tatdetails erlangt haben, diese aber anschließend schlechter wiedererkennen als die schuldigen, dann dürften sie insgesamt auch schwächere Reaktionen auf die relevanten Items zeigen.

Aber auch bei Kontrollfragentests sollten potentielle **Gedächtniseffekte kontrolliert** werden. Die Untersuchung von O'Toole et al. (1994, S. 262) deutete darauf hin, daß unter Umständen die Entdeckbarkeit Schuldiger mittels des KFT durch Erinnerungsdefizite in bezug auf die Tat beeinträchtigt werden kann. Solche Einflüsse waren auch im Hinblick auf den hier verwendeten **DLT** zu beachten. Per Definition ist eine Täuschung stets vom subjektiven Kenntnisstand der kommunizierenden Person abhängig (vgl. Abschnitt 2.1). Nur wenn ein Pb entgegen besserem Wissen eine Falschaussage macht, handelt es sich um eine Lüge. Falls z. B. ein Pb das von ihm entwendete Schmuckstück nicht mehr erinnern bzw. wiedererkennen kann, dann dürften auch die Unterschiede zwischen den entsprechenden tatbezogenen Fragen (Relevant-Täuschung, Relevant-Aufrichtigkeit) nivelliert werden.

Die **Überprüfung der mnestischen Reproduktionsleistung** orientierte sich an den Studien von Bradley und Mitarbeitern (z. B. Bradley & Rettinger, 1992, S. 57; Bradley & Warfield, 1984, S. 686) sowie O'Toole et al. (1994, S. 257). Zuerst sollten die Pbn einen Wiedererinnerungstest mit Hinweisreizen und anschließend einen Wiedererkennungstest absolvieren („cued-recall test" und „recognition test", vgl. O'Toole et al., 1994, S. 257). Es wurde festgelegt, daß die Datensätze jener Vpn, die weniger als zwei Drittel der kritischen Tatdetails (s. u.) erinnern bzw. wiedererkennen konnten, von der Auswertung auszuschließen waren.

Nach den Gedächtnistests sollte jeweils die **Täuschungsmotivation** der Vpn und die von ihnen **subjektiv eingeschätzte Treffsicherheit** der psychophysiologischen Aussagebeurteilung erfaßt werden. Anhand dieser Ratings wollte man insbesondere überprüfen, ob die Versuchsgruppen diesbezüglich Unterschiede zeigten. Potentielle Kovariationen könnten sich als relevant erweisen. Eine hohe Motivation, als glaubwürdig eingestuft zu werden, und die Überzeugung der Pbn, die Methode sei zuverlässig, gelten in der psychophysiologischen Aussageforschung mitunter als wichtige Voraussetzungen für valide Testergebnisse (zusammenfassend Ben-Shakhar & Furedy, 1990, S. 60ff.; Steller, 1987, S. 61ff.; vgl. auch Abschnitt 2.10.1). Außerdem wurde hier – wie bei Scheinverbrechen-Experimenten üblich – durch entsprechende Maßnahmen (finanzieller Anreiz, Instruktion, Stimulationstest) eine Erhöhung der Täuschungsmotivation und der subjektiv eingeschätzten Treffsicherheit angestrebt. Somit dienten die entsprechenden Beurteilungsverfahren auch der Kontrolle, inwiefern man diese Zielsetzung realisieren konnte.

Als **Nachbefragung** war ein strukturiertes, halbstandardisiertes Interview vorgesehen. Eine Frage zielte auf **eventuelle Manipulationsversuche** der Vpn zur Vermeidung einer erfolgreichen „Lügendetektion" ab (z. B. Entspannungstechniken, gedankliche Ablenkung, Herbeiführen körperlicher Reaktionen etc.). Falls die Vpn diese Frage bejahten, wurden sie zusätzlich gebeten, ihre Strategien näher zu erläutern. Sowohl der KFT als auch der TWT haben sich als anfällig gegenüber körperlichen und mentalen Beeinflussungsversuchen („countermeasures") erwiesen, die schuldige Pbn unter Umständen durchführen, um einen irrtümlich negativen Befund („glaubwürdig") herbeizuführen (z. B. Honts et al., 1994, S. 257, 1996, S. 90). Obwohl die Vpn der vorliegenden Untersuchung nicht über Manipulationstechniken informiert wurden, mußte man aufgrund der angestrebten Erhöhung der Täuschungsmotivation damit rechnen, daß einige Pbn Maßnahmen gegen eine „erfolgreiche" Testung ergreifen würden. Unter der Annahme, daß diese Manipulationsversuche auch beim DLT und GAT Auswirkungen auf die Testergebnisse haben können, sollten sie hier ebenfalls berücksichtigt und kontrolliert werden.

In einer weiteren Frage sollten die Pbn angeben, ob sie beim Lügen und Wahrheitsagen **differentielle Reaktionen** bzw. körperliche Empfindungen bemerkt hätten, und wenn ja, welche. Aus den Schilderungen erhoffte man sich zusätzliche qualitative Informationen hinsichtlich der introspektiv wahrgenommenen Begleiterscheinungen nach Darbietung der unterschiedlichen Fragen- bzw. Itemtypen, die über die quantitativ erhobenen Befindensäußerungen (s. o.) hinausgingen.

Abschließend wurde das **Versuchserleben** der Vpn in bezug auf die gesamte Untersuchung erfaßt. Hierbei war von besonderem Interesse, wie der experimentelle Kontext aus Sicht der Pbn beurteilt wurde. Jene Charakteristika der äußeren Bedingungen, die eventuell als unangenehm oder störend erlebt wurden, könnten als Anhaltspunkte für methodische Verbesserungen nachfolgender Untersuchungen fungieren.

4.2 Versuchspersonen

Die **Anwerbung** der Vpn erfolgte überwiegend durch direktes Ansprechen sowie über Plakate bzw. Handzettel, die auf dem Campus der Johannes Gutenberg-Universität, auf dem Gelände der Universitätsklinik und im Stadtgebiet Mainz aufgehängt bzw. verteilt wurden. Weitere Teilnehmer meldeten sich auf Inserate in einer lokalen Studentenzeitung und in einer Internet-Newsgroup. Darüber hinaus wurden einige Pbn durch Tageszeitungs- und Rundfunkberichte auf das Experiment aufmerksam. Im Rahmen eines ersten telefonischen Kontakts erhielten die Interessenten eine grobe Beschreibung der Untersuchung (Scheinverbrechen-Experiment mit anschließender „Lügendetektion“), ohne daß dabei auf Details eingegangen wurde. Allen Teilnehmern stellte man eine Belohnung von 20 DM für das erfolgreiche Bestehen eines „Lügendetektortests“ in Aussicht.

Insgesamt nahmen 122 Männer im Alter zwischen 19 und 70 Jahren freiwillig an der Untersuchung teil. Zwei Pbn brachen das Experiment vor dem Scheinverbrechen ab. Bei zwei weiteren Teilnehmern wurde die Messung von seiten des Versuchsleiters (Vl) abgebrochen. Einer der beiden Pbn zeigte überhaupt keine Hautleitfähigkeitsreaktionen. Im anderen Fall lagen starke Herzrhythmusstörungen vor, die eine reguläre Aufzeichnung der kardiovaskulären Reaktionen verhinderten.

Von den restlichen 118 Datensätzen wurden weitere 38 aus unterschiedlichen Gründen von der Auswertung **ausgeschlossen**. Bei drei Teilnehmern mußte der jeweilige Test (wegen mangelndem Instruktionsverständnis bzw. nach technischen Systemfehlern) abgebrochen und neu gestartet werden. Zwei Pbn bewegten sich während den physiolo-

gischen Messungen mehrfach so heftig, daß die Aufzeichnungen aufgrund unkorrigierbarer Bewegungsartefakte nicht sinnvoll auswertbar waren. In sieben weiteren Fällen wurde die Auswertung der kardiovaskulären Daten durch gravierende Herzrhythmusstörungen beeinträchtigt. Zehn Vpn gaben während des DLT bzw. GAT bei mindestens einem Trial eine falsche oder keine Antwort. Zwei Teilnehmer ließen bei der Bearbeitung der Reaktionsratings einige Items aus. Die Daten von sieben Pbn wurden ausgeschlossen, weil sie in den Gedächtnistests mehr als 4 von 12 kritischen Details der Scheinverbrechen nicht erinnern bzw. wiedererkennen konnten. Bei drei Teilnehmern lag die NSR-Anzahl genau auf dem Median von 29, so daß keine eindeutige Zuordnung der Variable Elektrodermale Labilität (Stabil vs. Labil) möglich war. Jeweils ein Datensatz aus den Bedingungskombinationen Ring-DLT-Labil und Ring-GAT-Labil sowie zwei Datensätze aus der Bedingungskombination Kette-DLT-Stabil wurden per Zufallsziehung eliminiert, um für alle Versuchsgruppen die gleiche Zellbesetzung von $n = 10$ zu erhalten (s. o.).

Die **verbleibende Stichprobe** bestand aus 80 Vpn, deren Alter zwischen 20 und 69 Jahren lag und im Durchschnitt ca. 28 Jahre betrug ($SD = 8.93$, Median = 26 Jahre). Unter den Pbn waren 49 Studenten aus unterschiedlichen Fachbereichen (durchschnittliche Semesterzahl: $M = 7.8$, $SD = 4.44$) und 31 nichtstudentische Vpn. Dabei handelte es sich weder um Psychologen noch Psychologiestudenten. In der Stichprobe waren nach Angaben der Vpn 73 Rechts- und sieben Linkshänder. Neunundvierzig Vpn berichteten eine Störung der Sehschärfe. Davon trugen 38 während der Untersuchung eine Brille oder Kontaktlinsen. Keiner der Teilnehmer äußerte Probleme beim Lesen der Instruktionen bzw. schriftlich dargebotenen Stimuli. Dreiundsiebzig Vpn sprachen Deutsch als Muttersprache. Die Deutschkenntnisse der anderen Pbn waren ebenfalls so gut, daß sich keine Verständnisschwierigkeiten zeigten. Fünfunddreißig Vpn gaben an, schon früher an psychologischen Experimenten teilgenommen zu haben, wobei es sich in keinem Fall um ein Experiment zum Thema „Lügendetektion“ handelte. Keine der Vpn besaß zum Zeitpunkt der Untersuchung genauere Informationen zum vorliegenden Experiment, außer den wenigen Angaben, die im Rahmen der Anwerbung publik gemacht worden waren.

4.3 Versuchsaufbau, Apparaturen und Stimuli

Die psychophysiologische Aussagebeurteilung wurde in einer elektrisch und akustisch abgeschirmten **Meßkabine** durchgeführt (Industrial Accustics Company, Typ 403-A). Im **Kabineninneren** befand sich ein Stuhl mit gepolsterter Sitzfläche sowie gepolsterten Rücken- und Armlehnen. Der Stuhl war so vor einem in der Wand der Meßkabine

befindlichen Fenster plaziert, daß eine bequem darauf sitzende Person bei gerade nach vorne gerichtetem Blick genau auf das Fenster sah. Dabei betrug der Abstand zwischen den Augen und der Glasscheibe ca. 115 bis 125 cm. Etwa 11 cm hinter der Glasscheibe befand sich ein Computermonitor. Über diesen Bildschirm erfolgte die schriftliche Präsentation der Fragen bzw. Items. In den beiden Raumecken hinter dem Stuhl standen zwei Lautsprecherboxen auf dem Boden. Die Boxen dienten zur auditiven Darbietung der Stimuli (s. u.).

Rechts neben dem Stuhl stand ein kleiner Tisch. Darauf war ein schwenkbares **Lesepult** befestigt, das zur Vergabe der schriftlichen Instruktionen benutzt wurde und mit variablem Abstand vor einer auf dem Stuhl sitzenden Person arretiert werden konnte. Rechts hinter dem Stuhl befand sich auf einem ca. 140 cm hohen Stativ eine **Videokamera** (Panasonic WV-3600E), deren Objektiv auf den Stuhl gerichtet war. Mittels dieser Kamera konnte die Vp während der Messung beobachtet werden. An der Wand links neben dem Stuhl war die Nebenstation einer **Wechselsprechanlage** befestigt (Hartig & Helling WRA 23).

Die **klimatischen Bedingungen** in der Kabine wurden jeweils zu Beginn und Ende der physiologischen Messungen anhand einer TempTec Thermo-Hygro-Clock (Mdl. 241, Conrad Electronic GmbH) bestimmt. Die Temperatur schwankte im Bereich von 21.3 bis 24.3 Grad Celsius. Die relative Luftfeuchtigkeit lag zwischen 36% und 68%. Das Kabineninnere wurde von zwei an den Wänden installierten Lampen beleuchtet. Die **Beleuchtung** war leicht gedämpft. Sie reichte aber zum problemlosen Lesen der Instruktionen und zum Betrieb der Überwachungskamera aus.

Die Aufzeichnung der Hautleitfähigkeit und die Erfassung der übrigen physiologischen Variablen (EKG, Atmung, Plethysmogramm sowie zusätzliche Kontrolle der Hauttemperatur und der Antworten der Pbn) erfolgten über zwei unterschiedliche Systeme. Der physiologische Datenrekorder **Varioport** (Kölner Vitaport-System, BECKER MEDITEC) sowie dessen Meßaufnehmer und Verstärker für das EKG, für den Atemgurt und für den Temperatursensor befanden sich in der Kabine. Das Varioport war unterhalb der linken Armlehne des Stuhls positioniert. An mehreren seiner Meßkanäle waren zusätzliche jeweils im Hause entwickelte Geräte angeschlossen. Dabei handelte es sich unter anderem um einen Verstärker mit eingebautem Mikrophon (Ulmann, Universität Mainz), das an der linken Wand neben dem Stuhl installiert war und zur Aufzeichnung der Antwortzeitpunkte diente. Dazu kamen noch ein regelbarer Plethysmographie-Verstärker (Juris, Universität Mainz) und ein Digital-Analog-Wandler (MARK-I, Ulmann, Universität Mainz), über den Markierungen in die Aufzeichnungen des Vario-

ports gesetzt wurden (s. u.). Der Plethysmographie-Verstärker und der Wandler waren außerhalb der Meßkabine plaziert.

Dort befand sich auch der größte Teil der Apparate für die **Versuchssteuerung**. Diese lagen somit außer Sicht- und Hörweite der Vpn. Die **Steuerung des Ablaufs** sowie die **Aufzeichnung der Hautleitfähigkeit** erfolgte über einen Computer (American Megatrend 486DX 33 MHz, Betriebssystem: MS-DOS 6.20), auf dem das im Hause entwickelte Programm EDA-MESS (Münch, Universität Mainz) implementiert war. An den Rechner waren auch der EDA-Verstärker (CEDA-12, Juris, Universität Mainz) und ein Bildschirm angeschlossen (EIZO Flexscan 9060S, 14-inch Colour Data Display), der eine On-line-Kontrolle der Hautleitfähigkeitsmessung ermöglichte. Dieser Computer steuerte über eine serielle Schnittstelle einen zweiten Rechner an (Award Pentium-S 100 MHz, Betriebssystem: MS-Windows 95), auf dem die Software zur Präsentation der Fragen und Items installiert war (Programm: SPEAK, Münch, Universität Mainz). Durch diese Verbindung signalisierte der Steuerungscomputer jeweils den Beginn und das Ende der Stimulusdarbietung. Außerdem markierte er deren Zeitpunkte in den Aufzeichnungen der Hautleitfähigkeit und sendete über die parallele Schnittstelle und den o. g. Digital-Analog-Wandler Signale in einen der Meßkanäle des Varioports, so daß die Stimulusdarbietungen auch in den Varioport-Aufzeichnungen registriert werden konnten. Die **Steuerung des Varioports** erfolgte mittels des Softwarepakets VitaGraph (Version 4.45), das auf einem dritten Rechner installiert war (Award Pentium-MMX 233 MHz, Betriebssystem: MS-Windows 95) und über einen Bildschirm (Iiyama Vision Master 17, Modell-Nr. MF-8617T) eine On-line-Überwachung der Aufzeichnungen gestattete. Die vom Varioport gemessenen Daten wurden zunächst auf der internen 20 MB-Memorycard gespeichert und erst nach der Messung per VitaGraph ausgelesen und entsprechend den Hautleitfähigkeitsdaten zur weiteren Verarbeitung auf Festplatte gespeichert.

Die **Präsentation der Fragen und Items** in der Kabine erfolgte sowohl visuell als auch auditiv. Die Stimuli wurden auf dem Bildschirm hinter dem Kabinenfenster im Fett-Format der Schriftart „Arial Black“ (Schriftgrad 22, schwarz auf hellgrauem Hintergrund) eingeblendet und parallel dazu per Soundkarte (Creative Labs Sound Blaster 16), Stereoanlage (Technics Stereo Integrated Amplifier SU-Z400 und Stereo Graphic Equalizer SH-8065) sowie über die o. g. Lautsprecherboxen (Jamo 7084) in normaler Zimmerlautstärke eingespielt. Die Sätze waren vor dem Experiment von einer männlichen Stimme vorgelesen und mittels des Programms Goldwave 4.01 aufgenommen, digital überarbeitet sowie mit einer Samplingrate von 16 kHz (16-bit, mono) abgespeichert worden. Die visuelle und auditive Stimulusdarbietung begann jeweils synchron. Die Dauer der schriftlichen Präsentation schwanke nach einem für die Vpn unvorher-

sagbaren Muster in Halbsekundenschritten zwischen 8 und 10 Sekunden (vgl. Abbildung 1; DLT: *M* = 8.93 Sekunden, *SD* = 0.70; GAT: *M* = 8.96 Sekunden, *SD* = 0.72). Die zeitliche Länge der auditiven Reize variierte ebenfalls, sie war jedoch stets kürzer als die visuelle Darbietung. Die Ausblendung der Fragen bzw. Items auf dem Bildschirm fungierte als imperativer Reiz für die Antwortgabe. Die Gesamtlänge eines Trials betrug stets 30 Sekunden. Zwischen den Trials benötigte EDA-MESS ca. 1 bis 2 Sekunden zur Abspeicherung der Hautleitfähigkeitsdaten auf Festplatte, so daß das Zeitintervall zwischen zwei aufeinanderfolgenden Einblendungen 31 bis 32 Sekunden betrug.

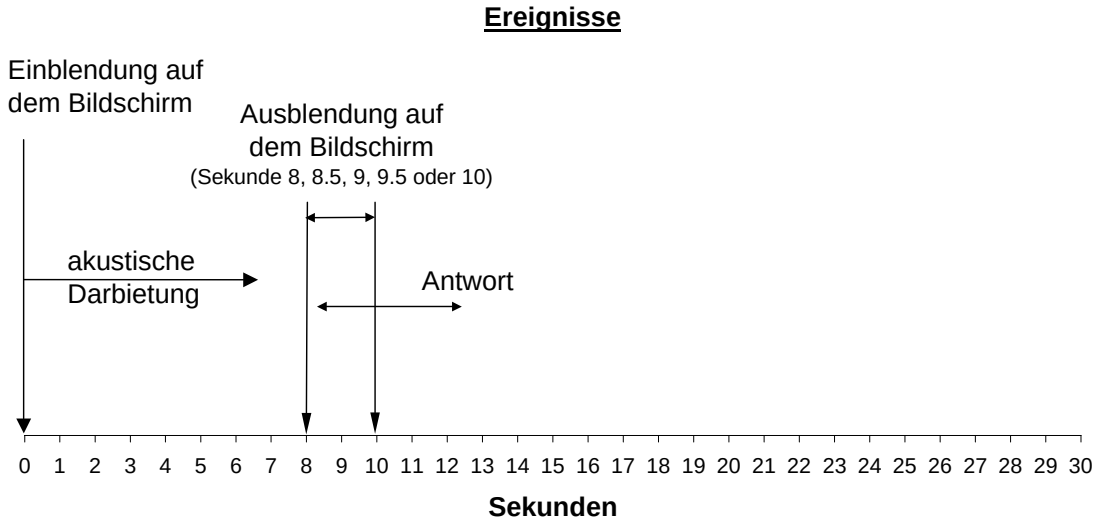


Abbildung 1. Zeitlicher Ablauf der Reizdarbietung.

Vom Anfang bis zum Ende eines Trials wurde die Hautleitfähigkeit kontinuierlich gemessen und aufgezeichnet. Die **Datenerfassung** des Varioports lief durchgehend, d. h. auch zwischen den Trials in den Speicherpausen von EDA-MESS. Die Ein- und Ausblendung der Fragen bzw. Items sowie die Aufzeichnung der physiologischen Daten erfolgten vollautomatisch computergesteuert. Die Aufgabe des Untersuchers bestand lediglich darin, die entsprechenden Programme zu starten und die Reaktionen der Vp zu überwachen. Während des Programmablaufs fand keine Interaktion zwischen Vl und Vp statt. Die visuelle und auditive Stimulusdarbietung konnte durch einen zum Monitor der Reizpräsentation synchron geschalteten Bildschirm (EIZO Flexscan 5500, s. o.) und durch die Hauptstation der Wechselsprechanlage außerhalb der Kabine vom Vl kontrolliert werden. Die Wechselsprechanlage ermöglichte zusätzlich eine Überprüfung der Antworten der Pbn. Ebenfalls im Blickfeld des Vl stand ein Bosch Picture Monitor (M 50 BB), an den die Überwachungskamera in der Kabine angeschlossen war.

4.4 Versuchsablauf

4.4.1 Rahmenbedingungen

Das **Experiment** wurde in der Zeit vom 6. Mai 1999 bis zum 30. August 1999 am Psychologischen Institut der Johannes Gutenberg-Universität Mainz in mehreren Räumen (02-133, 02-521, 02-524, 02-525 und 02-628) der 2. Etage des Gebäudes Staudingerweg 9 in Form von Einzeluntersuchungen durchgeführt. Die Dauer einer Untersuchung mit anschließender Aufklärung des jeweiligen Pb betrug ca. 90 bis 120 Minuten. Keine der Vpn begann ihr Experiment vor 9.00 Uhr bzw. nach 19.00 Uhr.

Am Versuch waren jeweils **zwei männliche Versuchsleiter** beteiligt (im folgenden als Vl1 und Vl2 bezeichnet). Die Rolle des Vl1 teilten sich tageweise zwei studentische Hilfskräfte (Psychologiestudenten im Hauptdiplomsabschnitt), während der Autor der vorliegenden Arbeit stets als Vl2 agierte.

Der Ablauf eines experimentellen Durchgangs läßt sich grob in **drei aufeinanderfolgende Abschnitte** untergliedern: (1) Vorbereitung und Durchführung des Scheinverbrechens, (2) Psychophysiologische Aussagebeurteilung und (3) Messung der erlebnisdeskriptiven Variablen, Gedächtnistests, Nachbefragung und Aufklärung.

4.4.2 Vorbereitung und Durchführung des Scheinverbrechens

Bei Ankunft im Psychologischen Institut wurde der jeweilige Pb vom Vl1 außerhalb des Laborbereichs begrüßt und in ein Büro (02-133) im Ostflügel des Gebäudes geführt (für eine Skizze der Lage der Räumlichkeiten vgl. auch Anhang C.2: Instruktion 1a). Dort erhielt die Vp zunächst eine kurze schriftliche **Information** zur Studie (Anhang C.1). Darin wurde darauf hingewiesen, daß es sich um ein Experiment zur „Lügendetektion“ handelte, und eine Belohnung von 20 DM in Aussicht gestellt, falls der Teilnehmer den „Lügendetektortest“ als „unschuldig“ bestünde. Des weiteren machte man den Pb auf die Freiwilligkeit der Teilnahme, die Einhaltung von Datenschutzrichtlinien und die spätere vollständige Aufklärung über die Untersuchung aufmerksam und verpflichtete ihn zur Verschwiegenheit.

Nachdem der Pb durch Unterschreiben einer **Einverständniserklärung** (Anhang C.1) die genannten Bedingungen akzeptiert hatte, wurde er der Tatbedingung (Ring vs. Kette) zugewiesen. Diese Aufteilung erfolgte bei den ersten 90 Vpn anhand einer Zufallszahlenliste. Später wurde dabei zusätzlich die Zielsetzung homogener Zellbeset-

zungen in den Versuchsgruppen berücksichtigt. Anschließend erhielt der Pb die jeweilige **Instruktion für das Scheinverbrechen** (Anhang C.2 bzw. C.3: Instruktion 1a bzw. 1b). Darin wurde ihm mitgeteilt, daß es im Experiment sowohl „schuldige" als auch „unschuldige" Teilnehmer gäbe und er per Zufall der Gruppe der Täter zugeteilt worden wäre. Als solcher sollte er einen simulierten Schmuckdiebstahl begehen.

Im folgenden wird der schematische **Ablauf der Scheinverbrechen** exemplarisch am **Ringdiebstahl** veranschaulicht (vgl. dazu Anhang C.2). Es sei vermerkt, daß man bei der Gestaltung der Szenarios besonders darauf geachtet hatte, hinreichend viele Einzelheiten einzuarbeiten, die sich als Items für den GAT eigneten. Alle kritischen Details wurden in den Instruktionen 1a bzw. 1b benannt und hervorgehoben.

Der Pb erhielt die Anweisung, den im Nordflügel der gleichen Etage gelegenen **Raum 02-525** aufzusuchen. Als Orientierungshilfe war eine Skizze des Weges in der Instruktion abgebildet. Der betreffende Raum war als Arbeitszimmer für Diplomanden und Tutoren deklariert. Auf dem Weg dorthin mußte eine Glastür mit dem Schild „Abteilung Allgemeine Experimentelle Psychologie – Laborbereich – Kein Durchgang" passiert werden. Der Pb sollte dieses Verbot ignorieren, sich jedoch für den Fall, daß ihn Mitarbeiter des Instituts auf den Grund seiner Anwesenheit hin ansprechen würden, eine plausible Rechtfertigung überlegen. Dabei durfte er auf keinen Fall angeben, er sei an einem psychologischen Experiment beteiligt. Der Raum 02-525 war zuvor so präpariert worden, daß er einem Institutsbüro ähnelte. Um den Eindruck zu vermitteln, das Arbeitszimmer werde tatsächlich genutzt und sei nur kurzfristig verlassen worden, befanden sich neben diversen Büroutensilien, Fachzeitschriften und Büchern auch Unterlagen auf den Tischen, an denen offenbar gearbeitet wurde. Die Lampe und der Computer auf dem Schreibtisch waren eingeschaltet. Daneben standen eine Tasse Kaffee und eine halbvolle Flasche Cola. Ferner hing eine Jacke am Kleiderständer, und an der Eingangstür klebte ein Zettel mit dem handschriftlichen Vermerk „Bin gleich zurück! K. Roggendorf".

Der im Raum befindliche Schreibtisch hatte mehrere Schubladen (vgl. Anhang D: Abbildung 39). Zwei dieser Schubladen waren mit Nummern versehen. In einer der Schubladen (**Nr. 3**) lag neben einem **Hefter** ein durchsichtiger **Plastikkasten** und darin eine **Kette** mit einem **gelben Stein** als Anhänger. An der Kette war ein Papieretikett mit dem Buchstaben „**A**" befestigt (siehe Anhang D: Abbildung 40). Der Pb sollte diese Schublade öffnen, sich deren Inhalt gut einprägen und sie danach wieder schließen, ohne einen der Gegenstände zu berühren. Daraufhin sollte er eine zweite Schublade (**Nr. 6**) öffnen. In dieser Schublade befand sich neben einem **Locher** eine **Glasschale** und darin ein **Ring** mit einem **blauen Stein**. Am Ring war ein Etikett mit dem Buch-

staben „**E**“ befestigt. Der Pb sollte die Glasschale öffnen, den Ring an sich nehmen und ihn in der Hosentasche verstauen. Außerdem erhielt er die Anweisung, nach dem Schließen der Schublade den Raum zu verlassen und zum Büro 02-133 zurückzukehren.

Für jene Vpn, die die Halskette entwenden mußten (**Tatbedingung Kette**), lief das Scheinverbrechen dementsprechend parallel ab. Lediglich die Reihenfolge beim Öffnen der Schubladen war invers (zuerst Nr. 6 und dann Nr. 3) und die Pbn wurden instruiert, den Plastikkasten zu öffnen und daraus die Kette zu entnehmen, wohingegen sie den Inhalt der Schublade Nr. 6 (Ring) unangetastet lassen sollten (vgl. Anhang C.3: Instruktion 1b).

Nach dem Lesen dieser Instruktion wurde der Pb aufgefordert, sich den **Ablauf gut einzuprägen** und die wichtigsten Punkte wiederzugeben. Falls dabei Fehler auftraten, wurde er vom Vl1 korrigiert und gebeten, die Instruktion nochmals zu lesen und sich die Details zu merken. Erst nach einer fehlerfreien Wiederholung folgten die weiteren Schritte. Der Vl1 händigte der Vp einen Zettel mit der Nummer des Scheinverbrechen-Raumes aus und beantwortete gegebenenfalls offene Fragen.

Danach beging der Pb den jeweiligen **Schmuckdiebstahl**. Der Vl2 befand sich in einem Nebenraum (02-521) des Zimmers 02-525 und beobachtete die dortigen Geschehnisse über eine verborgene Computerkamera und ein Laptop. Nachdem die Vp zum Büro 02-133 zurückgekehrt war, kontrollierte der Vl1, ob sie das entsprechende Schmuckstück in der Hosentasche deponiert hatte. Anschließend erhielt sie die Instruktion 2 (Anhang C.4). Darin wurde sie über den bevorstehenden „Lügendetektortest“ in Kenntnis gesetzt und aufgefordert, gegenüber dem Vl2 ihre Unschuld hinsichtlich des Scheinverbrechens zu beteuern und unter allen Umständen den Tatvorwurf abzustreiten. Darüber hinaus wurde der Pb erneut darauf aufmerksam gemacht, daß er nur dann die Belohnung von 20 DM erhalten könnte, wenn man ihn als „unschuldig“ einstufen würde.

Danach führte der Vl1 die Vp wieder in den **Laborbereich**, wobei sie einen anderen Korridor als beim Scheinverbrechen benutzten. Im Raum 02-524 wartete bereits der Vl2, um die psychophysiologische Aussagebeurteilung durchzuführen.

4.4.3 Psychophysiologische Aussagebeurteilung

Der **Laborraum 02-524** war unterteilt in eine Kabine, in der sich die Vp während der Untersuchung aufhielt, und in einen äußeren Bereich, in dem sich die Apparaturen für die Versuchssteuerung und der Vl2 während den Messungen befanden. Eine wichtige

Zielsetzung der vorliegenden Studie war die erhöhte Standardisierung. Der Ablauf der psychophysiologischen Aussagebeurteilung sollte für alle Pbn äquivalent gestaltet sein, d. h., der DLT und der GAT liefen bei allen Vpn der entsprechenden Testbedingung konstant ab. Darüber hinaus wurde ein großer Wert auf schriftliche Instruktionen und die vollautomatische Versuchssteuerung gelegt. Die Vpn wußten nicht, daß der Vl2 die Tatbedingung kannte, und der Vl2 verhielt sich diesbezüglich unauffällig. Die Tatsache, daß der Untersucher Kenntnis von dem jeweils durchgeführten Scheinverbrechen hatte, war in diesem Experiment insofern unproblematisch, als man im Gegensatz zu den üblichen Analogstudien keine direkte individualdiagnostische Zielsetzung verfolgte, sondern die Reaktionsunterschiede quantitativ untersuchen wollte. Aufgrund des standardisierten Ablaufs und der objektiven Auswertung hatte der Vl2 trotz seines Wissens keine Möglichkeit, darauf Einfluß zu nehmen, so daß diesbezüglich keine gravierenden Effekte zu erwarten waren.

Der Vl1 übergab die Vp an den Vl2 und verließ den Raum. Der Vl2 stellte sich vor. Dann bat er die Vp, zunächst außerhalb der Meßkabine Platz zu nehmen, und händigte ihr die Instruktion 3 (Anhang C.5) aus. Darin wurde kurz das **Prinzip der „Lügendetektion“ erklärt**. Um die subjektiv eingeschätzte Treffsicherheit der Methode zu erhöhen, gab man dem Pb die fingierte Information, daß jede Lüge zu körperlichen Veränderungen führen würde (vgl. Untersuchung zum DLT von Honts & Raskin, 1988, S. 58) und daß wissenschaftliche Studien zur „Lügendetektion“ hohe Trefferquoten erbracht hätten. Zusätzlich zum finanziellen Anreiz sollte die Täuschungsmotivation durch die Instruktion erhöht werden, daß nur besonders intelligente, emotional kontrollierte Personen in der Lage seien, den „Lügendetektor“ zu überlisten (vgl. auch Elaad & Ben-Shakhar, 1997, S. 589; Gustafson & Orne, 1963, S. 409). Darüber hinaus wurde der Ablauf der Untersuchung vom Anlegen der Meßfühler bis zum „Lügendetektortest“ skizziert. Die Zeit, in der sich die Vp die schriftliche Anweisung vergegenwärtigte, nutzte der Vl2 zum Kalibrieren der EDA-Meßapparatur und zur Vorbereitung der Elektroden.

Nach dem Lesen der Instruktion forderte man den Pb auf, sich am Waschbecken im Laborraum mit handwarmem Wasser und ohne Seife die Hände zu waschen. Es folgte das **Anbringen der EKG- und EDA-Elektroden** (vgl. die Abschnitte 4.5.1 und 4.5.2). Die Haut an den Ableitorten wurde zuvor mit Äthylalkohol (70%) gereinigt.

Im Anschluß daran führte der Vl2 die Vp in die **Kabine** und bat sie, auf dem Stuhl eine möglichst bequeme Sitzhaltung einzunehmen. Zu diesem Zweck bot man ihr auch die Möglichkeit, die gepolsterten Armlehnen zu verstellen und die Füße auf einer Schaumstoffunterlage abzulegen. Der Monitor, auf dem später die schriftliche Reizpräsentation

erfolgte, war zu diesem Zeitpunkt bereits eingeschaltet. Er zeigte jedoch über die ganze Fläche nur den hellgrauen Hintergrund. Der sitzenden Vp wurde der Atemgurt angelegt (vgl. Abschnitt 4.5.4). Nachdem der Vl2 die EDA- und EKG-Elektroden an die Meßvorrichtung angeschlossen hatte, brachte er den Photoplethysmographie- sowie den Hauttemperatursensor an (vgl. die Abschnitte 4.5.3 und 4.5.5) und schaltete das Mikrophon ein. Hierauf verließ der Vl2 die Meßkabine, um zu überprüfen, ob die Signale störungsfrei abgeleitet und übertragen wurden. War dies nicht der Fall, wurden die entsprechenden Mängel behoben. Anschließend betrat der Vl2 erneut die Meßkabine und erfaßte dort einige soziodemographische Daten der Vp (Alter, Studienfach bzw. Beruf, dominante Hand etc.) sowie die Temperatur und Luftfeuchtigkeit in der Kabine. Hierzu wurde ein vorgefertigter Befragungsbogen verwendet (siehe Anhang C.18), auf dem der Vl2 alle Informationen schriftlich festhielt.

Danach erhielt die Vp auf dem Lesepult die schriftliche Instruktion 4 (Anhang C.6). Darin wurde sie aufgefordert, während der darauffolgenden **Ruhemessung** bequem und entspannt zu sitzen, die Augen geöffnet zu halten und sich möglichst wenig zu bewegen. Außerdem machte man die Vp auf die Überwachungskamera in der Kabine aufmerksam. Währenddessen bereitete der Vl2 die EDA-Messung vor und startete die Varioport-Aufzeichnung. Nach dem Lesen der Instruktion kündigte der Vl2 den Beginn der Ruhemessung an und sagte, daß er nach sechs Minuten wieder in die Kabine kommen werde, um den weiteren Ablauf der Untersuchung zu erklären. Sobald der Vl2 die Kabinentür hinter sich geschlossen hatte, startete er mit dem Programm EDA-MESS die Ruhemessung. Diese dauerte insgesamt sechs Minuten. In dieser Zeit wurde die Hautleitfähigkeit (ebenso wie die restlichen physiologischen Daten) kontinuierlich gemessen und aufgezeichnet. Die erste Minute der Ruhemessung fungierte als Beruhigungsphase, während der die Vp sich an die Versuchssituation gewöhnen und an die physikalischen Verhältnisse in der Meßkabine adaptieren konnte. Die letzten fünf Minuten der Ruhemessung dienten zur Erfassung der Anzahl elektrodermaler Spontanfluktuationen (NSRs). Während der Ruhemessung wurde außerdem der Plethysmographie-Verstärker eingestellt (vgl. Abschnitt 4.5.3).

Im Anschluß an die Ruhemessung gestattete das Programm SPONDOS (Münch & Ulmann, Universität Mainz) eine automatische **Grobauszählung der NSRs**. Auf deren Basis war – unter Berücksichtigung des aktuellen Medians – eine vorläufige Grobklassifikation der Vpn in elektrodermal Stabile und Labile möglich. Später wurde die EDA-Ruhemessung (off-line) nochmals genauer ausgewertet, die NSR-Anzahl korrigiert und der Median kontinuierlich aktualisiert. Die ersten 30 Vpn wurden anhand einer Zufallszahlenliste den Testbedingungen (DLT vs. GAT) zugewiesen. Bei den restlichen Pbn

erfolgte die Zuweisung pseudorandomisiert unter Beachtung der EL und der Zielsetzung einheitlicher Zellbesetzungen.

Nach Beendigung der Ruhemessung betrat der Vl2 wieder die Meßkabine und legte der Vp die Instruktion 5 (vgl. Anhang C.7) für den **Stimulationstest** vor. Es handelte sich dabei um einen Zahlentest, der in Anlehnung an die DLT-Untersuchungen von Honts und Raskin (1988, S. 58) sowie Horowitz et al. (1997, S. 110) gestaltet war (vgl. „Known-Solution Stimulation Test", Matte, 1996, S. 311). Entsprechend deren Empfehlung fand der Stimulationstest vor der Besprechung der eigentlichen Testfragen statt. Der Pb sollte eine Zahl zwischen 11 und 16 auswählen und sie dem Untersucher mitteilen. Anschließend wurde ähnlich wie beim GAT eine Multiple-Choice-Frage nach der Zahl gestellt (siehe Tabelle 9). Die Darbietung der Frage und der Items erfolgte computergesteuert per Monitor und Lautsprecher. Der Pb wurde aufgefordert, die Items erst nach deren Ausblendung auf dem Bildschirm zu verneinen. Folglich mußte er auf das Item der gewählten Zahl lügen, sollte dabei aber möglichst glaubwürdig erscheinen. Diese Prozedur wurde damit erklärt, daß man die typischen Reaktionen beim Lügen bestimmen wollte, um damit die Meßinstrumente einzustellen und zu überprüfen, ob sich der Pb für eine Untersuchung mit dem „Lügendetektor" eignete. Unabhängig von seinen körperlichen Reaktionen gab der Vl2 dem Pb nach dem Stimulationstest die fingierte Rückmeldung, daß er auf das Item nach der gezogenen Zahl am stärksten reagiert habe und er somit für die „Lügendetektion" geeignet sei (vgl. auch Abschnitt 2.5).

Tabelle 9. Frage und Items des Stimulationstests

Trial	Frage und Items
	Welche Zahl haben Sie gewählt? War es
1.	die 11?
2.	die 12?
3.	die 13?
4.	die 14?
5.	die 15?
6.	die 16?

Anmerkung. Alle Items wurden verneint.

In Abhängigkeit davon, welcher **Testbedingung** sie zugeordnet war (DLT vs. GAT), erhielt die Vp anschließend die Instruktion 6a bzw. 6b (Anhang C.8 bzw. C.9). Darin hieß es, das Psychologische Institut der Universität Mainz habe ein neuartiges Programm zur „Lügendetektion" entwickelt und der Zweck des Experiments sei es, dessen Treffsicherheit zu überprüfen. Außerdem wurden die Befragungstechniken jeweils erklärt.

In der **Instruktion zum DLT** wurden die fünf Fragentypen nacheinander vorgestellt und erläutert. Zu Beginn ging man auf die irrelevanten Fragen näher ein. Und dann wurden entsprechend der von Honts und Raskin (1988, S. 58) für das Vortest-Interview vorgeschlagenen Reihenfolge die relevanten Fragen vor den Kontrollfragen erörtert. Man instruierte den Pb, die irrelevanten Fragen nach dem momentanen Aufenthaltsort aufrichtig zu bejahen, den Tatvorwurf der relevanten Fragen abzustreiten und die Lügen- bzw. Wahrheit-Kontrollfragen wahrheitswidrig bzw. wahrheitsgemäß zu verneinen. Die Anweisungen zu den Kontrollfragen (dort als Vergleichsfragen bezeichnet) waren an die üblichen Instruktionen des DLT angelehnt (vgl. Abschnitt 2.5). Bei der Verneinung sollte sich der Pb den Wahrheitsgehalt seiner Antwort vergegenwärtigen, an konkrete Situationen der angesprochenen Verfehlungen denken und auf seine Emotionen achten. Darüber hinaus erhielt er die Information, daß man seine Reaktionen auf die tatbezogenen Fragen mit denen auf die Kontrollfragen vergleichen und damit seine Glaubwürdigkeit bestimmen könne. Falls sich keine angemessenen Reaktionen auf die Vergleichsfragen zeigen würden, bestünde ein hohes Risiko, daß man ihn als „schuldig“ einstufen würde. Ferner machte man den Pb darauf aufmerksam, daß man die insgesamt 15 Fragen in einem zweiten Durchgang in veränderter Reihenfolge wiederholt darbieten werde.

Die **Instruktion zum GAT** benannte die Themengebiete, auf die sich die Multiple-Choice-Fragen bezogen (gestohlene Schmuckstücke, Aufbewahrungsgefäße, Nummern der Schubladen, Büroutensilien in den Schubladen, Farben der Schmucksteine, Buchstaben auf den Etiketten) und kündigte an, daß es zu jeder Frage sechs Alternativen gab. Entsprechend der üblichen Vorgehensweise des TWT bzw. GAT waren die Items selbst nicht aufgelistet.

Beide Instruktionen (sowohl zum DLT als auch zum GAT) verwiesen nochmals auf die in Aussicht gestellte Belohnung. Der Pb wurde erneut daran erinnert, die Fragen bzw. Items erst nach Ausblendung auf dem Bildschirm zu beantworten und ansonsten nichts zu sagen. Er sollte versuchen, stets glaubwürdig zu erscheinen und sich möglichst wenig zu bewegen. Außerdem wurde behauptet, daß eventuelle Versuche, die Messung zu stören, entdeckbar wären und zum Verlust der Belohnung führen würden. Zum Abschluß kündigten die Instruktionen jeweils einen Übungsdurchgang an.

Nachdem der Pb die betreffende Anweisung gelesen hatte und eventuell bestehende Fragen geklärt waren, verließ der Vl2 die Meßkabine, schloß die Tür und startete den entsprechenden **Probedurchgang**. Beim DLT wurde für jeden der fünf Fragentypen ein Beispiel präsentiert (vgl. Tabelle 10). Beim GAT bestand die Übungsphase aus einer Multiple-Choice-Frage, die sich auf die Nummer des Scheinverbrechen-Büros bezog

(vgl. Tabelle 11). Keine dieser Fragen war Bestandteil der eigentlichen Tests. Die Daten der Probedurchgänge wurden ebenso wie die des Stimulationstests nicht ausgewertet.

Tabelle 10. Fragen und Antworten des Probedurchgangs zum DLT sowie Fragentypen in Abhängigkeit von der Tatbedingung

Trial	Frage	Antwort
1.	Sitzen Sie auf einem Stuhl?	J
2.	Haben Sie die Schublade Nr. 6 geöffnet?	N
3.	Sind Sie immer aufrichtig?	N
4.	Haben Sie die Schublade Nr. 3 geöffnet?	N
5.	Sind Sie manchmal unaufrichtig?	N

Anmerkungen. Antwort: J = Ja, N = Nein.

Tabelle 11. Frage und Items des Probedurchgangs zum GAT

Trial	Frage und Items
	In welchem Raum haben Sie den Schmuckdiebstahl begangen? War es
1.	Raum 02-522?
2.	Raum 02-528?
3.	Raum 02-525?
4.	Raum 02-523?
5.	Raum 02-527?
6.	Raum 02-521?

Anmerkungen. Alle Items wurden verneint. Trial 3 beinhaltete das relevante Item.

Nach dem Probedurchgang betrat der Vl2 erneut die Kabine. Falls Probleme aufgetreten waren (z. B. hinsichtlich Antwortart und -zeitpunkt), wurden Teile der Instruktion mündlich wiederholt bzw. nochmals schriftlich dargeboten und diesbezügliche Unklarheiten beseitigt. Vor Verlassen der Kabine sagte der Untersucher, daß gleich der eigentliche Test beginne. Er schloß die Tür hinter sich und startete je nach Testbedingung den **Versuchsdurchgang für den DLT bzw. GAT**.

Der **DLT** beinhaltete 15 unterschiedliche Fragen. Wie bei Kontrollfragentests üblich, bot man die Fragen in veränderter Reihenfolge mehrfach dar. Im vorliegenden Experiment wurden sie in zwei Durchgängen wiederholt. Daraus resultierten insgesamt 30 Trials (vgl. Tabelle 12). Die Abfolge war für alle Pbn der Testbedingung DLT konstant. Die Fragensequenz ließ sich in sechs Blöcke unterteilen. Pro Block gab es eine irrelevante Frage und jeweils ein zusammengehöriges relevantes Fragenpaar, die sich auf den Ring- bzw. Kettendiebstahl bezogen, sowie ein (Lügen- vs. Wahrheit-) Kontrollfragenpaar. Der Block begann stets mit einer irrelevanten Frage. Danach wechselten sich relevante und Kontrollfragen ab, wobei die Reihenfolge der entsprechenden Fragentypen zwischen den Blöcken variierte. Bei der Wiederholung der Fragen im zweiten Durch-

gang, der sich ohne Unterbrechung an den ersten anschloß, tauschten die zusammengehörigen relevanten Fragenpaarlinge (Ring vs. Kette) ihre Position untereinander. Sie behielten aber die Reihenfolge in bezug auf die Blöcke bei, d. h., in beiden Durchgängen wurden zunächst die relevanten Fragen nach dem gestohlenen Schmuckstück, dann nach dem Aufbewahrungsgefäß (Glasschale vs. Plastikkasten) und abschließend nach dem bei sich getragenen Schmuckstück gestellt. Im Gegensatz dazu wechselten die irrelevanten Fragen und die Kontrollfragenpaare die Position zwischen den Blöcken. Diese Vorgehensweise orientierte sich an der gängigen Methodik von Kontrollfragentests, die Position der irrelevanten und Kontrollfragen über die Durchgänge hinweg zu rotieren, während die Position der relevanten Fragen konstant bleibt (vgl. Horowitz et al., 1997, S. 111; Matte, 1996, S. 368).

Der **GAT** bestand aus sechs Multiple-Choice-Fragen à sechs Items, d. h. insgesamt 36 Trials (vgl. Tabelle 13). Die Einleitungsfrage und das erste irrelevante Item waren jeweils in einem Trial zusammengefaßt. Die Positionen der relevanten Items variierten pseudozufällig von Frage zu Frage. Sie nahmen aber nie die erste Position in einem Multiple-Choice-Block ein. Die inhaltlich parallelisierten relevanten Items, die sich auf den Ring- oder Kettendiebstahl bezogen, standen jeweils gleich häufig vor- und nacheinander. Die Anzahl der zwischengeschalteten irrelevanten Items schwanke zwischen null und zwei. Die Sequenz wurde für alle Vpn der Testbedingung GAT konstant gehalten. Bei der Auswahl der irrelevanten Alternativen hatte man darauf geachtet, daß sie von den relevanten Details gut zu unterscheiden waren und nicht im Kontext der Scheinverbrechen auftraten (vgl. Lykken, 1998, S. 288).

Während des Versuchsablaufs überwachte der Vl2 die Reaktionen der Vp. Die schriftliche Darbietung und die Ausblendung der Stimuli wurden auf einem Monitor außerhalb der Meßkabine kontrolliert. Die auditive Reizdarbietung und die Antworten der Vp konnten über die Wechselsprechanlage mitgehört werden. Gegebenenfalls notierte der Vl2 Besonderheiten oder Störungen im Versuchsablauf (z. B. verfrühte oder falsche Antworten, Bewegungen, Husten oder Räuspern der Vp, Störungen der Meßsignale etc.) auf eigens dafür vorgesehenen Protokollbögen (Anhang C.19 bzw. C.20).

Tabelle 12. Fragen und Antworten des DLT, Fragentypen in Abhängigkeit von der Tatbedingung

Trial	Frage	Ant-wort	Fragentyp Tat: Ring	Kette
1.	Befinden Sie sich momentan in Mainz?	J	IR	
2.	Haben Sie den Ring gestohlen?	N	RT	RA
3.	Sind Sie immer ehrlich?	N	WK	
4.	Haben Sie die Kette gestohlen?	N	RA	RT
5.	Sind Sie manchmal unehrlich?	N	LK	
6.	Befinden Sie sich momentan an der Universität?	J	IR	
7.	Haben Sie den Plastikkasten geöffnet?	N	RA	RT
8.	Handeln Sie manchmal illegal?	N	LK	
9.	Haben Sie die Glasschale geöffnet?	N	RT	RA
10.	Handeln Sie stets legal?	N	WK	
11.	Befinden Sie sich momentan am Psychologischen Institut?	J	IR	
12.	Sagen Sie manchmal die Unwahrheit?	N	LK	
13.	Tragen Sie die gestohlene Kette bei sich?	N	RA	RT
14.	Sagen Sie immer die Wahrheit?	N	WK	
15.	Tragen Sie den gestohlenen Ring bei sich?	N	RT	RA
16.	Befinden Sie sich momentan an der Universität?	J	IR	
17.	Haben Sie die Kette gestohlen?	N	RA	RT
18.	Handeln Sie stets legal?	N	WK	
19.	Haben Sie den Ring gestohlen?	N	RT	RA
20.	Handeln Sie manchmal illegal?	N	LK	
21.	Befinden Sie sich momentan am Psychologischen Institut?	J	IR	
22.	Sagen Sie immer die Wahrheit?	N	WK	
23.	Haben Sie die Glasschale geöffnet?	N	RT	RA
24.	Sagen Sie manchmal die Unwahrheit?	N	LK	
25.	Haben Sie den Plastikkasten geöffnet?	N	RA	RT
26.	Befinden Sie sich momentan in Mainz?	J	IR	
27.	Sind Sie manchmal unehrlich?	N	LK	
28.	Tragen Sie den gestohlenen Ring bei sich?	N	RT	RA
29.	Sind Sie immer ehrlich?	N	WK	
30.	Tragen Sie die gestohlene Kette bei sich?	N	RA	RT

Anmerkungen. Antwort: J = Ja, N = Nein. Fragentyp: RT = Relevant-Täuschung, RA = Relevant-Aufrichtigkeit, LK = Lügen-Kontroll, WK = Wahrheit-Kontroll, IR = Irrelevant.

Tabelle 13. Fragen und Items des GAT, Itemtypen in Abhängigkeit von Tatbedingung

Trial	Frage bzw. Item	Tat:	Itemtyp Ring	Itemtyp Kette
	Welches Schmuckstück haben Sie gestohlen? War es			
1.	eine Krawattennadel?		I1	
2.	eine Uhr?		IV	
3.	eine Brosche?		IV	
4.	ein Ring?		RT	RA
5.	ein Armreif?		IV	
6.	eine Kette?		RA	RT
	Aus welcher Schublade haben Sie den Schmuck gestohlen? War es			
7.	die Schublade Nr. 5?		I1	
8.	die Schublade Nr. 3?		RA	RT
9.	die Schublade Nr. 4?		IV	
10.	die Schublade Nr. 1?		IV	
11.	die Schublade Nr. 6?		RT	RA
12.	die Schublade Nr. 2?		IV	
	In welchem Gefäß befand sich das von Ihnen gestohlene Schmuckstück? Befand es sich			
13.	in einem Porzellanbehälter?		I1	
14.	in einer Pappschachtel?		IV	
15.	in einem Plastikkasten?		RA	RT
16.	in einer Holzschatulle?		IV	
17.	in einer Metalldose?		IV	
18.	in einer Glasschale?		RT	RA
	Welcher Gegenstand befand sich noch in der Schublade, aus der Sie den Schmuck gestohlen haben? War es			
19.	ein Bleistift?		I1	
20.	ein Füller?		IV	
21.	ein Locher?		RT	RA
22.	ein Hefter?		RA	RT
23.	ein Lineal?		IV	
24.	eine Schere?		IV	
	Welche Farbe hat der Stein des von Ihnen gestohlenen Schmuckstücks? Ist der Stein			
25.	schwarz?		I1	
26.	gelb?		RA	RT
27.	weiß?		IV	
28.	blau?		RT	RA
29.	grün?		IV	
30.	rot?		IV	
	Wie lautet der Buchstabe auf dem Etikett des von Ihnen gestohlenen Schmuckstücks? Ist es			
31.	ein Y?		I1	
32.	ein E?		RT	RA
33.	ein U?		IV	
34.	ein I?		IV	
35.	ein A?		RA	RT
36.	ein O?		IV	

Anmerkungen. Die Frage und das erste irrelevante Item (I1) waren jeweils zu einem Trial zusammengefaßt. Alle Items wurden verneint. Itemtyp: RT = Relevant-Täuschung, RA = Relevant-Aufrichtigkeit, IV = Irrelevant-Vergleich, I1 = Irrelevant-1.

Zum **Abschluß des Versuchsdurchgangs** beendete der Vl2 das EDA-MESS-Programm und die Aufzeichnungen des Varioports. Nach Betreten der Kabine notierte er die Temperatur und die relative Luftfeuchtigkeit auf dem Befragungsbogen (Anhang C.18). Der Vp wurden die Temperatur- und Plethysmographiesensoren sowie der Atemgurt abgenommen und die EDA- bzw. EKG-Elektroden von der Meßapparatur abgeklemmt. Das Entfernen der Elektroden erfolgte außerhalb der Kabine. Daraufhin kündigte der Vl2 die anschließende Nachbefragung an und begleitete die Vp in das Büro 02-628, wo bereits der Vl1 wartete. Der Vl2 begab sich wieder in den Laborraum und startete das Auslesen der Varioport-Memorycard. Außerdem führte er eine Kurzauswertung der Hautleitfähigkeitsdaten durch, auf der die Diagnose und die Entscheidung hinsichtlich der Belohnung basierten. Dazu wurden ähnlich wie bei Ben-Shakhar und Dolev (1996, S. 276) die Hautleitfähigkeitskurven für die unterschiedlichen Fragen- bzw. Itemtypen separat gemittelt und die Amplituden der mittleren Hautleitfähigkeitsreaktionen auf die Stimulusdarbietungen (unter Beachtung eines Latenzzeitkriteriums von mindestens einer Sekunde nach Reizbeginn) ausgemessen (Programme: EDA-SUM und EDA-VIEW, Münch, Universität Mainz). Falls die mittlere Reaktionsamplitude auf die Fragen bzw. Items nach dem jeweils begangenen Scheinverbrechen (Relevant-Täuschung) höher war als auf die anderen Fragen- bzw. Itemtypen (ausgenommen die irrelevanten Fragen des DLT und die ersten irrelevanten Items des GAT), dann wurde der Pb als „unglaubwürdig“ eingestuft. In allen anderen Fällen erhielt er die Belohnung von 20 DM.

4.4.4 Messung der erlebnisdeskriptiven Variablen, Gedächtnistests, Nachbefragung und Aufklärung

Der Vl1 legte der Vp je nach Testbedingung (DLT vs. GAT) die Fragebögen zur Erfassung der **erlebnisdeskriptiven abhängigen Variablen** vor. Zunächst sollte die Vp retrospektiv die Stärke ihrer körperlichen Reaktionen auf die einzelnen Fragen bzw. Items einstufen (vgl. Anhang C.10 und C.11: Rating 1a und 1b). Anschließend erfolgte die subjektive Einschätzung der Bedeutsamkeit der unterschiedlichen Fragen bzw. Items (siehe Anhang C.12 und C.13: Rating 2a und 2b). Der Pb sollte angeben, wie wichtig diese seiner Ansicht nach für das Ergebnis des „Lügendetektortests“ waren. Die Instruktionen für die Ratings betonten, daß die Fragebögen nicht zur Beurteilung der Glaubwürdigkeit herangezogen wurden. Die Vp konnte also frei antworten, ohne ihre Täterschaft verbergen zu müssen.

Nach den Ratings überprüfte man ebenfalls in Form von Papier-und-Bleistift-Tests die Behaltensleistung in bezug auf die kritischen Tatdetails. Die entsprechenden Verfahren

waren für beide Testbedingungen gleich. Zunächst bearbeitete die Vp einen **Gedächtnistest** zur hinweisreizbedingten Reproduktion (Anhang C.14: Gedächtnistest 1) und anschließend einen Wiedererkennungstest (Anhang C.15: Gedächtnistest 2). Der Vl1 legte den zweiten Testbogen erst nach Entgegennahme des ersten vor, so daß die Vp keine nachträglichen Korrekturen durchführen konnte.

Dann erhielt die Vp die Instruktion für die **Postexperimentellen Ratings** (Anhang C.16). Darin wurde sie aufgefordert, die eigene Täuschungsmotivation sowie die Treffsicherheit der „Lügendetektion“ zu beurteilen. Nachdem die Vp die beiden Einschätzungen auf dem dafür vorgesehenen Bogen vorgenommen hatte, folgte noch eine mündliche **Nachbefragung** in Form eines halbstandardisierten Interviews. Darin sollte die Vp Angaben über etwaige Täuschungsmanöver sowie über Selbstwahrnehmungen während der vorangegangenen Untersuchung machen. Außerdem wurde der Vp an dieser Stelle die Möglichkeit gegeben, an dem Experiment Kritik zu üben. Der Vl1 orientierte sich bei der Nachbefragung an einem vorgefertigten Befragungsbogen, auf dem die Angaben der Vp protokolliert wurden (siehe Anhang C.17).

Im Anschluß daran bat der Vl1 um die Rückgabe des entwendeten Schmuckstücks und begann, die Vp über die Zielsetzung der Untersuchung aufzuklären. Dabei wurde v. a. darauf geachtet, daß die fingierten Behauptungen der Instruktionen (z. B. Zusammenhang zwischen Intelligenz, emotionaler Kontrolle und Trefferquoten) zurückgenommen und klargestellt wurden. Am Ende der **Versuchsaufklärung** betrat der Vl2 das Büro und gab das Ergebnis der Kurzauswertung bekannt. Die Auswertung und die Diagnose wurden dem Pb anhand von Computerausdrucken der gemittelten Hautleitfähigkeitskurven erklärt und eventuell offene Fragen beantwortet. Im Falle eines negativen Befundes erhielt der Pb die Belohnung von 20 DM und unterschrieb eine Empfangsbestätigung. Von den insgesamt 122 Teilnehmern war dies bei 40 Pbn der Fall. Die anderen Vpn bekamen einen Trostpreis in Form von Süßigkeiten. Abschließend händigte man jedem Teilnehmer eine Kopie seiner Einverständniserklärung aus. Die Versuchsleiter bedanken sich nochmals beim Pb und erinnerten ihn daran, nichts von der Untersuchung weiterzuerzählen.

4.5 Physiologische Messungen

4.5.1 Hautleitfähigkeit

Die **Hautleitfähigkeit** wurde im Konstant-Spannungsverfahren mit 0.5 Volt gemessen (Lykken & Venables, 1971, S. 671). Die Ableitung erfolgte bipolar von der thenaren

und hypothenaren Erhebung der linken Hand (Vossel & Zimmer, 1998, S. 53). Dabei verwendete man Silber/Silberchlorid-(Ag/AgCl-)Elektroden der Firma Marquette Hellige (effektive Elektrodenfläche: ca. 0.8 cm^2) und eine 0.05-molare Natriumchlorid-Elektrolytpaste auf „Unibase“-Grundlage (vgl. Fowles et al., 1981, S. 235). Die Leitfähigkeitswerte wurden von einem im Psychologischen Institut entwickelten Transducer (CEDA-12) frequenzmoduliert und vom EDA-MESS-Steuerungscomputer mit einer Abtastrate von 10 Hz und einer Auflösung von 0.001 µS registriert.

4.5.2 Elektrokardiogramm (EKG)

Das **EKG** wurde mit Hilfe von EKG-Elektroden (Ag/AgCl) und Elektroden-Gel der Firma Marquette Hellige abgeleitet. Die Messung erfolgte als Brustwandableitung zwischen dem Manubrium sterni und dem linken untersten Rippenbogen, bei Erdung auf dem rechten untersten Rippenbogen. Das Signal wurde an den EKG-Verstärker des Varioports angelegt und mit einer Abtastrate von 512 Hz digitalisiert und gespeichert.

4.5.3 Photoplethysmogramm[6]

Die **Durchblutungsänderungen im Finger** wurden photoplethysmographisch ermittelt. Der Aufnehmer arbeitete nach dem Reflexionsprinzip mit Infrarot-Leuchtdioden als Sender (siehe Vossel & Zimmer, 1998, S. 76). Der Sensor war in eine anatomisch gewölbte Kunststoffschale eingearbeitet, die mittels eines Elektrodenkleberings und eines Klettbands an der distalen Phalanx des rechten Zeigefingers befestigt wurde. Sowohl der Signalaufnehmer als auch der regelbare Verstärker waren im Hause entwickelt worden (Juris, Universität Mainz). Da die Bestimmung der Fingerpulsvolumenamplitude (FPA) im Vordergrund stand, wählte man eine kurze Zeitkonstante von 0.115 Sekunden, um tonische Schwankungen des Blutvolumens zu unterdrücken (Jennings, Tahmoush & Redmond, 1980, S. 71, S. 80). Die Verstärkung wurde während der Ruhemessung so eingestellt, daß die FPA ca. 1 Volt Ausgangsspannung am Verstärker betrug. Das Signal wurde auf einen Kanal des Varioports gegeben und aufgezeichnet. Die Aufzeichnungsrate betrug 512 Hz. Da sich bei photoplethysmographischen

[6] Die photoplethysmographische Messung der peripheren Durchblutung zählt inzwischen zu den Standards der psychophysiologischen Aussagebeurteilung (vgl. Abschnitt 2.3). Darum wurde sie auch in der eigenen Studie erhoben. Die Parametrisierung und Ergebnisdarstellung sind jedoch nicht Gegenstand der vorliegenden Arbeit, zumal der Pilotcharakter dieser Untersuchung im Hinblick auf die quantitative Auswertung der Reaktionen zunächst eine eindringliche explorative Datenanalyse erforderlich macht.

Messungen eine Kontrolle der Hauttemperatur empfiehlt, wurde diese zusätzlich erhoben (vgl. Abschnitt 4.5.5).

4.5.4 Atmung[7]

Die **Atembewegungen** wurden mit dem Varioport-Atemgurt gemessen. Nach Walschburger (1976) überwiegt bei einer niedrigen bis mittleren Aktiviertheit die Bauchatmung gegenüber der Brustatmung. Dennoch erfaßte man hier entsprechend anderen Untersuchungen zur Respiration im Kontext der psychophysiologischen Aussageforschung die Änderungen des Brustumfangs beim Ein- und Ausatmen (vgl. Ben-Shakhar & Dolev, 1996, S. 275; Bradley & Rettinger, 1992, S. 57; Timm, 1982, S. 393). Demgemäß positionierte man den Atemgurt im Brustbereich (Vossel & Zimmer, 1998, S. 111). Das Signal wurde vom Varioport mit einer Zeitkonstanten von 20 Sekunden verstärkt und mit einer Aufzeichnungsrate von 32 Hz registriert.

4.5.5 Hauttemperatur

Zusätzlich zu den o. g. physiologischen Variablen bestimmte man kontinuierlich die **Hauttemperatur** an der Hand. Dazu wurde der Temperaturfühler des Varioport-Systems mit einem Elektrodenklebering an der distalen Phalanx des rechten Mittelfingers angebracht und das Signal mit einer Speicherrate von 1 Hz aufgezeichnet. Die Hauttemperatur war jedoch nicht als eigenständige abhängige Variable in die Versuchsplanung einbezogen. Sie diente lediglich als Kontrollvariable für die photoplethysmographische Messung der peripheren Durchblutung. Insofern sind die Ergebnisse auch nicht Gegenstand der vorliegenden Arbeit.

4.6 Erfassung der erlebnisdeskriptiven abhängigen Variablen

Nach der psychophysiologischen Aussagebeurteilung sollten die Vpn **Selbstratings der Reaktionsstärke** und eine subjektive **Einschätzung der Relevanz** der Fragen bzw. Items für das Testergebnis abgeben. Hierzu waren auf den entsprechenden Ratingbögen

[7] Die respiratorischen Maße wurden ebenfalls berücksichtigt, da es sich um eine Standardvariable der psychophysiologischen Aussagebeurteilung handelt (vgl. Abschnitt 2.3). Ähnlich wie beim Photopletysmogramm bestehen jedoch noch Unklarheiten hinsichtlich der Parametrisierung der Atemkurve, die eine gesonderte Aufbereitung der Daten voraussetzen. Aus diesem Grund ist die Darstellung der Auswertung und der Ergebnisse nicht Gegenstand der vorliegenden Arbeit.

für den DLT bzw. GAT (Anhang C.10 bis C.13) alle Fragen bzw. Items der Tests nochmals abgedruckt. Die Vpn markierten für jede Frage bzw. jedes Item auf siebenstufigen Selbstbeurteilungsskalen, wie stark sie ihrer Meinung nach darauf reagiert hatten und wie wichtig sie die Stimuli für die Beurteilung ihrer Glaubwürdigkeit bewerteten. Diese Methode war an die Vorgehensweise von Horowitz et al. (1997, S. 111) angelehnt.

4.7 Erfassung der Kontrollvariablen

Die **Gedächtnistests** orientierten sich an den Studien von Bradley und Mitarbeitern (z. B. Bradley & Warfield, 1984, S. 686). Zuerst absolvierten die Pbn einen Test zum Erinnern mit Hinweisreizen. Dabei wurden gezielt offene Fragen nach den kritischen Details der Scheinverbrechen gestellt (vgl. Anhang C.14). Die Pbn sollten ihre Antworten in die entsprechenden Felder eintragen. Der anschließende Wiedererkennungstest ähnelte dem Aufbau des GAT, wobei für jedes relevante Item der begangenen vs. nicht begangenen Tat eine Multiple-Choice-Frage mit den entsprechenden Alternativen konstruiert werden mußte (vgl. Anhang C.15). Die Pbn sollten die richtigen Lösungen auf dem Bogen ankreuzen. Die Reihenfolge der beiden Tests war für alle Vpn gleich. Es erschien nicht angemessen, den Wiedererkennungstest dem Wiedererinnern voranzustellen, da die Darbietung der Multiple-Choice-Alternativen eine Aktivierung der entsprechenden Gedächtnisinhalte bewirken kann (im Sinne eines „Priming“, vgl. Baddeley, 1997, S. 123). Solche Primingeffekte hätten möglicherweise dazu beitragen können, daß die relevanten Items im Anschluß an eine Überprüfung des Wiedererkennens besser erinnert worden wären. Die beiden Tests wurden auf getrennten Bögen dargeboten und die Pbn erhielten den zweiten Test erst nach Abgabe des ersten, damit keine nachträglichen Korrekturen möglich waren.

Die Messung der **Täuschungsmotivation** und der subjektiv eingeschätzten **Treffsicherheit** geschah anhand siebenstufiger Ratingskalen (Anhang C.16). Die Vpn sollten beurteilen, wie stark ausgeprägt ihre Motivation gewesen war, dem Untersucher die „Lügendetektion“ zu erschweren, und wie gut es dem Untersucher gelingen würde, nur anhand der körperlichen Reaktionen wahrheitsgemäße Antworten und Lügen zu unterscheiden.

Zum Abschluß der Untersuchung befragte man die Vpn im Rahmen eines **halbstandardisierten Interviews**, ob sie irgendeine Strategie, Taktik oder Technik angewendet hatten, um dem Untersucher die „Lügendetektion“ zu erschweren, und ob sie beim Lügen und Wahrheitsagen unterschiedliche Empfindungen, Gefühle, körperliche Reak-

tionen o. ä. verspürt hatten. Zudem wurde jede Vp angehalten, etwaige Anmerkungen, Kommentare und Anregungen zum Experiment abzugeben, z. B. ob ihr etwas angenehm oder unangenehm aufgefallen sei. Dazu las der Vl1 die jeweiligen Fragen vor, hakte je nach Antwort der Vp weiter nach und notierte die Angaben stichwortartig auf dem Bogen (vgl. Anhang C.17).

4.8 Auswertung, Datenreduktion und Parameterabstraktion

Eine wichtige Zielsetzung des Experiments bestand darin, die Effekte der Fragen- und Itemtypen auf die körperlichen Veränderungen direkt zu analysieren. Folglich wurden im Gegensatz zu den meisten Untersuchungen der psychophysiologischen Aussageforschung die Reaktionen **objektiv und quantitativ**, statt semi-objektiv und numerisch ausgewertet.

4.8.1 Hautleitfähigkeitsreaktionen (SCRs)

Die **Auswertung** der SCR-Amplituden erfolgte mittels eines interaktiven Auswertungsprogramms (EDA-VIEW, Münch, Universität Mainz). Die Hautleitfähigkeitsreaktionen waren generell über ein Latenzzeit-Kriterium von mindestens einer Sekunde definiert, d. h., als reizbedingte SCRs wurden artefaktfreie phasische Änderungen der Hautleitfähigkeit gewertet, die mindestens eine Sekunde nach dem Beginn bzw. Ende der Reizdarbietung (Ein- bzw. Ausblendung der Fragen oder Items auf dem Bildschirm) einen Anstiegspunkt zeigten. Die Breite der Zeitfenster, in denen die Fußpunkte der Reaktionen liegen mußten, waren von der jeweiligen Auswertungsmethode abhängig. Diese unterschiedlichen Kennwertbildungen werden im Ergebnisteil (Abschnitt 5.1.2) genauer spezifiziert.

Die **Amplitude** einer SCR ergab sich rechnerisch als Änderung der Hautleitfähigkeit (in µS) zwischen Reaktionsminimum und dem darauffolgenden Maximum. Die Extrempunkte wurden nicht mathematisch anhand Ableitungen, sondern stets näherungsweise unter visueller Inspektion bestimmt. In jenen Zweifelsfällen, wenn z. B. eine Spontanfluktuation ohne eine merkliche Absenkung von einer nachfolgenden Reaktion überlagert wurde, dienten auch deutlich erkennbare Übergänge der Anstiegsflanken (z. B. „Einkerbung“ bzw. Steigungsänderungen des Kurvenverlaufs) als Minima (vgl. Boucsein, 1988, S. 160). Die SCRs mußten einen Amplitudenwert von mindestens 0.003 µS aufweisen. Hautleitfähigkeitsänderungen, welche die o. g. Kriterien nicht erfüllten, waren als Nullreaktionen definiert.

4.8.2 Herzschlagfrequenz (HR)

Die auszuwertenden Bereiche der Varioport-Aufzeichnungen wurden zunächst mittels VitaGraph in ein ASCII-Datenformat konvertiert. Daraus wurden dann die EKG-Daten extrahiert, binärkonvertiert und in einzelne Trials zerlegt (Programm: EKGEX, Münch, Universität Mainz). Danach unterzog man die EKG-Verläufe einer computergestützten **R-Zacken-Erkennung** (Programm: EKG-VIEW, Münch, Universität Mainz; Wellenformanalyse unter Berücksichtigung eines Amplitudenkriteriums; vgl. auch Schandry, 1998, S. 138). Die anschließende Betrachtung der analog dargestellten EKG-Kurven und Triggerpunkte auf einem Grafikbildschirm bot die Möglichkeit, die Aufzeichnungen auf Artefakte hin zu untersuchen und die Kongruenz zwischen den R-Zacken und den automatisch gesetzten Triggern zu überprüfen. Fehler bei der Triggerung (z. B. Auslassungen oder Doppel-Triggerungen) konnten mittels einer interaktiven Option des Programms EKG-VIEW am Bildschirm visuell kontrolliert und manuell per Tastatur korrigiert werden.

Die vom Programm ermittelten zeitlichen R-Zacken-Abstände (RR-Intervalle in Millisekunden [ms]) wurden gemäß der von Velden und Wölk (1987, S. 173) bzw. Velden und Graham (1988, S. 292) vorgeschlagenen Formel in entsprechende Werte der Herzschlagfrequenz in Schläge pro Minute (HR in beats per minute [bpm]) auf eine **Sekunden-Echtzeitskalierung** transformiert (Programm: BPMS, Münch, Universität Mainz). Dieser Umrechnung liegt die modellhafte Annahme zugrunde, daß die HR-Schwankungen in einem linearen Zusammenhang die kontinuierlichen modulierenden Einflüsse des autonomen Nervensystems auf die diskret auftretenden Ereignisse der Herzaktivität wiedergeben (de Boer, Karemaker & Strackee, 1985, S. 150f.). Zudem ermöglicht die Echtzeitskalierung eine einfachere Verrechnung der Daten über Trials und Personen hinweg (z. B. bei Mittelungen, „Averaging“, vgl. Schandry, 1998, S. 147). Im Gegensatz zu den von Velden und Wölk (1987, S. 175) empfohlenen 500 ms betrug die Intervallänge auf der Echtzeitskala im vorliegenden Fall 1000 ms. Erfahrungsgemäß führen bei einer solchen Skalierung Arrhythmien in einem geringeren Umfang zu „Ausreißern“ in den HR-Verläufen. Außerdem erbringt nach Velden (1999, S. 93) die Verkürzung des Zeitintervalls von 1000 ms auf 500 ms nur einen geringen Genauigkeitszuwachs.

Ausgehend von der Annahme, daß reizbedingte phasische Veränderungen der HR vor dem Hintergrund tonischer Herzaktivität auftreten, wurde für jeden einzelnen Trial der Verlauf der HR in den Poststimulus-Intervallen (d. h. nach Ein- und Ausblendung der Fragen oder Items auf dem Bildschirm) als Abweichung vom Prästimulus-Wert der letzten Sekunde vor Reizbeginn (Einblendung) berechnet. Das Computerprogramm

BPMS bildete die **Differenzwerte** (ΔHR), indem es von allen HR-Werten der Poststimulus-Sekunden den HR-Wert in der Sekunde vor der Stimulusdarbietung (Prästimulus-Baseline) subtrahierte. Die weitere Auswertung der HR-Daten ist im Ergebnisteil (Abschnitt 5.2.2) ausführlich beschrieben.

4.8.3 Ratings, Nachbefragung und Gedächtnistests

Die von den Vpn auf den siebenstufigen **Ratingskalen** zur Reaktionsstärke, Bedeutsamkeit, Täuschungsmotivation und Treffsicherheit angekreuzten Werte wurden jeweils ausgezählt. Außerdem erfaßte man für jede einzelne Vp die Antwortkategorie (Bejahung oder Verneinung) hinsichtlich der beiden **Fragen** nach eventuellen Manipulationsversuchen bzw. nach etwaigen differentiellen Empfindungen beim Lügen und Wahrheitsagen.

Zur Auswertung der **Gedächtnistests** wurde jeweils die Anzahl der erinnerten bzw. wiedererkannten Details der Scheinverbrechen bestimmt (unter Beachtung der durchgeführten vs. nicht verübten Tat). Da jeder Gedächtnistest 12 Fragen beinhaltete (pro Tat jeweils 6), waren zwischen 0 und 12 Punkte möglich. Bei der Überprüfung des Erinnerns mit Hinweisreizen wurden die Antworten nicht nach dem genauen Wortlaut beurteilt. Entscheidend war, daß sie sinngemäß korrekt ausfielen. So wertete man eine Lösung auch dann als richtig, wenn der Pb z. B. eine „Kunststoffbox“ statt eines „Plastikkastens“ als Behältnis der Kette nannte. Mehrdeutige Antworten, wie etwa „durchsichtiges Gefäß“, galten jedoch als Fehler. Im Gegensatz dazu waren die korrekten Alternativen beim Wiedererkennungstest eindeutig definiert. Bei beiden Verfahren wurden Auslassungen bzw. Mehrfachnennungen ebenfalls als Fehler gewertet.

4.9 Statistische Auswertung

Die statistischen Analysen wurden mit dem Programmpaket SPSS (Version 10.0 für Windows) durchgeführt. Das A-priori-Alpha-Signifikanzniveau der Hypothesentests betrug 5%. Bei Varianzanalysen („analysis of variance“, ANOVA), die Meßwiederholungsfaktoren mit mehr als zwei Stufen umfaßten, wendete man zur Kompensation von Verletzungen der Sphärizitätsannahme eine Anpassung der Freiheitsgrade nach Greenhouse und Geisser (1959, S. 101) an (vgl. auch Vorberg & Blankenberger, 1999, S. 161). Im Ergebnisteil sind für die entsprechenden Meßwiederholungsfaktoren und deren Interaktionen stets die unkorrigierten Freiheitsgrade der F-Werte (F) sowie die entsprechenden p-Werte (p) und Epsilon (ε) gemäß der Greenhouse-Geisser-Korrektur

angegeben. Darüber hinaus ist Eta-Quadrat (η^2) als deskriptives Maß für die Varianzaufklärung dokumentiert, anhand dessen die Effektgröße veranschaulicht werden kann (Bortz, 1993, S. 236f.; Bortz & Döring, 1995, S. 571ff., S. 594f.). Zur genaueren Untersuchung signifikanter Haupt- und Interaktionseffekte führte man paarweise Einzelvergleiche (zweiseitige t-Tests) bzw. reduzierte ANOVAs durch. Da für viele Mittelwertsunterschiede A-priori-Hypothesen formuliert waren, wurde keine Adjustierung des Alpha-Fehlerniveaus vorgenommen, zumal die entsprechenden Verfahren (z. B. Bonferroni-Korrektur) v. a. bei abhängigen Tests (z. B. Meßwiederholungsfaktoren) sehr konservative Entscheidungen und damit einen erheblichen Teststärkeverlust nach sich ziehen können (Bortz, 1993, S. 125, S. 249f.).

5. Ergebnisse

5.1 Hautleitfähigkeitsreaktionen (SCRs)

5.1.1 Verlaufsmorphologie der gemittelten SCRs

Um die charakteristischen Verlaufsformen der SCRs auf die im vorliegenden Experiment dargebotenen Reizkonfigurationen näher bestimmen zu können, wurde eine **Mittelung** der Hautleitfähigkeitsdaten nach Ein- und Ausblendung der visuellen Stimuli durchgeführt (Programm: EDASUM, Münch, Universität Mainz). Dieses „Averaging“ erfolgte für den DLT und den GAT getrennt, d. h., die SCRs wurden jeweils über alle Personen und Trials der beiden Testbedingungen hinweg gemittelt. Die entsprechenden „Grand Averages“ sind in den Abbildungen 2 und 3 dargestellt.

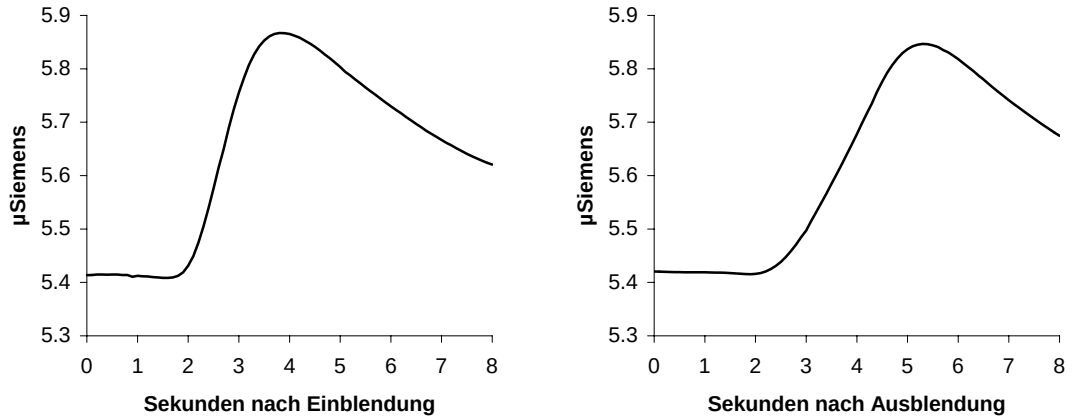


Abbildung 2. DLT: Grand Averages der SCRs nach Ein- und Ausblendung der Stimuli auf dem Bildschirm.

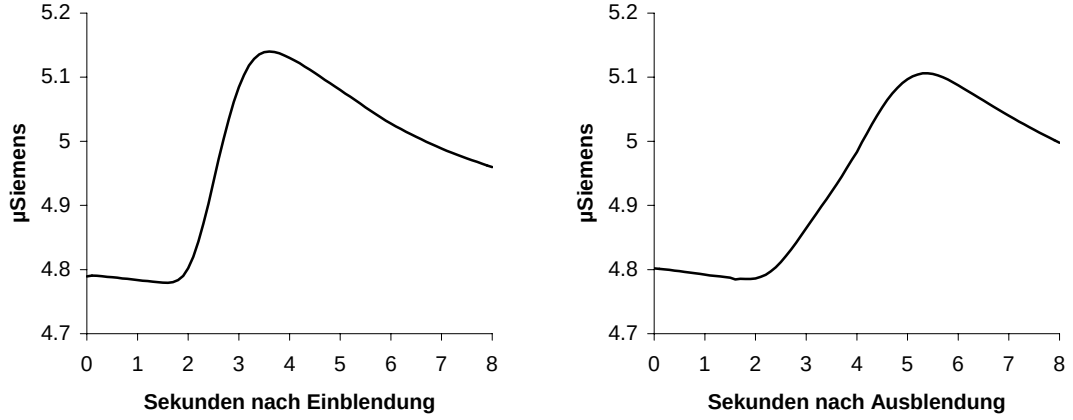


Abbildung 3. GAT: Grand Averages der SCRs nach Ein- und Ausblendung der Stimuli auf dem Bildschirm.

Sowohl beim DLT als auch beim GAT zeigten sich in den **mittleren Kurvenverläufen** SCRs im Anschluß an die Ein- und Ausblendung der Fragen bzw. Items. Die Fußpunkte dieser Reaktionen lagen im Bereich von 1 – 3 Sekunden nach Beginn bzw. Ende der Stimuluspräsentation. Dies stellt einen typischen Wertebereich der SCR-Latenz dar (Venables & Christie, 1980, S. 18f.) und stützt somit die zuvor getroffene Festlegung des Latenzzeitkriteriums auf mindestens eine Sekunde. Unter Berücksichtigung des Kriteriums ist anzunehmen, daß sowohl die Darbietung der Fragen bzw. Items als auch deren Ausblendung (imperativer Reiz der Antwortgabe) SCRs evozierten. Die gemittelten Reaktionen zeigten keine besonderen morphologischen Auffälligkeiten (vgl. „Idealfall“, „Typ 1“, Boucsein, 1988, S. 159).

5.1.2 Spezifikation der Parametrisierung

Im folgenden werden zunächst die **Auswertungsverfahren** hinsichtlich der SCRs auf die Darbietung der Fragen bzw. Items erläutert. Anschließend wird die Kennwertbildung bei den SCRs auf die Ausblendung geschildert.

Bezüglich der Reaktionen auf die Darbietung der Fragen bzw. Items wählte man in einem ersten Auswertungsschritt ein **konservatives Latenzfenster von 1 – 3 Sekunden nach Einblendung**. Ein solches Zeitintervall, in dem die Fußpunkte der Reaktionen liegen müssen, gilt für die meisten Untersuchungen zur SCR als angemessen (Vossel & Zimmer, 1998, S. 55). Der Reaktionsbeginn wurde als Minimum gewertet und die Amplitude zum darauffolgenden Reaktionsgipfel bestimmt (Maximum, definiert als eindeutiger Hochpunkt mit anschließender Steigungsumkehr).

In der Forschung und Anwendung der psychophysiologischen Aussagebeurteilung werden bei der EDA-Auswertung in der Regel breitere Zeitfenster gewählt. Dies liegt unter anderem daran, daß die Fragen und Items der Testverfahren komplexer und von längerer Dauer sind als das relativ einfache Reizmaterial, das häufig in Laborstudien zu elektrodermalen Reaktionen verwendet wird (z. B. kurze Töne). Darum ist im Rahmen der psychophysiologischen Aussagebeurteilung auch eher mit später einsetzenden bzw. mehrgipfligen SCRs zu rechnen. Dementsprechend wurde die zweite Auswertung **weniger konservativ** gefaßt. Man orientierte sich dabei an einer Reihe von Studien zum KFT, TWT und GAT, die das Zeitfenster auf bis zu 10 Sekunden nach Beginn der Fragen- bzw. Itemdarbietung ausweiteten (z. B. Bradley & Ainsworth, 1984, S. 66; Bradley, MacLaren & Carle, 1996, S. 156; Elaad, 1990, S. 523; Giesen & Rollison, 1980, S. 7). Unter Berücksichtigung des Latenzzeitkriteriums von einer Sekunde wurde die SCR mit der größten Amplitude innerhalb von **1 – 10 Sekunden nach Einblendung**

gewertet. Minimum und Maximum der SCR mußten in diesem Intervall liegen. Darüber hinaus war zu beachten, daß bei jenen Trials mit einer nur acht- bis neunsekündigen Bildschirmdarbietung bereits die Reaktion auf die Ausblendung im Zeitfenster liegen konnte. In solchen Fällen mußte also überprüft werden, ob die auszuwertende SCR einen Reaktionsgipfel im Bereich zwischen 9 und 10 Sekunden zeigte, und wenn ja, ob sie eventuell auf die Ausblendung zurückzuführen war. Nur jene Reaktionen, bei denen man dies ausschließen konnte, wurden gewertet. Andernfalls wurde die Amplitude der zweitstärksten Reaktion innerhalb des Zeitfensters herangezogen. Es sei darauf hingewiesen, daß die Amplitudenwerte der zweiten Auswertungsmethode nicht kleiner ausfallen konnten als die der ersten, sondern nur größer oder gleich. Das breitere Zeitintervall schloß nämlich das konservative Latenzfenster ein, und es wurde stets die höchste Reaktion darin bestimmt.

Da es sich bei der **Ausblendung der Fragen bzw. Items** auf dem Bildschirm um ein distinktes Ereignis handelte, legte man für die dadurch ausgelösten SCRs wiederum ein konservatives Latenzzeitkriterium von 1 – 3 Sekunden nach Reizende fest. Die Amplitude wurde auch hier ausgehend vom ersten Fußpunkt in Relation zum darauffolgenden Reaktionsgipfel gemessen.

5.1.3 Logarithmische Transformation der SCR-Amplituden

Aus statistischen Gründen wurde zur Annäherung an eine Normalverteilung zusätzlich eine **logarithmische Transformation** der Amplitudenwerte nach der von Venables und Christie (1980, S. 17) vorgeschlagenen Formel *log (1 + x)* durchgeführt. Die in der Regel signifikante Linksschiefe und Steilheit der SCR-Verteilung wird durch diese Logarithmierung weitgehend normalisiert, und aufgrund der Addition von eins sind auch Nullreaktionen mathematisch definiert (vgl. auch Boucsein, 1988, S. 179f.).

Um die Verteilungen vergleichen zu können, wurden die untransformierten und die logarithmierten Amplitudenwerte, die aus den drei o. g. Auswertungen resultierten, einer statistischen **Verteilungsanalyse** unterzogen. Die Ergebnisse sind im Anhang E in Form von Histogrammen graphisch veranschaulicht (Abbildungen 41 – 46). Eine zusammenfassende Gegenüberstellung der jeweiligen Verteilungskennwerte Schiefe und Exzeß (Kurtosis) findet sich in Tabelle 14.

Für eine Normalverteilung erwartet man dabei einen Schiefe- (*S*) und einen Exzessivitätskennwert (*E*) von jeweils null (Bortz, 1993, S. 46). Positive Werte indizieren **Linksschiefe** und **Schmalgipfligkeit** (Hyperexzessivität). Die Verteilungen der nicht

logarithmierten Rohwerte fielen somit erwartungsgemäß asymmetrisch linksschief und schmalgipflig aus (vgl. auch Anhang E).

Tabelle 14. Die Kennwerte Schiefe (*S*) und Exzeß (*E*) für die Verteilungen der SCR-Amplituden, getrennt nach den drei Auswertungen und untransformierten vs. logarithmierten Daten

Auswertung	untransformiert	logarithmiert
SCRs nach Einblendung	$S = 2.27$	$S = 1.17$
(konservatives Latenzfenster: 1 – 3 s)	$E = 7.00$	$E = 0.80$
SCRs nach Einblendung	$S = 2.21$	$S = 1.12$
(breites Latenzfenster: 1 – 10 s)	$E = 6.70$	$E = 0.71$
SCRs nach Ausblendung	$S = 2.73$	$S = 1.43$
(Latenzfenster: 1 – 3 s)	$E = 11.61$	$E = 2.00$

Anmerkung: s = Sekunden.

Durch die logarithmische Transformation wurde die Linksschiefe zwar **in Richtung einer Normalverteilung korrigiert**, die anomale Asymmetrie konnte jedoch nicht eliminiert werden, was auf einen relativ hohen Anteil von Nullreaktionen zurückzuführen war, deren Logarithmus entsprechend der Formel *log (1 + 0) = log 1* ebenfalls null betrug. Die Exzessivitätskennwerte der logarithmierten Daten deuteten darauf hin, daß die Verteilungen infolge der Transformation breiter wurden. Die Hyperexzessivität blieb jedoch erhalten.

Auf zusätzliche Korrekturverfahren (z. B. Ausgangswert- oder Bereichskorrektur) verzichtete man, zumal deren Zweckmäßigkeit und praktische Umsetzung umstritten sind (Vossel, 1990, S. 58ff.). Für jede Vp und Auswertungsmethode separat wurden die Amplitudenwerte über die Trials der jeweiligen Fragen- bzw. Itemtypen gemittelt (unter Berücksichtigung von Nullreaktionen) und die resultierenden mittleren Amplituden (SCR-Magnituden) **inferenzstatistisch analysiert**.

Im Rahmen der **Ergebnisdarstellung** werden zunächst nur die Resultate der **logarithmierten SCR-Amplituden** referiert, weil deren Verteilungscharakteristika deutlich vorteilhafter im Sinne einer Annäherung an die Normalverteilung ausfielen. Sofern sich bedeutende Unterschiede zwischen den Befunden der untransformierten und logarithmierten Daten zeigten, wird an entsprechender Stelle näher darauf eingegangen.

5.1.4 SCRs nach Einblendung der Fragen bzw. Items

Da die Auswertungen der SCRs nach Einblendung der Stimuli für die beiden unterschiedlichen Zeitfenster weitgehend gleichartige Ergebnisse erbrachten, werden hier lediglich die **Resultate der konservativen Auswertung** (Latenzfenster: 1 – 3 Sekunden) im Detail berichtet und die wesentlichen Diskrepanzen zur Amplitudenbestimmung im breiten Latenzfenster (1 – 10 Sekunden) anschließend kurz erläutert.

Es wurden für die beiden Tests (DLT vs. GAT) getrennt **3-faktorielle Varianzanalysen** mit den Gruppenfaktoren Tatbedingung (Ring vs. Kette) und Elektrodermale Labilität (Stabil vs. Labil) und dem Meßwiederholungsfaktor Fragen- bzw. Itemtyp durchgeführt.

Beim **DLT** resultierte lediglich ein signifikanter Haupteffekt des Fragentyps, $F(4,144) = 23.78$, $p < .001$, $\varepsilon = .584$, $\eta^2 = .40$. Da sich weder für die Tatbedingung noch die EL Haupt- oder Interaktionseffekte ergaben, wurden die Daten über die Stufen der beiden Gruppenfaktoren aggregiert. Es sei jedoch vermerkt, daß erwartungsgemäß elektrodermal Labile ($M = 0.1853$ log µS, $SD = 0.0884$) im Durchschnitt stärker reagierten als Stabile ($M = 0.1374$ log µS, $SD = 0.1041$); dieser Unterschied fiel aber nicht statistisch bedeutsam aus, $F(1,36) = 2.34$, $p = .135$, $\eta^2 = .06$. Die Mittelwerte für die fünf Fragentypen sind in Tabelle 15 wiedergegeben und in Abbildung 4 veranschaulicht.

Tabelle 15. DLT – SCRs nach Einblendung der Fragen: Mittelwerte (*M*) und Standardabweichungen (*SD*) der SCR-Magnituden (in log µS) der fünf Fragentypen

	Fragentyp				
	Relevant-Täuschung	Relevant-Aufrichtigkeit	Lügen-Kontroll	Wahrheit-Kontroll	Irrelevant
M	0.2275_{a}	0.1734_{b}	0.1227_{c}	$0.1357_{c,d}$	0.1475_{d}
SD	0.1457	0.1075	0.0975	0.0901	0.0977

Anmerkung. Mittelwerte mit unterschiedlichem Subskript unterscheiden sich auf dem 5%-Signifikanzniveau.

Paarweise Vergleiche (zweiseitige t-Tests für abhängige Stichproben ohne Korrektur des Alpha-Fehlerniveaus) zeigten, daß die beiden Arten von relevanten Fragen jeweils signifikant stärkere Reaktionen auslösten als die anderen Fragentypen (für die betreffenden t-Werte [*t*s] und p-Werte [*p*s] galt: *t*s[39] > 2.26, *p*s < .029; vgl. auch die Mittelwerte in Tabelle 15). Entsprechend den Erwartungen war die SCR-Magnitude unter der Bedingung Relevant-Täuschung höher als unter der Bedingung Relevant-Aufrichtigkeit, $t(39) = 3.53$, $p = .001$. Die beiden Kontrollfragentypen (Lügen-Kontroll vs. Wahrheit-

Kontroll) unterschieden sich jedoch nicht signifikant voneinander, $t(39) = 1.73$, $p = .092$. Die irrelevanten Fragen evozierten insgesamt stärkere Reaktionen als die Lügen-Kontrollfragen, $t(39) = 3.11$, $p = .003$, wohingegen der Vergleich zu Wahrheit-Kontroll nicht signifikant ausfiel, $t(39) = 1.23$, $p = .225$.

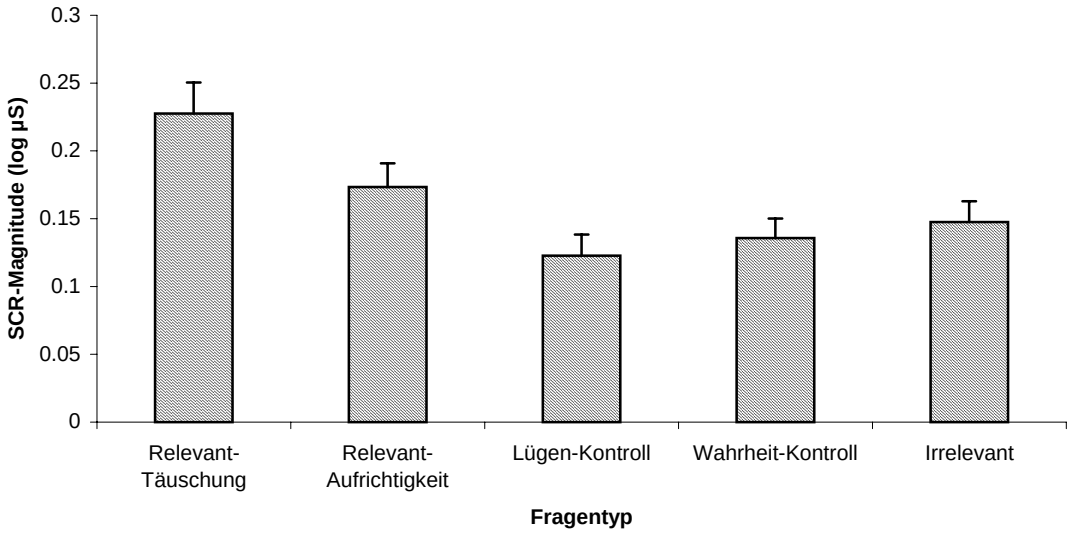


Abbildung 4. DLT – SCRs nach Einblendung der Fragen: Vergleich der SCR-Magnituden der fünf Fragentypen; dargestellt sind Mittelwerte (Säulen) und deren Standardfehler (aufgesetzte vertikale Linien).

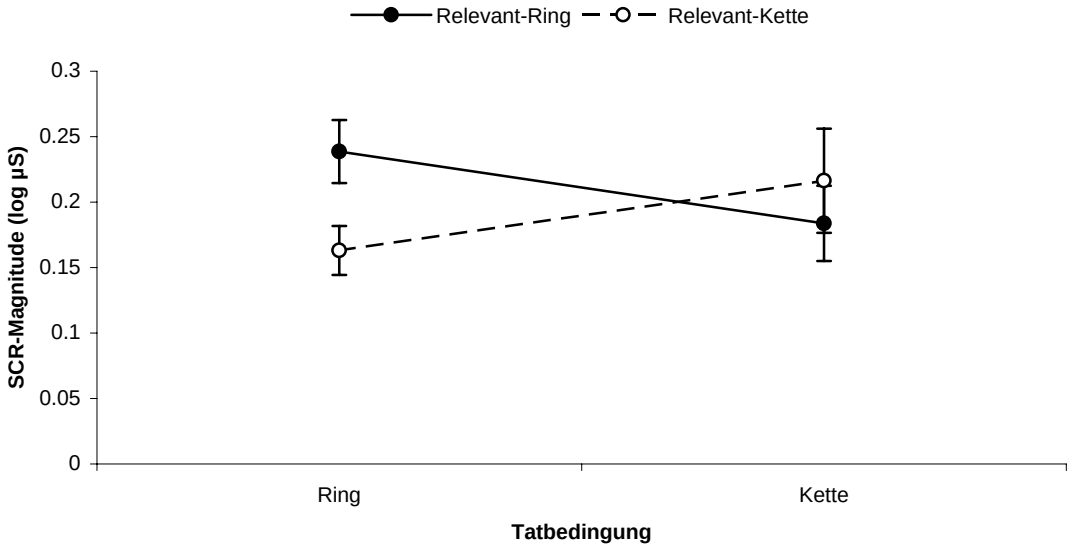


Abbildung 5. DLT – SCRs nach Einblendung der Fragen: Vergleich der SCR-Magnituden der Fragentypen Relevant-Ring und -Kette, getrennt für die beiden Tatbedingungen; dargestellt sind Mittelwerte (Punkte) und deren Standardfehler (vertikale Linien).

Zur genaueren Untersuchung der **Effekte der relevanten Fragen** wurde eine 2 × 2 × 2-ANOVA mit den Gruppenfaktoren Tatbedingung (Ring vs. Kette) und EL (Stabil vs. Labil) und dem Meßwiederholungsfaktor Tatbezug der relevanten Fragen (Relevant-Ring vs. Relevant-Kette) gerechnet. Hypothesenkonform fand man eine **disordinale Interaktion** zwischen Tatbedingung und Tatbezug, $F(1,36) = 13.28$, $p = .001$, $\eta^2 = .27$, die in Abbildung 5 ersichtlich ist. Die Pbn reagierten insgesamt stärker auf die Fragen nach dem eigenen Scheinverbrechen als auf die inhaltlich parallelisierten Fragen nach der nicht begangenen Tat (Tatbedingung Ring: Relevant-Ring: $M = 0.2386$ log µS, $SD = 0.1079$, Relevant-Kette: $M = 0.1630$ log µS, $SD = 0.0836$; Tatbedingung Kette: Relevant-Ring: $M = 0.1838$ log µS, $SD = 0.1284$, Relevant-Kette: $M = 0.2164$ log µS, $SD = 0.1779$). Ansonsten resultierten aus dieser Analyse keine weiteren signifikanten Befunde.

Beim **GAT** erbrachte die 2 × 2 × 4- (Tatbedingung × EL × Itemtyp-) ANOVA signifikante Haupteffekte der EL, $F(1,36) = 19.71$, $p < .001$, $\eta^2 = .35$, und des Itemtyps, $F(3,108) = 27.30$, $p < .001$, $\varepsilon = .757$, $\eta^2 = .43$, sowie eine Interaktion zwischen dem Gruppen- und dem Meßwiederholungsfaktor, $F(3,108) = 3.51$, $p = .029$, $\varepsilon = .757$, $\eta^2 = .09$. In Tabelle 16 sind die Mittelwerte und Standardabweichungen der Itemtypen insgesamt und getrennt für Stabile und Labile angegeben. Abbildung 6 veranschaulicht die hybride (semi-disordinale) Wechselwirkung.

Tabelle 16. GAT – SCRs nach Einblendung der Items: Mittelwerte (*M*) und Standardabweichungen (*SD*) der SCR-Magnituden (in log µS) der vier Itemtypen, gesamt und getrennt für elektrodermal stabile und labile Pbn

		Itemtyp			
EL		Relevant-Täuschung	Relevant-Aufrichtigkeit	Irrelevant-Vergleich	Irrelevant-1
Stabil	*M*	0.0937	0.0676	0.0432	0.0927
	SD	0.0829	0.0542	0.0329	0.0593
Labil	*M*	0.2485	0.1881	0.1567	0.2603
	SD	0.1460	0.1393	0.1045	0.1414
Gesamt	*M*	0.1711_a	0.1278_b	0.1000_c	0.1765_a
	SD	0.1409	0.1209	0.0957	0.1366

Anmerkung. Gesamtmittelwerte mit unterschiedlichem Subskript unterscheiden sich auf dem 5%-Signifikanzniveau.

Paarvergleiche (zweiseitige t-Tests für abhängige Stichproben) ergaben, daß sich bis auf die Bedingungen Relevant-Täuschung und Irrelevant-1 ($t < 1$) alle anderen Gesamt-

mittelwerte der Itemtypen signifikant voneinander unterschieden (*ts*[39] > 3.84, *p*s < .001; vgl. auch Tabelle 16).

Um den Haupteffekt des Itemtyps angemessen interpretieren zu können, mußte jedoch die **Interaktion** des Faktors mit der EL näher analysiert werden. In Abbildung 6 ist zu erkennen, daß elektrodermal Labile über alle Itemtypen hinweg höhere SCR-Magnituden aufwiesen als Stabile (vgl. Haupteffekt der EL). Sowohl Stabile als auch Labile reagierten auf die relevanten Itemtypen (Relevant-Täuschung und -Aufrichtigkeit) jeweils stärker als auf die Irrelevant-Vergleich-Items. Zudem war für beide Gruppen die SCR-Magnitude unter der Bedingung Relevant-Täuschung höher als unter der Bedingung Relevant-Aufrichtigkeit. Beim Vergleich zwischen den Itemtypen Relevant-Täuschung und Irrelevant-1 waren jedoch Diskrepanzen erkennbar. Während Labile im Durchschnitt etwas stärker auf die ersten irrelevanten Items der Multiple-Choice-Fragen reagierten, war bei den Stabilen eine umgekehrtes Reaktionsmuster mit leicht stärkeren Reaktionen auf die Relevant-Täuschung-Items erkennbar (vgl. auch Tabelle 16). Dies legte nahe, daß die Interaktion zwischen der EL und dem Itemtyp z. T. auf die Irrelevant-1-Items zurückzuführen war. Da bei der üblichen Auswertung des GAT die entsprechenden Reaktionen nicht berücksichtigt werden, führte man eine reduzierte 2 × 2 × 3- (Tatbedingung × EL × Itemtyp-) ANOVA unter **Ausschluß der Bedingung Irrelevant-1** durch. Daraus resultierten nur noch signifikante Haupteffekte der EL, $F(1,36) = 17.23$, $p < .001$, $\eta^2 = .32$, und des Itemtyps, $F(2,72) = 29.46$, $p < .001$, $\varepsilon = .851$, $\eta^2 = .45$, wohingegen die Wechselwirkung zwischen den beiden Faktoren nicht mehr signifikant ausfiel, $F(2,72) = 2.78$, $p = .078$, $\varepsilon = .851$, $\eta^2 = .07$. Dies stützte die Annahme, daß die o. g. Interaktion insbesondere auf dem Itemtyp Irrelevant-1 fußte.

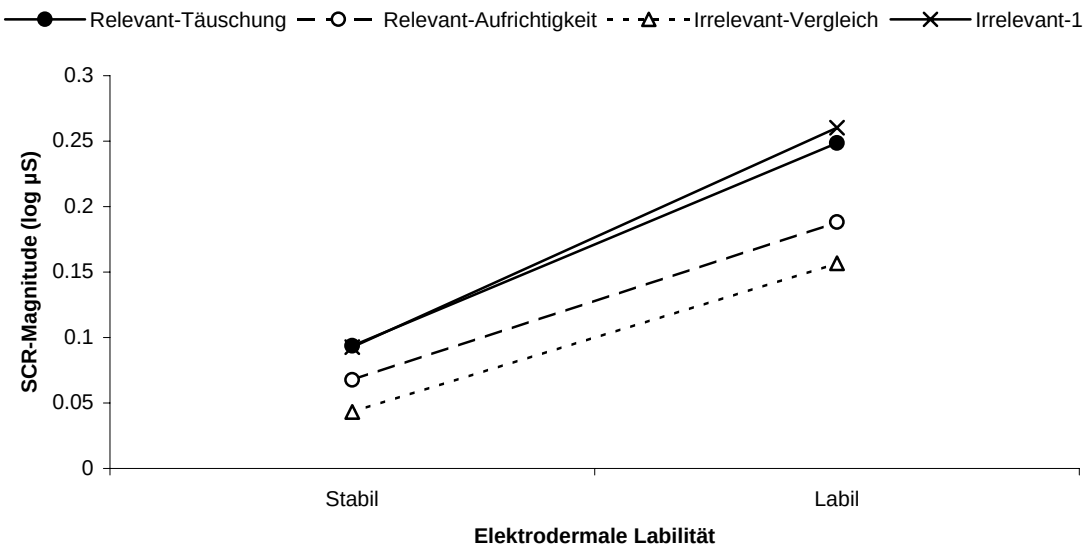


Abbildung 6. GAT – SCRs nach Einblendung der Items: Vergleich der SCR-Magnituden der vier Itemtypen, getrennt für elektrodermal stabile und labile Pbn.

Um die **Effekte der relevanten Items** genauer analysieren zu können, wurde eine 2 × 2 × 2-ANOVA mit den Gruppenfaktoren Tatbedingung (Ring vs. Kette) und EL (Stabil vs. Labil) und dem Meßwiederholungsfaktor Tatbezug der relevanten Items (Relevant-Ring vs. Relevant-Kette) gerechnet. Es manifestierten sich lediglich ein Labilitätseffekt, $F(1,36) = 15.56$, $p < .001$, $\eta^2 = .30$, der darauf beruhte, daß Labile ($M = 0.2183$ log µS, $SD = 0.1376$) insgesamt stärker auf die relevanten Items reagierten als Stabile ($M = 0.0807$ log µS, $SD = 0.0656$), und eine disordinale Interaktion zwischen der Tatbedingung und dem Meßwiederholungsfaktor, $F(1,36) = 18.14$, $p < .001$, $\eta^2 = .34$. Die Pbn zeigten stärkere Reaktionen auf Items, die das begangene Scheinverbrechen thematisierten, als auf Details der nicht begangenen Tat (vgl. Abbildung 7; Tatbedingung Ring: Relevant-Ring: $M = 0.1704$ log µS, $SD = 0.1244$, Relevant-Kette: $M = 0.1178$ log µS, $SD = 0.0933$; Tatbedingung Kette: Relevant-Ring: $M = 0.1379$ log µS, $SD = 0.1451$, Relevant-Kette: $M = 0.1718$ log µS, $SD = 0.1591$).

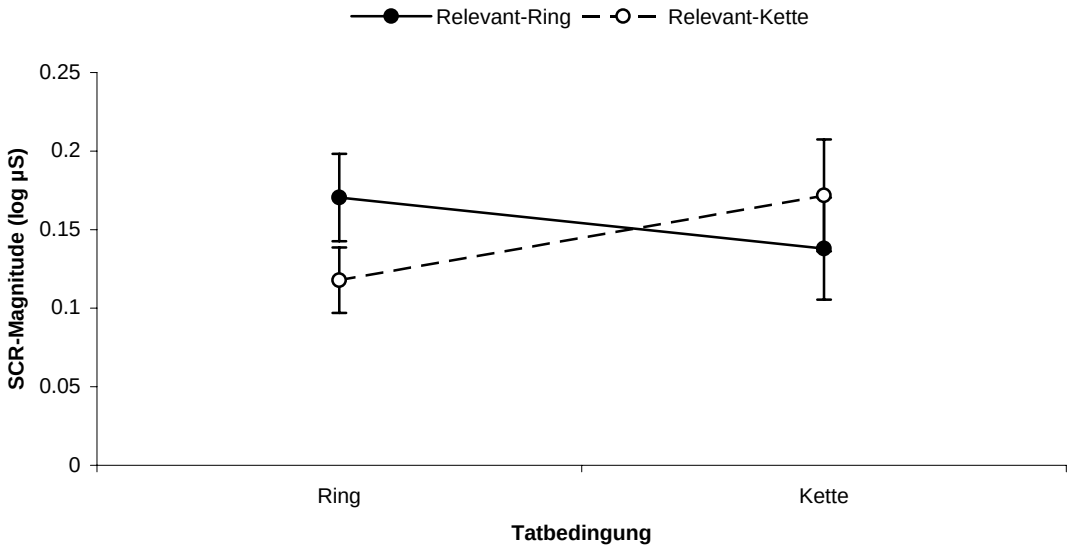


Abbildung 7. GAT – SCRs nach Einblendung der Items: Vergleich der SCR-Magnituden der Itemtypen Relevant-Ring und -Kette, getrennt für die beiden Tatbedingungen; dargestellt sind Mittelwerte (Punkte) und deren Standardfehler (vertikale Linien).

Die varianzanalytische Auswertung der **untransformierten Amplitudenwerte** der SCRs mit einer Latenz von 1 – 3 Sekunden nach Einblendung erbrachte überwiegend analoge Ergebnisse. Lediglich beim **GAT** resultierte aus der 3-faktoriellen (2 × 2 × 4-) ANOVA eine nur marginal signifikante ($p < .1$) Interaktion zwischen der EL und dem Itemtyp, $F(3,108) = 2.79$, $p = .072$, $\varepsilon = .631$, $\eta^2 = .07$. Diese Wechselwirkung verfehlte auch bei der SCR-Auswertung im **breiten Latenzfenster** (1 – 10 Sekunden) die Signifikanzgrenze von 5%, und zwar sowohl für die logarithmierten ($F[3,108] = 2.76$, $p = .061$, $\varepsilon = .771$, $\eta^2 = .07$) als auch die untransformierten Amplitudenwerte ($F[3,108] = 2.45$, $p = .093$, $\varepsilon = .674$, $\eta^2 = .06$). Alle anderen Effekte waren (unabhängig

von Latenzfenster oder Transformation) durchweg mit den o. g. Ergebnissen vergleichbar.

5.1.5 SCRs nach Ausblendung der Fragen bzw. Items

Die 3-faktorielle Varianzanalyse (Tatbedingung × EL × Fragentyp) der logarithmierten Amplitudenwerte ergab für den **DLT** einen signifikanten Haupteffekt des Fragentyps, $F(4,144) = 2.84$, $p = .031$, $\varepsilon = .906$, $\eta^2 = .07$, und eine Interaktion 2. Ordnung zwischen den beiden Gruppenfaktoren und dem Meßwiederholungsfaktor, $F(4,144) = 3.05$, $p = .023$, $\varepsilon = .906$, $\eta^2 = .08$. In Tabelle 17 sind die Mittelwerte und Standardabweichungen der fünf Fragentypen insgesamt und für die vier Gruppen getrennt angegeben. Ähnlich wie bei den Reaktionen auf die Einblendung fand man auch hier für die elektrodermal labilen Pbn ($M = 0.1601$ log µS, $SD = 0.0713$) im Durchschnitt höhere SCR-Magnituden als für die stabilen ($M = 0.1072$ log µS, $SD = 0.0939$); dieser Effekt verfehlte aber knapp die Signifikanzgrenze, $F(1,36) = 4.07$, $p = .051$, $\eta^2 = .10$.

Tabelle 17. DLT – SCRs nach Ausblendung der Fragen: Mittelwerte (*M*) und Standardabweichungen (*SD*) der SCR-Magnituden (in log µS) der Fragentypen, gesamt und getrennt für die vier Gruppen

		Fragentyp				
EL		Relevant-Täuschung	Relevant-Aufrichtigkeit	Lügen-Kontroll	Wahrheit-Kontroll	Irrelevant
		Tatbedingung Ring				
Stabil	*M*	0.1256	0.0935	0.1249	0.1183	0.1046
	SD	0.0786	0.0693	0.0761	0.1040	0.0875
Labil	*M*	0.2238	0.1822	0.1976	0.1782	0.1608
	SD	0.1392	0.0870	0.0995	0.1019	0.0925
		Tatbedingung Kette				
Stabil	*M*	0.1187	0.1243	0.0928	0.0664	0.1028
	SD	0.0861	0.0857	0.0698	0.0538	0.0928
Labil	*M*	0.1415	0.1089	0.1233	0.1346	0.1500
	SD	0.0685	0.0954	0.0843	0.1059	0.1056
Gesamt	*M*	0.1524_a	0.1272_b	0.1346_b	0.1244_b	0.1295_b
	SD	0.1025	0.0884	0.0889	0.0990	0.0949

Anmerkung. Die Gesamtmittelwerte der Fragentypen mit unterschiedlichem Subskript unterscheiden sich auf dem 5%-Signifikanzniveau.

Abbildung 8 illustriert den Haupteffekt des Fragentyps. **Paarweise Vergleiche** (einfache zweiseitige t-Tests) zeigten, daß die Fragen der Bedingung Relevant-Täuschung signifikant stärkere Reaktionen auslösten als die restlichen Fragentypen (*t*s[39] > 2.13, *p*s < .039; vgl. auch die Gesamtmittelwerte in Tabelle 17). In allen anderen Paarvergleichen manifestierten sich keine statistisch bedeutsamen Unterschiede (*t*s[39] < 1.17, *p*s > .249). Der Haupteffekt ist jedoch aufgrund der signifikanten 3-fachen Interaktion zu relativieren. Die Abbildungen 9 und 10 veranschaulichen die SCR-Magnituden der einzelnen Fragentypen nach Tatbedingung und EL getrennt.

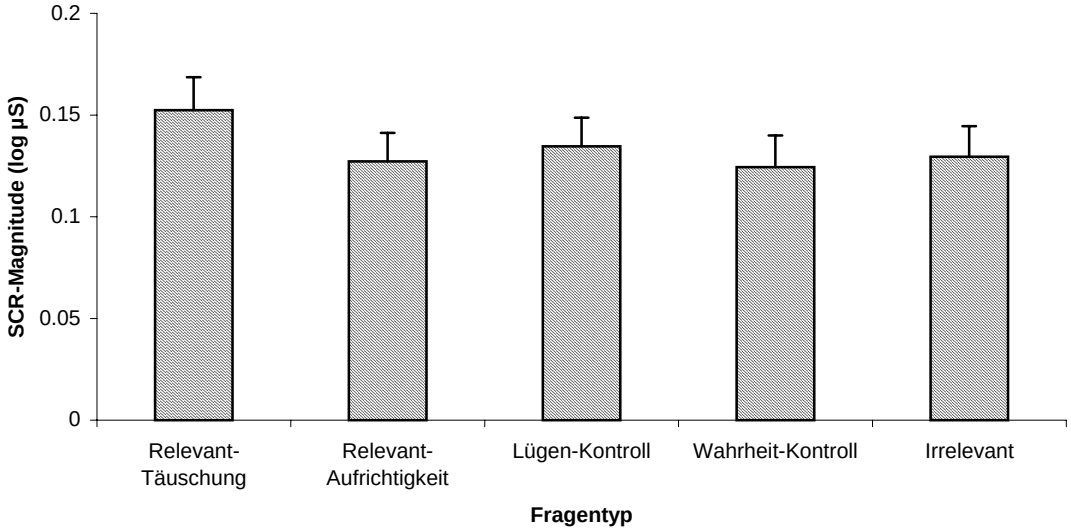


Abbildung 8. DLT – SCRs nach Ausblendung der Fragen: Vergleich der SCR-Magnituden der fünf Fragentypen; dargestellt sind Mittelwerte (Säulen) und deren Standardfehler (vertikale Linien).

Zur weiteren **Auflösung dieser komplexen Wechselwirkung** wurden reduzierte 2-faktorielle ANOVAs (EL × Fragentyp) separat für die Tatbedingungen Ring und Kette gerechnet. Deren Ergebnisse sind in Tabelle 18 aufgelistet.

Tabelle 18. DLT – SCRs nach Ausblendung der Fragen: Ergebnisse der reduzierten 2-fachen Varianzanalysen, nach Tatbedingung (Ring vs. Kette) getrennt

Faktor	Ring	Kette
EL	n. s.	n. s.
Fragentyp	$p < .05$	n. s.
EL × Fragentyp	n. s.	$p < .05$

Anmerkung. n. s. = nicht signifikant.

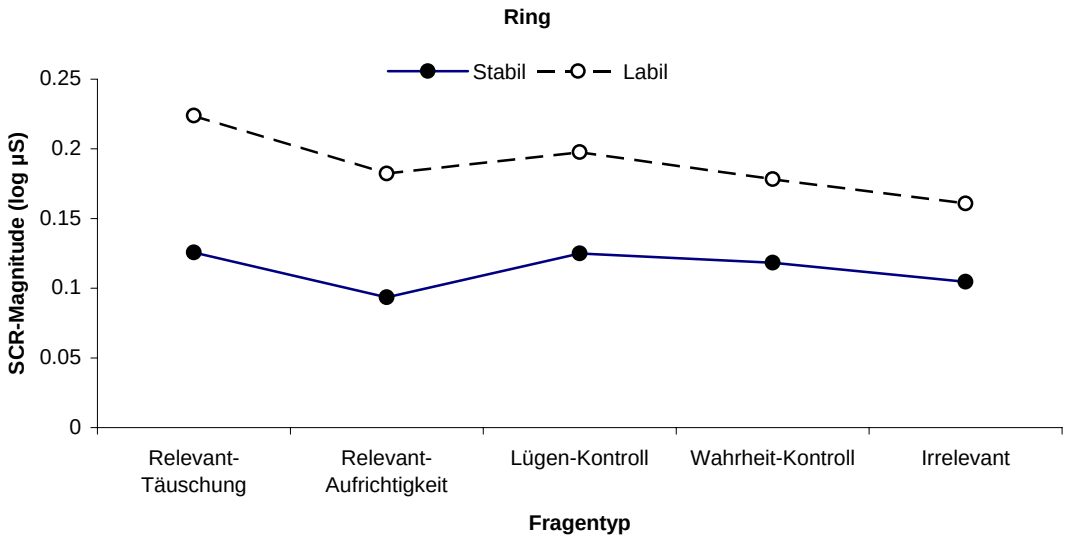


Abbildung 9. DLT – SCRs nach Ausblendung der Fragen: Vergleich der SCR-Magnituden der fünf Fragentypen, getrennt für elektrodermal stabile und labile Pbn der Tatbedingung Ring.

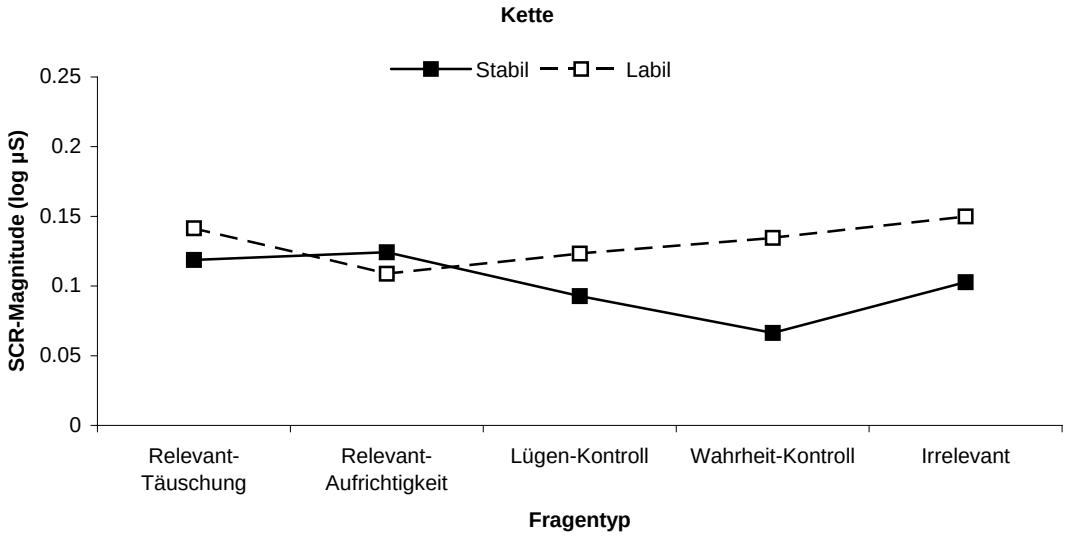


Abbildung 10. DLT – SCRs nach Ausblendung der Fragen: Vergleich der SCR-Magnituden der fünf Fragentypen, getrennt für elektrodermal stabile und labile Pbn der Tatbedingung Kette.

Der Haupteffekt des Fragentyps schlug sich nur in der Ring-Gruppe nieder, während unter der Bedingung Kette lediglich die Interaktion des Meßwiederholungsfaktors mit der EL signifikant wurde. Dementsprechend wird aus Abbildung 9 ersichtlich, daß unter der Bedingung Ring die SCR-Magnituden von Stabilen und Labilen über die Fragentypen hinweg weitgehend parallel verliefen. Im Gegensatz dazu traten unter der Bedingung Kette stärkere Abweichungen auf (vgl. Abbildung 10). Es fielen dort insbesondere

die Reaktionsunterschiede bei den relevanten Fragentypen (Labile reagierten erwartungsgemäß stärker auf Relevant-Täuschung als auf Relevant-Aufrichtigkeit, Stabile zeigten ein umgekehrtes Muster) und eine deutliche Diskrepanz zwischen den Gruppen bei den Wahrheit-Kontrollfragen auf. Wenn man die letzten beiden Fragentypen im Rahmen reduzierter 2 × 2 × 4- (Tatbedingung × EL × Fragentyp-) ANOVAs jeweils eliminierte, wurde die 3-fache Interaktion nicht mehr signifikant (Ausschluß Relevant-Aufrichtigkeit: $F[3,108] = 2.03$, $p = .118$, $\varepsilon = .955$, $\eta^2 = .05$; Ausschluß Wahrheit-Kontroll: $F[3,108] = 2.58$, $p = .060$, $\varepsilon = .962$, $\eta^2 = .07$). Diese Befunde deuteten insgesamt darauf hin, daß die 3-fache Interaktion der Initialanalyse v. a. auf die Reaktionsunterschiede zwischen stabilen und labilen Pbn der Tatbedingung Kette bei den Relevant-Aufrichtigkeit- und Wahrheit-Kontroll-Fragen zurückzuführen war.

Zur detaillierteren Prüfung der **Effekte der relevanten Fragen** wurde wiederum eine 2 × 2 × 2-ANOVA (Tatbedingung × EL × Tatbezug der relevanten Fragen) gerechnet. Der einzige daraus resultierende signifikante Effekt war eine 2-fache semi-disordinale Wechselwirkung der Faktoren Tatbedingung und Tatbezug, $F(1,36) = 7.31$, $p = .01$, $\eta^2 = .17$ (vgl. Abbildung 11). Die Pbn reagierten insgesamt stärker auf die Fragen nach dem eigenen Scheinverbrechen als nach der nicht begangenen Tat (Tatbedingung Ring: Relevant-Ring: $M = 0.1747$ log µS, $SD = 0.1210$, Relevant-Kette: $M = 0.1378$ log µS, $SD = 0.0891$; Tatbedingung Kette: Relevant-Ring: $M = 0.1166$ log µS, $SD = 0.0886$, Relevant-Kette: $M = 0.1301$ log µS, $SD = 0.0766$), wobei der Reaktionsunterschied bei den Pbn, die die Kette entwendet hatten, deutlich kleiner war als bei der Ring-Gruppe.

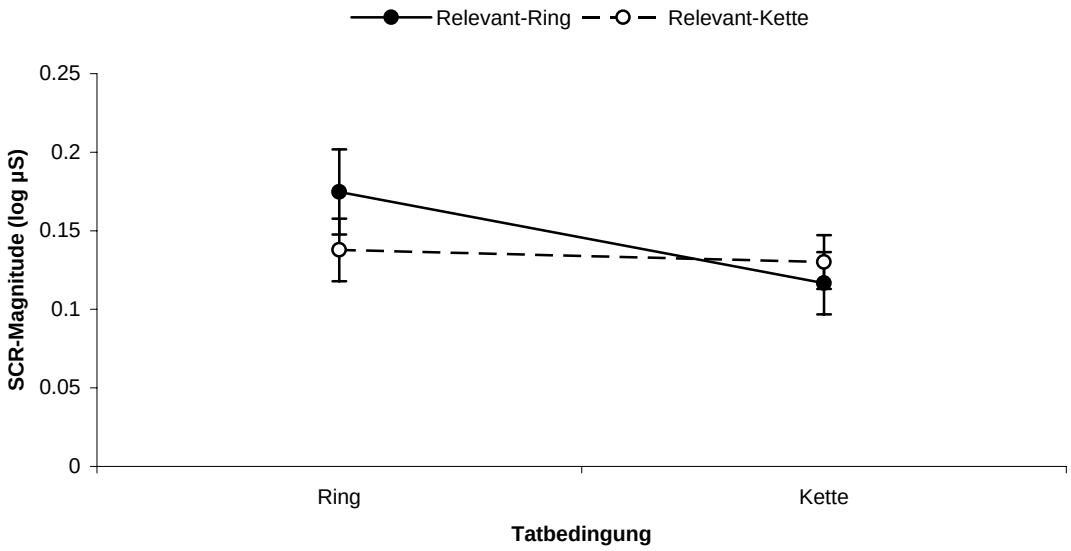


Abbildung 11. DLT – SCRs nach Ausblendung der Fragen: Vergleich der SCR-Magnituden der Fragentypen Relevant-Ring und -Kette, getrennt für die beiden Tatbedingungen; dargestellt sind Mittelwerte (Punkte) und deren Standardfehler (vertikale Linien).

Im Gegensatz zu den logarithmierten Werten erreichte bei der entsprechenden 2 × 2 × 5-ANOVA der **untransformierten SCR-Amplituden** keiner der o. g. Effekte das Signifikanzniveau. Dies galt sowohl für den Haupteffekt des Fragentyps, $F(4,144) = 2.02$, $p = .105$, $\varepsilon = .880$, $\eta^2 = .05$, als auch für die Interaktion aller drei Faktoren, $F(4,144) = 2.27$, $p = .073$, $\varepsilon = .880$, $\eta^2 = .06$. Indes fiel bei der 2 × 2 × 2-ANOVA die Wechselwirkung zwischen der Tatbedingung und dem Tatbezug der relevanten Fragen gleichsam wie bei den logarithmierten Daten signifikant aus, $F(1,36) = 5.34$, $p = .027$, $\eta^2 = .13$.

Im Hinblick auf den **GAT** erbrachte die 2 × 2 × 4- (Tatbedingung × EL × Itemtyp-) ANOVA der logarithmierten Amplitudenwerte signifikante Haupteffekte der EL, $F(1,36) = 22.99$, $p < .001$, $\eta^2 = .39$, und des Itemtyps, $F(3,108) = 4.60$, $p = .007$, $\varepsilon = .849$, $\eta^2 = .11$. Elektrodermal Labile ($M = 0.1384$ log µS, $SD = 0.0840$) reagierten insgesamt stärker als Stabile ($M = 0.0393$ log µS, $SD = 0.0352$). Die Mittelwerte der Itemtypen sind in Tabelle 19 abgedruckt und in Abbildung 12 veranschaulicht.

Tabelle 19. GAT – SCRs nach Ausblendung der Items: Mittelwerte (*M*) und Standardabweichungen (*SD*) der SCR-Magnituden (in log µS) der vier Itemtypen

	Itemtyp			
	Relevant-Täuschung	Relevant-Aufrichtigkeit	Irrelevant-Vergleich	Irrelevant-1
M	0.1074_a	0.0822_b	0.0799_b	0.0859_b
SD	0.0918	0.0798	0.0821	0.0944

Anmerkung. Mittelwerte mit unterschiedlichem Subskript unterscheiden sich auf dem 5%-Signifikanzniveau.

Einzelvergleiche (einfache zweiseitige t-Tests) ergaben, daß die Relevant-Täuschung-Items signifikant stärkere Reaktionen auslösten als die restlichen Itemtypen ($ts[39] > 2.16$, $ps < .037$). In allen anderen Paarvergleichen manifestierten sich keine statistisch bedeutsamen Unterschiede (jeweils $t < 1$).

Um die **Effekte der relevanten Items** genauer zu untersuchen, führte man eine 2 × 2 × 2-ANOVA mit den Gruppenfaktoren Tatbedingung (Ring vs. Kette) und EL (Stabil vs. Labil) und dem Meßwiederholungsfaktor Tatbezug der relevanten Items (Relevant-Ring vs. Relevant-Kette) durch. Es resultierte ein Haupteffekt der EL, $F(1,36) = 18.84$, $p < .001$, $\eta^2 = .34$. Labile ($M = 0.1423$ log µS, $SD = 0.0825$) zeigten bei den relevanten Items insgesamt eine höhere SCR-Magnitude als Stabile ($M = 0.0473$ log µS, $SD = 0.0479$). Die ebenfalls signifikante **disordinale Interaktion** zwischen der Tatbedingung und dem Meßwiederholungsfaktor, $F(1,36) = 9.05$, $p = .005$, $\eta^2 = .20$,

beruhte darauf, daß die Pbn stärker auf Items des eigenen Scheinverbrechens als auf Details der nicht begangenen Tat reagierten (vgl. Abbildung 13; Tatbedingung Ring: Relevant-Ring: M = 0.1055 log µS, SD = 0.0784, Relevant-Kette: M = 0.0807 log µS, SD = 0.0682; Tatbedingung Kette: Relevant-Ring: M = 0.0837 log µS, SD = 0.0917, Relevant-Kette: M = 0.1093 log µS, SD = 0.1055).

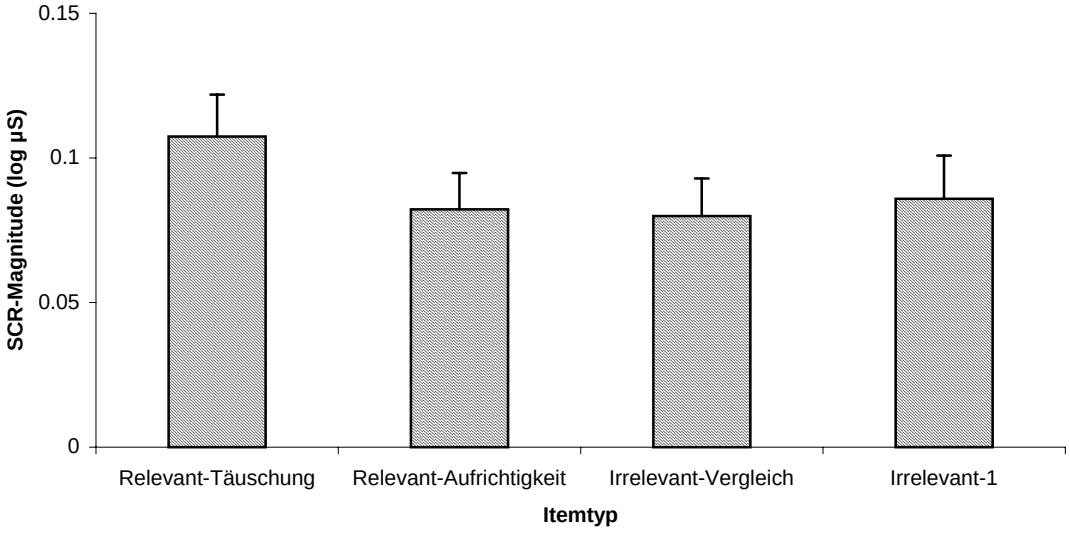


Abbildung 12. GAT – SCRs nach Ausblendung der Items: Vergleich der SCR-Magnituden der vier Itemtypen; dargestellt sind Mittelwerte (Säulen) und deren Standardfehler (aufgesetzte vertikale Linien).

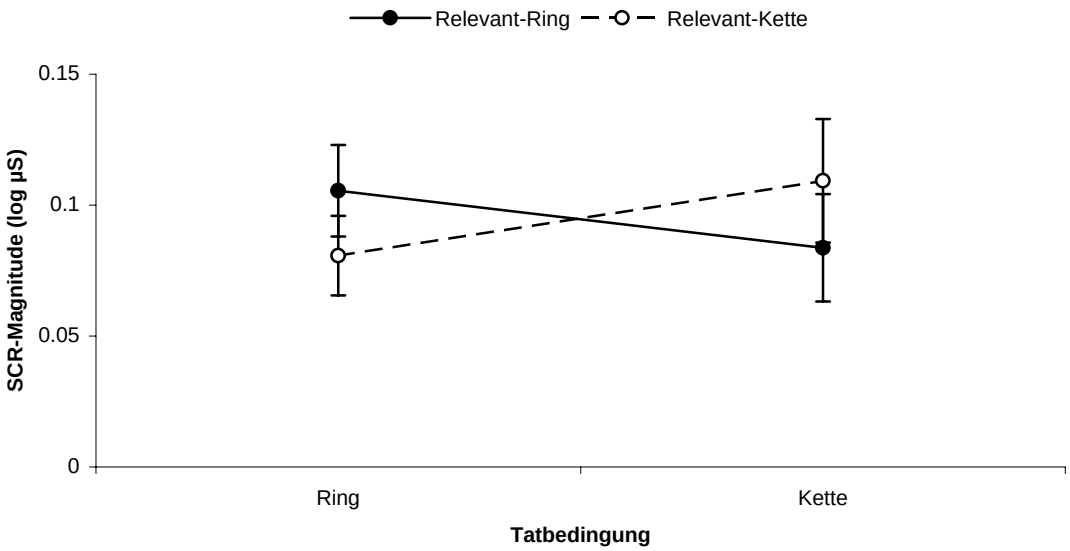


Abbildung 13. GAT – SCRs nach Ausblendung der Items: Vergleich der SCR-Magnituden der Itemtypen Relevant-Ring und -Kette, getrennt für die beiden Tatbedingungen; dargestellt sind Mittelwerte (Punkte) und deren Standardfehler (vertikale Linien).

Die 2 × 2 × 4- (Tatbedingung × EL × Itemtyp-) ANOVA der **untransformierten Amplitudenwerte** des GAT erbrachte vergleichbare Ergebnisse (Haupteffekt EL: $F[1,36] = 17.09$, $p < .001$, $\eta^2 = .32$; Haupteffekt Itemtyp: $F[3,108] = 4.03$, $p = .014$, $\varepsilon = .847$, $\eta^2 = .10$). Ebenso ergab die 2 × 2 × 2- (Tatbedingung × EL × Tatbezug-) ANOVA neben dem Labilitätseffekt, $F(1,36) = 13.96$, $p = .001$, $\eta^2 = .28$, eine signifikante Interaktion zwischen den Faktoren Tatbedingung und Tatbezug der relevanten Items, $F(1,36) = 8.15$, $p = .007$, $\eta^2 = .19$.

5.2 Herzschlagfrequenz (HR)

5.2.1 Verlaufsmorphologie der gemittelten HR-Reaktionen

Die **Auswertung der HR-Reaktionen** basierte auf deren mittleren Verläufen über die Sekunden hinweg. Dazu wurden die Daten der beiden Tests (DLT vs. GAT) separat einem „Averaging“ unterzogen. Die in Relation zur Prästimulus-Baseline bestimmten Differenzwerte der Herzschlagfrequenz (ΔHR) wurden – für die Poststimulus-Sekunden nach Ein- und Ausblendung getrennt – über die Trials der Fragen- bzw. Itemtypen und anschließend über die Pbn der jeweiligen Testbedingungen gemittelt. Die Ergebnisse dieser Prozedur (Grand Averages) finden sich in den Abbildungen 14 und 15. Darin sind ebenso wie bei den folgenden Darstellungen der ΔHR-Verläufe die Punkte in die Mitte der Sekundenintervalle gezeichnet, da ein HR-Wert auf der Echtzeitskala als ein gewichteter Durchschnitt der Schlagfrequenz im entsprechenden Zeitraum anzusehen ist (vgl. Velden & Wölk, 1987, S. 175).

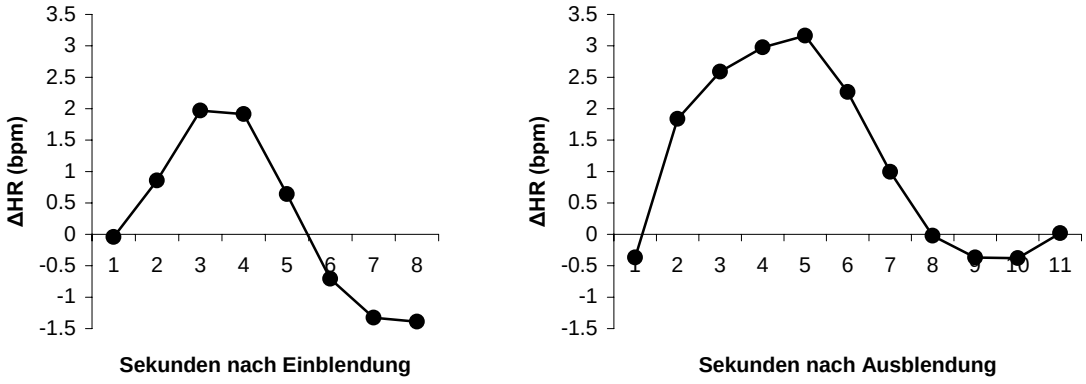


Abbildung 14. DLT: Grand Averages der HR-Reaktionen nach Ein- und Ausblendung der Stimuli auf dem Bildschirm.

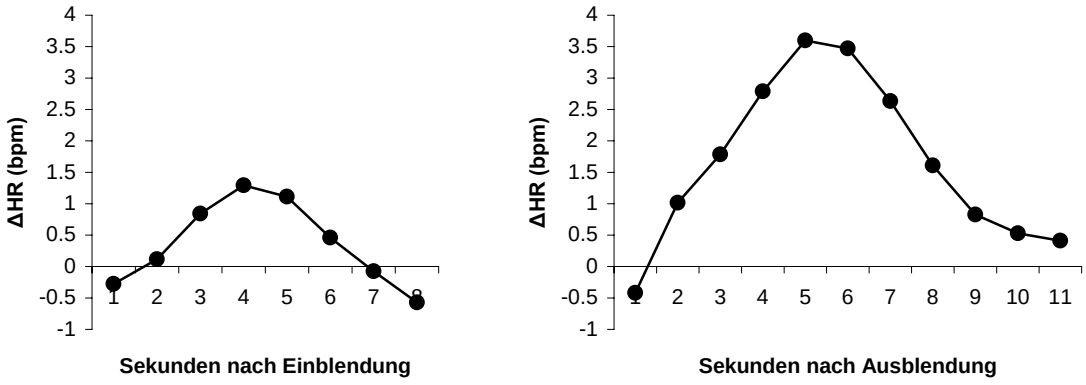


Abbildung 15. GAT: Grand Averages der HR-Reaktionen nach Ein- und Ausblendung der Stimuli auf dem Bildschirm.

Die mittleren Verläufe ließen mehrere Phasen bzw. **Komponenten der HR-Reaktionen** erkennen:

Für beide Tests zeigte sich in der ersten Poststimulus-Sekunde **nach Einblendung** eine initiale Abnahme (Dezeleration) der HR, die allerdings beim DLT nur sehr schwach ausgeprägt war. Daraufhin folgte ab Sekunde 2 ein Anstieg der Schlagfrequenz (Akzeleration) in Relation zur Baseline, der je nach Test bis zur Sekunde 5 (DLT) bzw. 6 (GAT) andauerte. Anschließend sank die HR wieder unter das Grundlinienniveau. Da die Ausblendung der Fragen bzw. Items im Bereich von 8 – 10 Sekunden nach Einblendung erfolgte, waren die HR-Reaktionen nur bis zur Poststimulus-Sekunde 8 sinnvoll auswertbar.

In der ersten Sekunde **nach Ausblendung** (imperativer Reiz der Antwortgabe) lag wiederum eine Dezeleration vor, gefolgt von einer akzelerativen Komponente. Letztere erstreckte sich beim DLT bis zur Poststimulus-Sekunde 7; der durchschnittliche HR-Verlauf erreichte in der Sekunde 8 in etwa die Grundlinie, sank dann leicht darunter, um anschließend wieder in Richtung Baseline zu konvergieren. Beim GAT hielt die Akzeleration länger an, was möglicherweise auch darauf zurückzuführen war, daß deren Maximum um ca. 0.5 bpm höher lag als das des DLT (vgl. die Abbildungen 14 und 15, jeweils Sekunde 5 nach Ausblendung). Ab der Sekunde 9 war allerdings beim GAT ein deutliches Abflachen der Kurve und ein allmähliches Zulaufen auf die Grundlinie erkennbar.

5.2.2 Spezifikation der Auswertung

Unter Berücksichtigung der Grand Averages legte man die **auszuwertenden Bereiche** der HR-Reaktionen jeweils auf das Intervall von Sekunde 1 – 8 nach Stimulusbeginn bzw. -ende fest. In bezug auf die Reaktionen im Anschluß an die Einblendung wurde das Zeitfenster dadurch begrenzt, daß bei einigen Trials bereits nach 8 Sekunden die Ausblendung erfolgte. Hinsichtlich der phasischen HR nach der Ausblendung deuteten die mittleren Verläufe an, daß ab der Poststimulus-Sekunde 9 im wesentlichen nur noch ein Ausklingen der Reaktionen zu beobachten war. Dies deckt sich auch mit der Feststellung von Schandry (1998, S. 144), wonach bei der Untersuchung reizbedingter Veränderungen der Herzschlagfrequenz in der Regel der Zeitraum bis ca. 15 Sekunden nach Beginn der Reizdarbietung (in diesem Fall also nach Einblendung) von besonderem Interesse ist. Abgesehen davon waren durch die Verwendung gleich großer Zeitfenster die Ergebnisse der beiden Tests und der Reaktionen nach Ein- und Ausblendung besonders gut gegenüberzustellen.

Für jede Vp wurden die ΔHR-Verläufe in den Poststimulus-Sekunden 1 – 8 nach Ein- bzw. Ausblendung über die Trials der fünf Fragen- bzw. vier Itemtypen gemittelt. Die Daten wurden dann für die beiden Testarten (DLT vs. GAT) und die beiden Zeitfenster getrennt anhand 4-facher **Varianzanalysen** mit den Gruppenfaktoren Tatbedingung und Elektrodermale Labilität sowie den Meßwiederholungsfaktoren Fragen- bzw. Itemtyp und Sekunden inferenzstatistisch ausgewertet.

5.2.3 HR-Reaktionen nach Einblendung der Fragen bzw. Items

Für den **DLT** ergab die 2 × 2 × 5 × 8- (Tatbedingung × EL × Fragentyp × Sekunden-) ANOVA signifikante Haupteffekte des Fragentyps, $F(4,144) = 7.58$, $p < .001$, $\varepsilon = .936$, $\eta^2 = .17$, und der Sekunden, $F(7,252) = 34.18$, $p < .001$, $\varepsilon = .267$, $\eta^2 = .49$, sowie eine Wechselwirkung zwischen den beiden Meßwiederholungsfaktoren, $F(28,1008) = 2.91$, $p = .004$, $\varepsilon = .286$, $\eta^2 = .08$, und eine Interaktion Tatbedingung × EL, $F(1,36) = 5.94$, $p = .020$, $\eta^2 = .14$. Abbildung 16 zeigt die durchschnittlichen ΔHR-Verläufe der fünf Fragentypen über die Sekunden hinweg. Der Sekundeneffekt wurde bereits anhand des Grand Averages im Abschnitt 5.2.1 veranschaulicht (Abbildung 14). Die Interaktion Fragentyp × Sekunden konnte man darauf zurückführen, daß die HR-Reaktionen der fünf Fragentypen nicht gänzlich parallel über die Sekunden verliefen, sondern sich speziell bei den Kontroll- und irrelevanten Fragen überschnitten (vgl. Abbildung 16). Von größerer Relevanz für die vorliegende Problemstellung war jedoch der Haupteffekt des Fragentyps (vgl. die über die Sekunden des Zeitfenster gemittelten ΔHR-Werte in Tabelle 20).

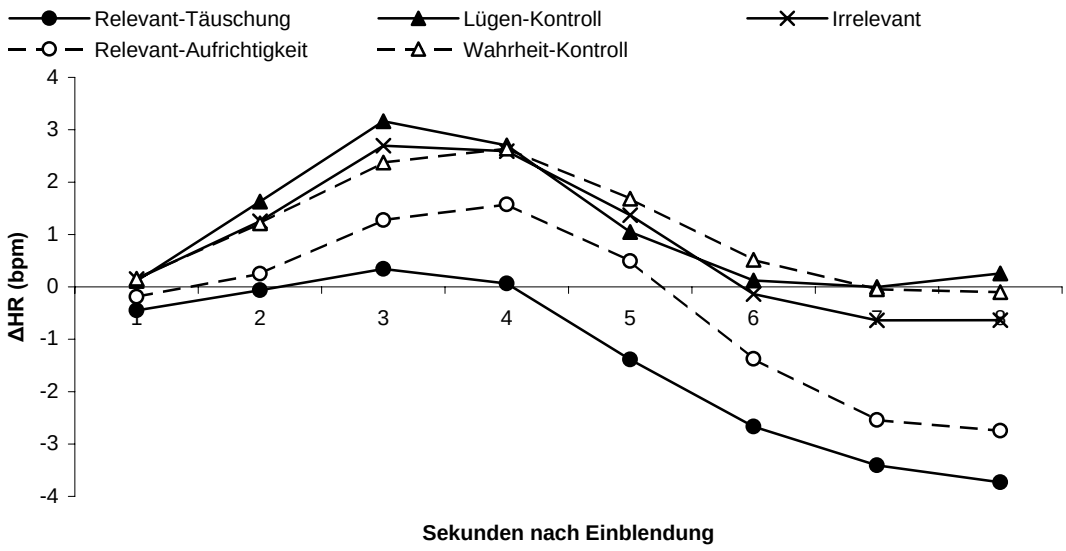


Abbildung 16. DLT – HR-Reaktionen nach Einblendung der Fragen: Vergleich der ΔHR-Verläufe der fünf Fragentypen.

Tabelle 20. DLT – HR-Reaktionen nach Einblendung der Fragen: Mittelwerte (*M* in bpm) und Standardabweichungen (*SD*) der fünf Fragentypen (gemittelt über die Sekunden 1 – 8)

	Fragentyp				
	Relevant-Täuschung	Relevant-Aufrichtigkeit	Lügen-Kontroll	Wahrheit-Kontroll	Irrelevant
M	-1.412_a	-0.408_a	1.128_b	1.055_b	0.831_b
SD	2.553	2.879	2.832	1.868	2.912

Anmerkung. Mittelwerte mit unterschiedlichem Subskript unterscheiden sich auf dem 5%-Signifikanzniveau.

Paarweise Vergleiche (zweiseitige t-Tests für abhängige Stichproben ohne Korrektur des Alpha-Fehlers) ergaben, daß sich die beiden Arten von relevanten Fragen jeweils signifikant von den anderen Reiztypen unterschieden (*t*s[39] > 2.16, *p*s < .037). Die Differenz der Bedingungen Relevant-Täuschung und Relevant-Aufrichtigkeit erreichte nicht das Signifikanzniveau, $t(39) = 1.96$, $p = .058$. Ebenso fielen die Vergleiche zwischen den restlichen drei Fragentypen nicht signifikant aus (jeweils $t < 1$). Insgesamt deuteten diese Befunde darauf hin, daß die tatbezogenen Stimuli schwächere Akzelerationen und stärkere Dezelerationen evozierten als die Kontroll- und irrelevanten Fragen (vgl. auch Abbildung 16).

Die disordinale **Interaktion der beiden Gruppenfaktoren** war darauf zurückzuführen, daß unter der Tatbedingung Ring der ΔHR-Mittelwert der elektrodermal Stabilen ($M = 0.895$ bpm, $SD = 1.129$) über dem der Labilen ($M = -0.148$ bpm, $SD = 1.106$) lag, wohingegen unter der Bedingung Kette ein inverses Verhältnis vorlag (Stabile: $M = -0.377$ bpm, $SD = 1.323$; Labile: $M = 0.585$ bpm, $SD = 1.585$).

Zur genaueren Untersuchung der **Effekte der relevanten Fragen** wurde eine 2 × 2 × 2 × 8-ANOVA mit den Faktoren Tatbedingung, EL, Tatbezug der relevanten Fragen und Sekunden gerechnet. Die erwartete Interaktion Tatbedingung × Tatbezug fiel zwar nur marginal signifikant aus, $F(1,36) = 3.61$, $p = .066$, $\eta^2 = .09$. Wie aber die Abbildungen 17 und 18 illustrieren, lagen – mit Ausnahme der Sekunde 2 unter der Bedingung Kette – in beiden Tatbedingungen die HR-Verläufe der Fragen nach dem durchgeführten Scheindiebstahl unterhalb der entsprechenden Kurve der Stimuli, die sich auf die nicht begangene Tat bezogen. D. h., die Ring-Gruppe reagierte im Durchschnitt mit einer niedrigeren ΔHR auf die Relevant-Ring-Fragen und die Kette-Gruppe entsprechend auf die Relevant-Kette-Fragen. Im übrigen resultierte aus dieser Varianzanalyse lediglich ein Sekundeneffekt, $F(7,252) = 25.16$, $p < .001$, $\varepsilon = .313$, $\eta^2 = .41$.

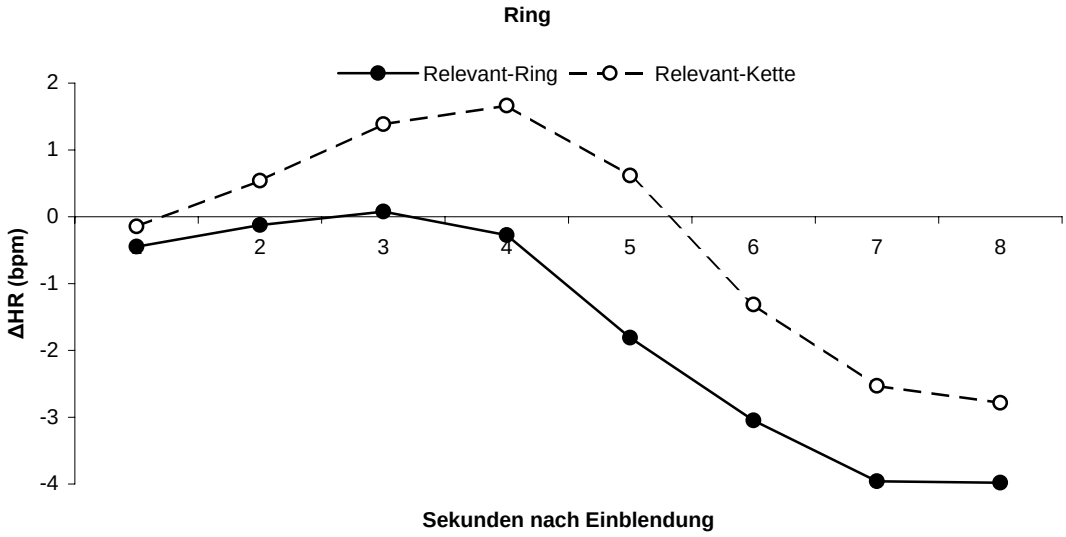


Abbildung 17. DLT – HR-Reaktionen nach Einblendung der Fragen: ΔHR-Verläufe der beiden Fragentypen Relevant-Ring und -Kette unter der Tatbedingung Ring.

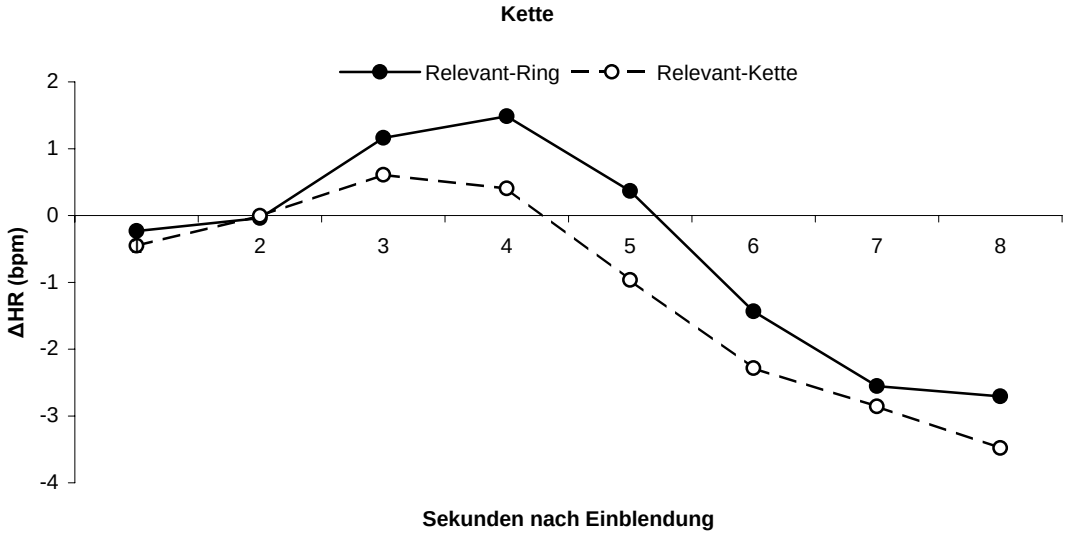


Abbildung 18. DLT – HR-Reaktionen nach Einblendung der Fragen: ΔHR-Verläufe der beiden Fragentypen Relevant-Ring und -Kette unter der Tatbedingung Kette.

Beim **GAT** erbrachte die 2 × 2 × 4 × 8- (Tatbedingung × EL × Itemtyp × Sekunden-) ANOVA nur statistisch bedeutsame Haupteffekte der Faktoren Itemtyp, $F(3,108) = 6.47$, $p = .002$, $\varepsilon = .757$, $\eta^2 = .15$, und Sekunden, $F(7,252) = 12.84$, $p < .001$, $\varepsilon = .341$, $\eta^2 = .26$. Der Sekundeneffekt wurde bereits in Abbildung 15 veranschaulicht. Abbildung 19 zeigt die ΔHR-Verläufe der vier Itemtypen. Die Mittelwerte der Itemtypen im entsprechenden Zeitintervall finden sich in Tabelle 21.

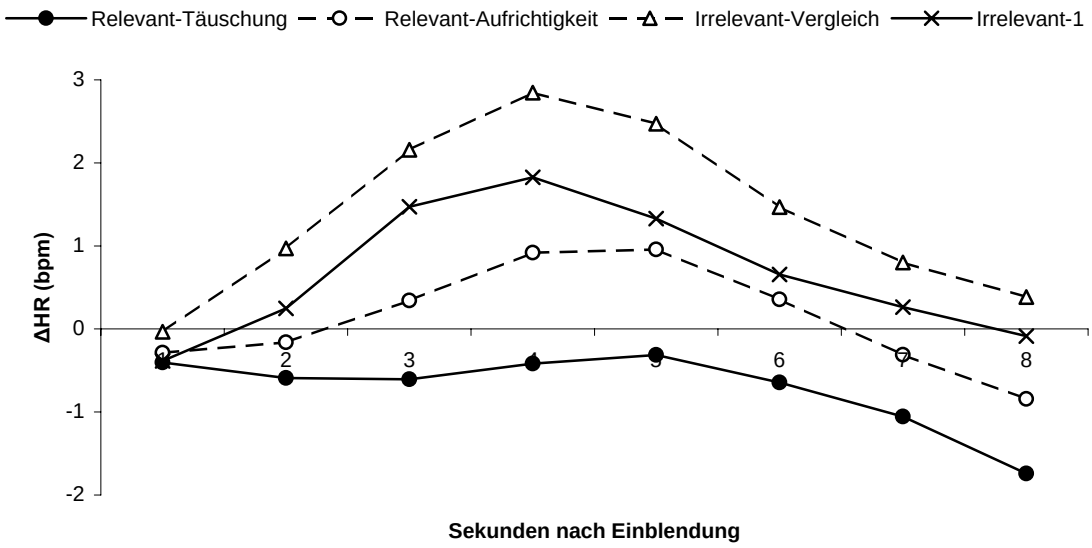


Abbildung 19. GAT – HR-Reaktionen nach Einblendung der Items: Vergleich der ΔHR-Verläufe der vier Itemtypen.

Tabelle 21. GAT – HR-Reaktionen nach Einblendung der Items: Mittelwerte (*M* in bpm) und Standardabweichungen (*SD*) der vier Itemtypen (gemittelt über die Sekunden 1 – 8)

	Itemtyp			
	Relevant-Täuschung	Relevant-Aufrichtigkeit	Irrelevant-Vergleich	Irrelevant-1
M	-0.722_{a}	$0.121_{a,b}$	1.383_{c}	$0.666_{b,c}$
SD	2.343	2.429	1.837	3.154

Anmerkung. Mittelwerte mit unterschiedlichem Subskript unterscheiden sich auf dem 5%-Signifikanzniveau.

Aus Abbildung 19 ist ersichtlich, daß der durchschnittliche ΔHR-Verlauf der Irrelevant-Vergleich-Items insgesamt am höchsten lag. Darunter verliefen jeweils nacheinander abgestuft die HR-Reaktionen der Bedingungen Irrelevant-1, Relevant-Aufrichtigkeit und Relevant-Täuschung. Allein in der ersten Poststimulus-Sekunde kam es zu leichten Überschneidungen der Reaktionen. Die Mittelwerte über die Sekunden 1 – 8 wurden jeweils anhand zweiseitiger t-Tests kontrastiert (vgl. Tabelle 21). Diese **Einzelvergleiche** ergaben signifikante Unterschiede zwischen den Relevant-Täuschung-Items und den beiden Arten von irrelevanten Items (*t*s[39] > 2.16, *p*s < .037). Die Mittelwerte der beiden relevanten Itemtypen unterschieden sich nicht statistisch bedeutsam voneinander, $t(39) = 1.67$, $p = .103$. Ebenso verfehlten die paarweisen Vergleiche der Bedingung Irrelevant-1 mit Relevant-Aufrichtigkeit ($t < 1$) bzw. mit Irrelevant-Vergleich ($t[39] = 1.51$, $p = .140$) das Signifikanzniveau. Ansonsten war nur die Diskrepanz zwi-

schen letzteren beiden Itemtypen überzufällig, $t(39) = 3.99$, $p < .001$. Diese Ergebnisse legten nahe, daß die tatbezogenen Items im Durchschnitt schwächere Akzelerationen und stärkere Dezelerationen auslösten als die irrelevanten Vergleichsstimuli, wobei die mittlere HR-Reaktion auf die Relevant-Täuschung-Items sogar während der gesamten 8 Sekunden unterhalb der Grundlinie blieb und somit rein dezelerativ ausfiel.

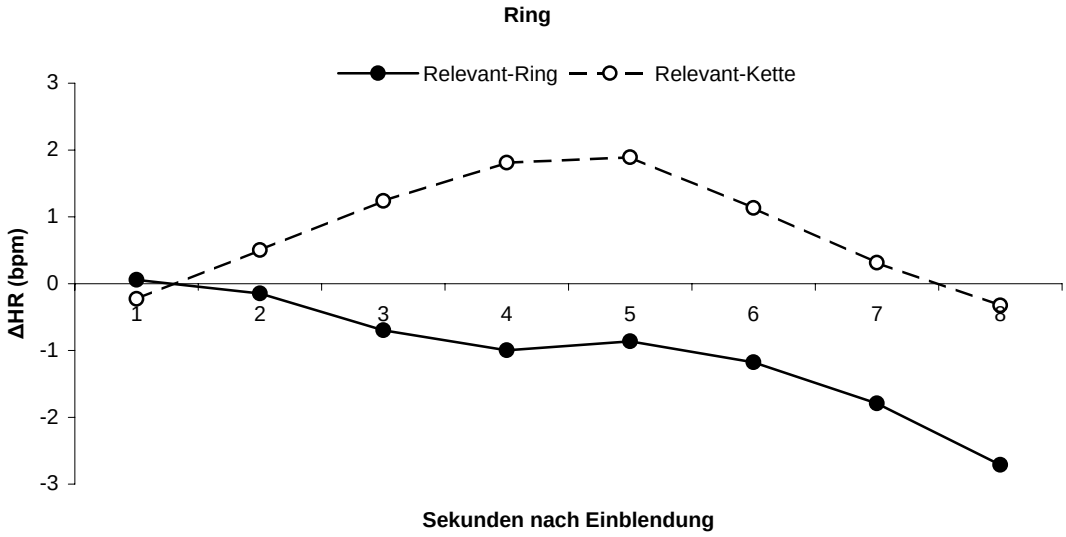


Abbildung 20. GAT – HR-Reaktionen nach Einblendung der Items: ΔHR-Verläufe der beiden Itemtypen Relevant-Ring und -Kette unter der Tatbedingung Ring.

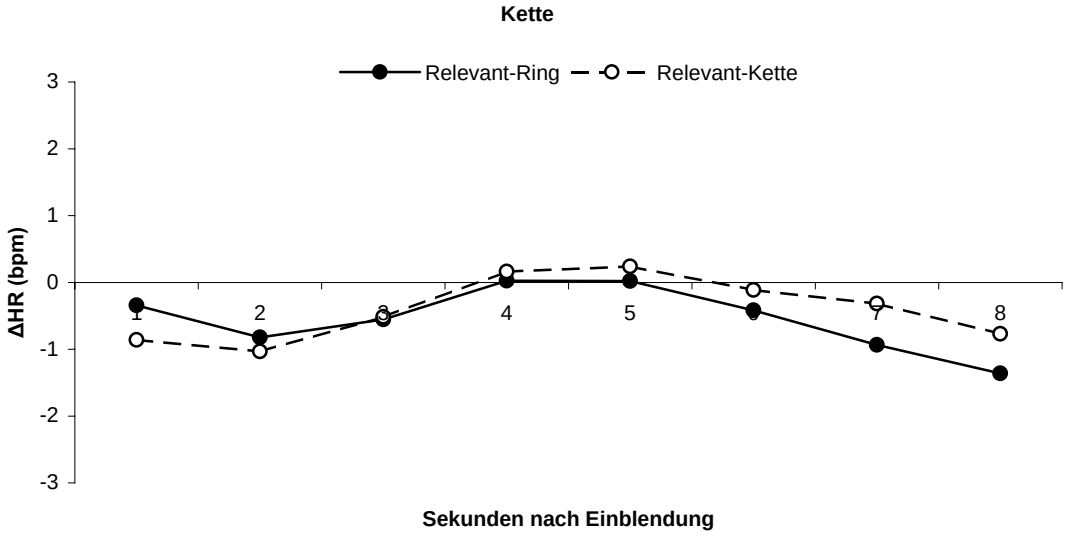


Abbildung 21. GAT – HR-Reaktionen nach Einblendung der Items: ΔHR-Verläufe der beiden Itemtypen Relevant-Ring und -Kette unter der Tatbedingung Kette.

Aus der 2 × 2 × 2 × 8-ANOVA (Tatbedingung × EL × Tatbezug der relevanten Items × Sekunden) zur näheren Analyse der **Effekte der relevanten Items** resultiert nur eine

statistische Tendenz ($p < .1$) zur Wechselwirkung Tatbedingung × Tatbezug, $F(1,36) = 3.16$, $p = .084$, $\eta^2 = .08$. Abgesehen von der ersten Sekunde reagierte die Ring-Gruppe insgesamt mit einer Dezeleration auf die Details ihrer Tat und weitgehend mit einer Akzeleration auf die Items des Kettendiebstahls (vgl. Abbildung 20). Die Kette-Gruppe hingegen zeigte kaum Unterschiede zwischen den beiden Itemtypen, wobei sich in den Sekunden 3 – 8 die mittlere HR-Reaktion der Bedingung Relevant-Kette sogar geringfügig oberhalb des Verlaufs von Relevant-Ring bewegte (vgl. Abbildung 21). Als signifikante Ergebnisse fand man nur Haupteffekte der Meßwiederholungsfaktoren Sekunden, $F(7,252) = 4.23$, $p = .011$, $\varepsilon = .364$, $\eta^2 = .11$, und Itemtyp, $F(1,36) = 4.35$, $p = .044$, $\eta^2 = .11$. Letzterer beruhte darauf, daß die ΔHR bei den Relevant-Kette-Items ($M = 0.195$ bpm, $SD = 2.106$) im Durchschnitt höher lag als bei den Relevant-Ring-Items ($M = -0.796$ bpm, $SD = 2.611$).

5.2.4 HR-Reaktionen nach Ausblendung der Fragen bzw. Items

Die 4-faktorielle Varianzanalyse (Tatbedingung × EL × Fragentyp × Sekunden) der HR-Reaktionen nach Ausblendung der Fragen des **DLT** auf dem Bildschirm ergab signifikante Haupteffekte des Fragentyps, $F(4,144) = 8.75$, $p < .001$, $\varepsilon = .753$, $\eta^2 = .20$, und der Sekunden, $F(7,252) = 19.67$, $p < .001$, $\varepsilon = .416$, $\eta^2 = .35$, sowie eine Interaktion Fragentyp × Sekunden, $F(28,1008) = 2.52$, $p = .011$, $\varepsilon = .293$, $\eta^2 = .07$, und eine Wechselwirkung zwischen den Gruppenfaktoren Tatbedingung und Elektrodermale Labilität, $F(1,36) = 7.73$, $p = .009$, $\eta^2 = .18$. Der Sekundeneffekt wurde bereits im Grand Average (Abschnitt 5.2.1, Abbildung 14) graphisch dargestellt. Die Interaktion der beiden Meßwiederholungsfaktoren konnte man darauf zurückführen, daß die ΔHR-Verläufe der fünf Fragentypen über die Sekunden nicht vollkommen parallel waren, sondern sich die Reaktionen der Bedingungen Relevant-Aufrichtigkeit, Lügen- und Wahrheit-Kontroll sowie Irrelevant überkreuzten (vgl. Abbildung 22). Allein der mittlere HR-Verlauf von Relevant-Täuschung lag durchweg unter dem der restlichen Fragen und hob sich somit deutlich davon ab.

Dieser Eindruck ist auch mit dem Haupteffekt des Fragentyps und den **paarweisen Vergleichen** der entsprechenden Mittelwerte vereinbar (siehe Tabelle 22). Der Mittelwert von Relevant-Täuschung unterschied sich signifikant von den anderen Fragentypen (ts[39] > 3.72, ps < .001). Ansonsten erreichte nur die Differenz zwischen Relevant-Aufrichtigkeit und Lügen-Kontroll das Signifikanzniveau, $t(39) = 2.11$, $p = .042$. Alle anderen Kontraste fielen nicht statistisch bedeutsam aus (ts[39] < 1.79, ps > .081).

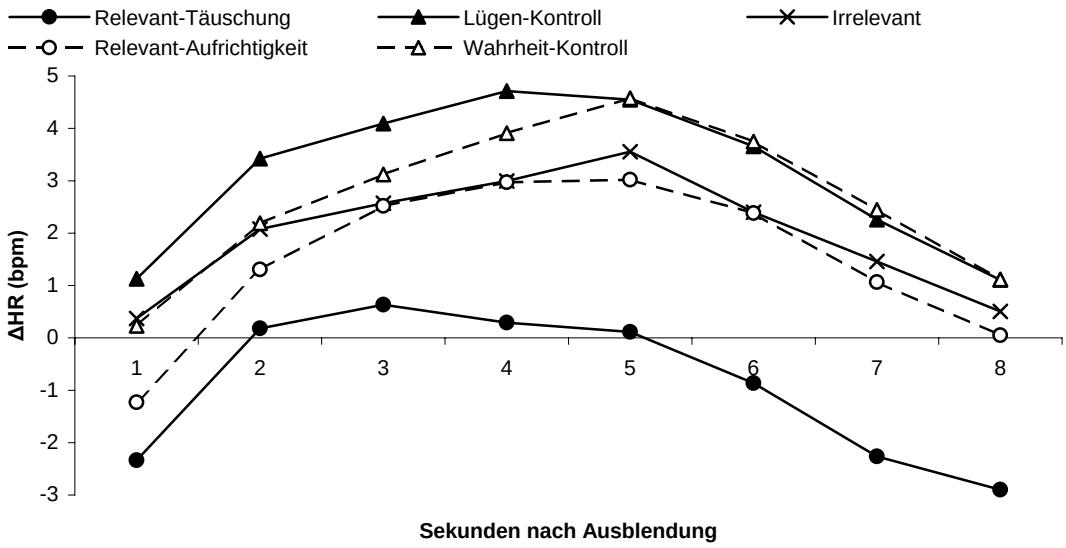


Abbildung 22. DLT – HR-Reaktionen nach Ausblendung der Fragen: Vergleich der ΔHR-Verläufe der fünf Fragentypen.

Tabelle 22. DLT – HR-Reaktionen nach Ausblendung der Fragen: Mittelwerte (*M* in bpm) und Standardabweichungen (*SD*) der fünf Fragentypen (gemittelt über die Sekunden 1 – 8)

	Fragentyp				
	Relevant-Täuschung	Relevant-Aufrichtigkeit	Lügen-Kontroll	Wahrheit-Kontroll	Irrelevant
M	-0.893_{a}	1.510_{b}	3.116_{c}	$2.667_{b,c}$	$1.992_{b,c}$
SD	3.433	3.304	3.052	3.726	3.638

Anmerkung. Mittelwerte mit unterschiedlichem Subskript unterscheiden sich auf dem 5%-Signifikanzniveau.

Die disordinale **Interaktion der beiden Gruppenfaktoren** war dadurch bedingt, daß unter der Tatbedingung Ring der ΔHR-Mittelwert der elektrodermal Stabilen (*M* = 2.788 bpm, *SD* = 1.445) über dem der Labilen (*M* = 1.075 bpm, *SD* = 1.014) lag, während sich unter der Bedingung Kette ein umgekehrtes Verhältnis zeigte (Stabile: *M* = 0.825 bpm, *SD* = 1.860; Labile: *M* = 2.025 bpm, *SD* = 2.097).

Die **Effekte des Tatbezugs der relevanten Fragen** wurden wiederum anhand einer 2 × 2 × 2 × 8- (Tatbedingung × EL × Tatbezug × Sekunden-) ANOVA analysiert. Die erwartete Interaktion Tatbedingung × Tatbezug war signifikant, $F(1,36) = 13.38$, $p = .001$, $\eta^2 = .27$. Darüber hinaus resultierte ein Sekundeneffekt, $F(7,252) = 15.50$, $p < .001$, $\varepsilon = .375$, $\eta^2 = .30$, und eine Wechselwirkung aller drei Faktoren, $F(7,252) = 6.96$, $p = .001$, $\varepsilon = .373$, $\eta^2 = .16$. In Abbildung 23 kann man erkennen, daß unter der

Tatbedingung Ring die HR-Reaktion auf die Relevant-Ring-Fragen unterhalb der Reaktion auf die Relevant-Kette-Fragen verliefen. In der Gruppe von Pbn, die die Kette entwendet hatten, manifestierte sich ein entsprechendes Muster mit einem niedrigeren ΔHR-Verlauf bei den Fragen nach dem Kettendiebstahl (vgl. Abbildung 24). Diese Ergebnisse deuteten also darauf hin, daß im Anschluß an die Fragen nach dem begangenen Scheinverbrechen mit schwächeren Akzelerationen bzw. stärkeren Dezelerationen reagiert wurde als bei den Fragen nach der nicht verübten Tat. Unter beiden Tatbedingungen verliefen die Reaktionen zwar übereinander, aber nicht parallel; sie drifteten teilweise auseinander, womit die signifikante 3-fache Interaktion zu erklären wäre.

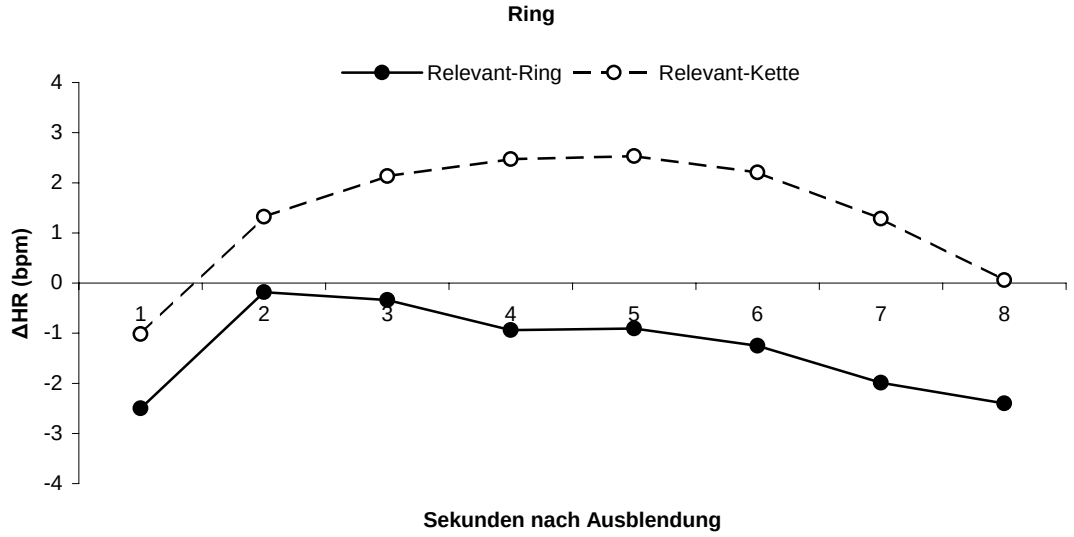


Abbildung 23. DLT – HR-Reaktionen nach Ausblendung der Fragen: ΔHR-Verläufe der beiden Fragentypen Relevant-Ring und -Kette unter der Tatbedingung Ring.

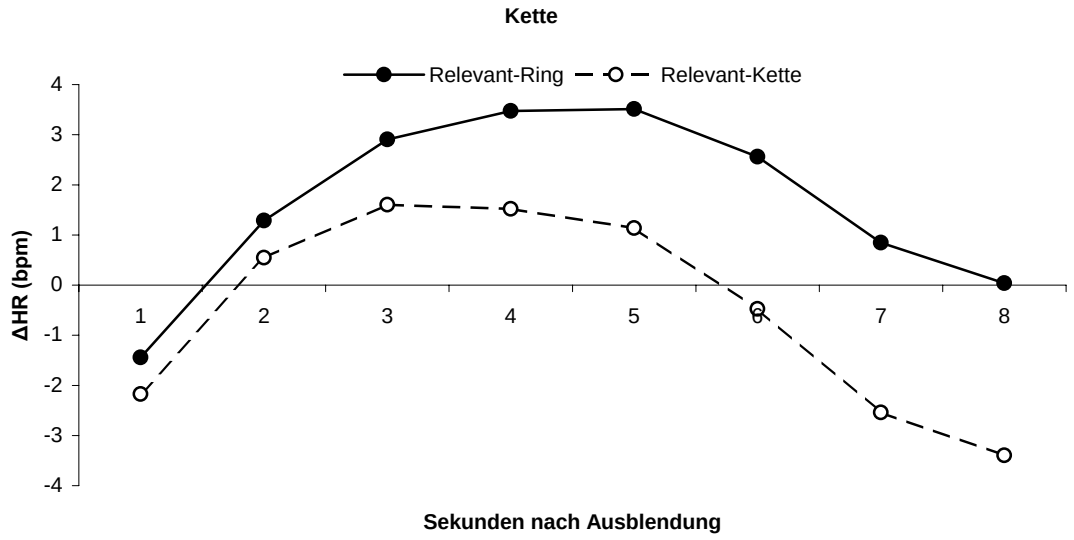


Abbildung 24. DLT – HR-Reaktionen nach Ausblendung der Fragen: ΔHR-Verläufe der beiden Fragentypen Relevant-Ring und -Kette unter der Tatbedingung Kette.

Beim **GAT** erbrachte die 2 × 2 × 4 × 8- (Tatbedingung × EL × Itemtyp × Sekunden-) ANOVA der phasischen HR nach Ausblendung der Items signifikante Haupteffekte der Faktoren Itemtyp, $F(3,108) = 12.26$, $p < .001$, $\varepsilon = .852$, $\eta^2 = .25$, und Sekunden, $F(7,252) = 28.16$, $p < .001$, $\varepsilon = .389$, $\eta^2 = .44$. Der Sekundeneffekt wurde bereits im Grand Average dargestellt (Abschnitt 5.2.1, Abbildung 15). Abbildung 25 veranschaulicht die ΔHR-Verläufe der vier Itemtypen. Die Mittelwerte der Itemtypen über die Sekunden 1 – 8 sind in Tabelle 23 abgedruckt.

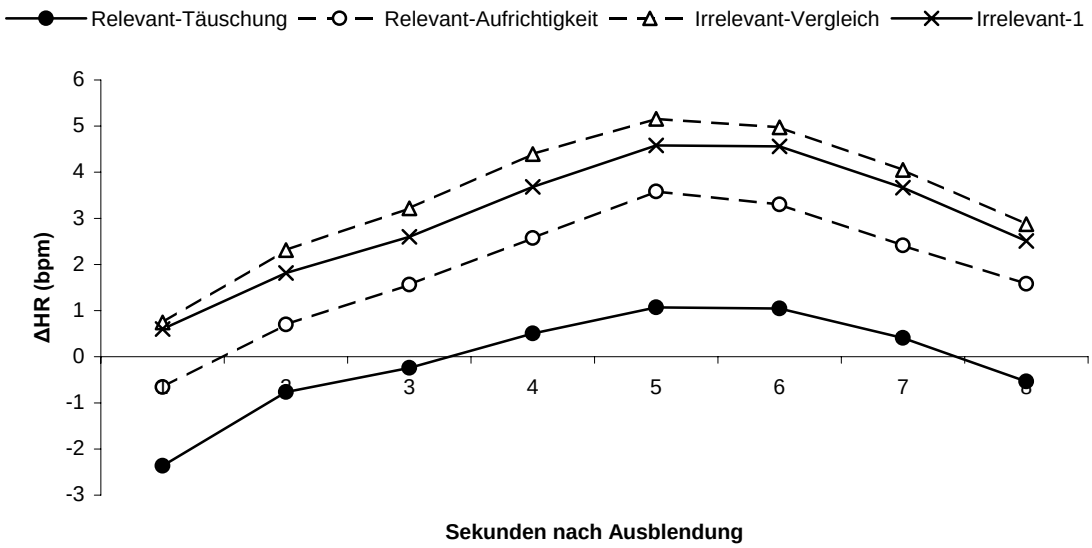


Abbildung 25. GAT – HR-Reaktionen nach Ausblendung der Items: Vergleich der ΔHR-Verläufe der vier Itemtypen.

Tabelle 23. GAT – HR-Reaktionen nach Ausblendung der Items: Mittelwerte (*M* in bpm) und Standardabweichungen (*SD*) der vier Itemtypen (gemittelt über die Sekunden 1 – 8)

	Itemtyp			
	Relevant-Täuschung	Relevant-Aufrichtigkeit	Irrelevant-Vergleich	Irrelevant-1
M	-0.110_a	1.881_b	3.465_c	$3.002_{b,c}$
SD	2.859	3.392	2.202	4.073

Anmerkung. Mittelwerte mit unterschiedlichem Subskript unterscheiden sich auf dem 5%-Signifikanzniveau.

Aus Abbildung 25 geht hervor, daß der durchschnittliche ΔHR-Verlauf der Relevant-Täuschung-Items am niedrigsten lag, und darüber die HR-Reaktionen der Bedingungen Relevant-Aufrichtigkeit, Irrelevant-1 und Irrelevant-Vergleich (vgl. Haupteffekt des Itemtyps). In den **Einzelvergleichen** der Mittelwerte über die Sekunden 1 – 8 unter-

schied sich die Relevant-Täuschung-Bedingung signifikant von allen anderen Itemtypen (*t*s[39] > 2.91, *p*s < .006). Des weiteren war nur noch die Diskrepanz zwischen Relevant-Aufrichtigkeit und Irrelevant-Vergleich statistisch bedeutsam, $t(39) = 3.45$, $p = .001$. Die paarweisen Gegenüberstellungen der Bedingung Irrelevant-1 mit Relevant-Aufrichtigkeit ($t[39] = 1.59$, $p = .120$) bzw. mit Irrelevant-Vergleich ($t < 1$) verfehlten das Signifikanzniveau. Diese Befunde sprachen zugunsten der Annahmen, daß die Relevant-Täuschung-Items im Durchschnitt eine schwächere Akzeleration bzw. stärkere Dezelerationen nach sich zogen als die anderen Itemtypen und daß die ΔHR der Relevant-Aufrichtigkeit-Items insgesamt unterhalb der Bedingung Irrelevant-Vergleich lag.

Im Hinblick auf die **Effekte der relevanten Items** ergab die 2 × 2 × 2 × 8- (Tatbedingung × EL × Tatbezug × Sekunden-) ANOVA eine signifikante Tatbedingung × Tatbezug-Interaktion, $F(1,36) = 8.67$, $p = .006$, $\eta^2 = .19$. Für beide Tatgruppen (Ring vs. Kette) verlief die durchschnittliche HR-Reaktion der Items, die das jeweils durchgeführte Scheinverbrechen thematisierten, unterhalb der mittleren Reaktion der Items, die sich auf die nicht verübte Tat bezogen. Unter der Tatbedingung Kette war allerdings die Diskrepanz geringer (vgl. die Abbildungen 26 und 27). Ferner resultierten ein signifikanter Sekundeneffekt, $F(7,252) = 16.79$, $p < .001$, $\varepsilon = .388$, $\eta^2 = .32$, und ein Labilitätseffekt, $F(1,36) = 5.45$, $p = .025$, $\eta^2 = .13$. Letzterer war darauf zurückzuführen, daß elektrodermal Stabile ($M = 1.692$ bpm, $SD = 2.466$) bei den Reaktionen nach Ausblendung der relevanten Items einen insgesamt höheren ΔHR-Mittelwert aufwiesen als Labile ($M = 0.079$ bpm, $SD = 1.776$).

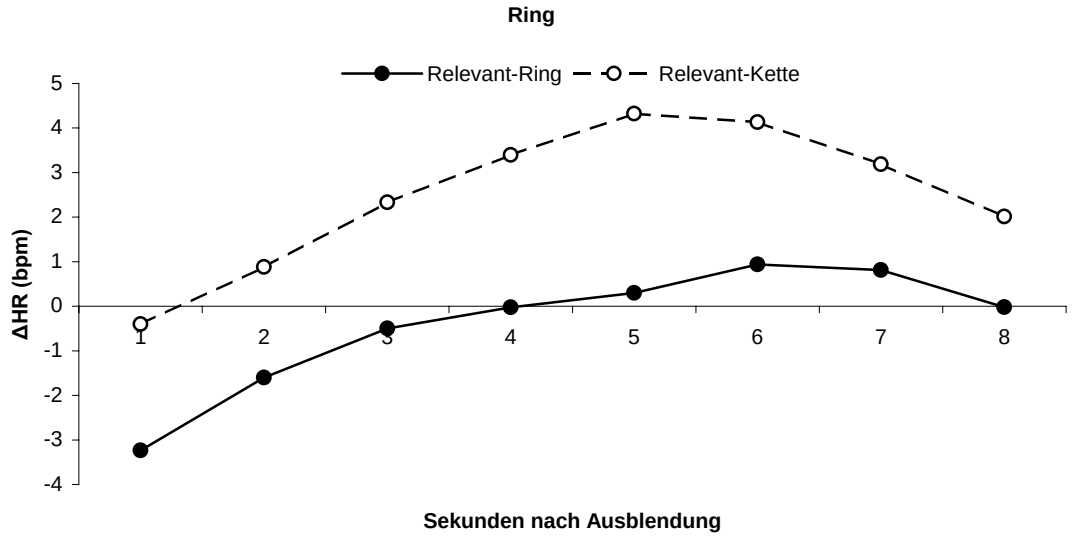


Abbildung 26. GAT – HR-Reaktionen nach Ausblendung der Items: ΔHR-Verläufe der beiden Itemtypen Relevant-Ring und -Kette unter der Tatbedingung Ring.

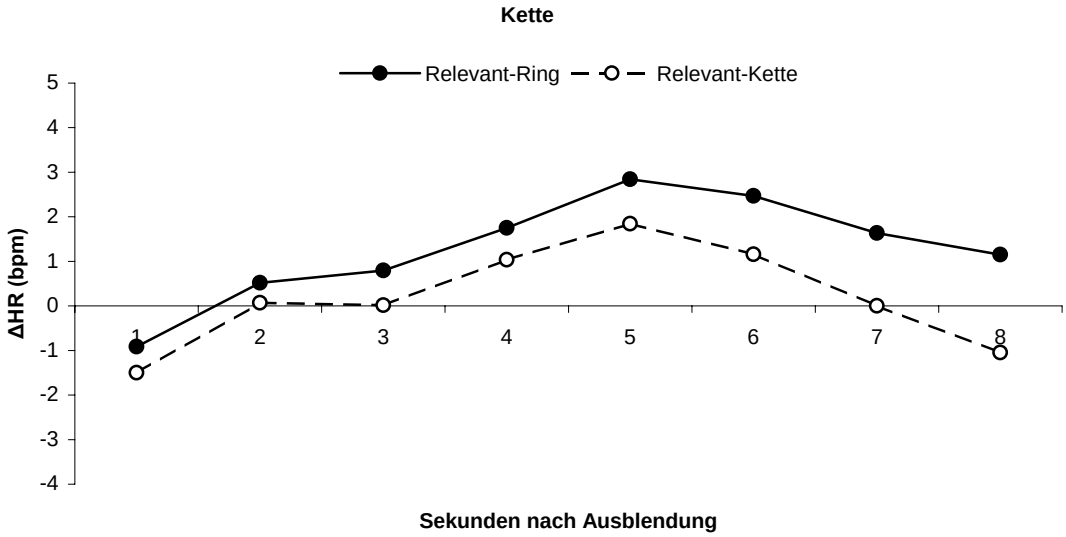


Abbildung 27. GAT – HR-Reaktionen nach Ausblendung der Items: ΔHR-Verläufe der beiden Itemtypen Relevant-Ring und -Kette unter der Tatbedingung Kette.

5.3 Erlebnisdeskriptive abhängige Variablen

Die **subjektiven Einschätzungen der Reaktionsstärke und der Bedeutsamkeit** der unterschiedlichen Fragen bzw. Items auf den siebenstufigen Ratingskalen wurden für jede Vp über die Trials der fünf Fragen- bzw. vier Itemtypen gemittelt. Diese Mittelwerte analysierte man nach Testart (DLT vs. GAT) getrennt anhand 3-faktorieller ANOVAs mit den Gruppenfaktoren Tatbedingung (Ring vs. Kette) und Elektrodermale Labilität (Stabil vs. Labil) und dem Meßwiederholungsfaktor Fragen- bzw. Itemtyp.

5.3.1 Subjektive Einschätzung der Reaktionsstärke

Beim **DLT** bestand das einzige signifikante Ergebnis der 3-fachen Varianzanalyse der Reaktionsratings in einem Haupteffekt des Fragentyps, $F(4,144) = 40.39$, $p < .001$, $\varepsilon = .703$, $\eta^2 = .53$. Die Mittelwerte der fünf Fragentypen wurden anhand einfacher zweiseitiger t-Tests gegenübergestellt (vgl. Tabelle 24 und Abbildung 28). Die Pbn stuften ihre Reaktionen unter der Bedingung Relevant-Täuschung stärker ein als bei allen anderen Fragentypen (*t*s[39] > 4.94, *p*s < .001). Zwischen Relevant-Aufrichtigkeit und den beiden Arten von Kontrollfragen bestanden keine signifikanten Unterschiede (stets $t < 1$). Diese drei Bedingungen hoben sich jedoch jeweils statistisch bedeutsam von den irrelevanten Fragen ab (*t*s[39] > 7.78, *p*s < .001), die nach Ansicht der Pbn die schwächsten Reaktionen ausgelöst hatten.

Tabelle 24. DLT – subjektive Einschätzung der Reaktionsstärke: Mittelwerte (*M*) und Standardabweichungen (*SD*) der fünf Fragentypen

	Fragentyp				
	Relevant-Täuschung	Relevant-Aufrichtigkeit	Lügen-Kontroll	Wahrheit-Kontroll	Irrelevant
M	4.30_a	3.43_b	3.34_b	3.33_b	2.22_c
SD	0.90	0.91	1.00	1.03	0.80

Anmerkung. Mittelwerte mit unterschiedlichem Subskript unterscheiden sich auf dem 5%-Signifikanzniveau.

Eine 2 × 2 × 2-ANOVA (Tatbedingung × EL × Tatbezug der relevanten Fragen) erbrachte eine **disordinale Wechselwirkung** der Faktoren Tatbedingung und Tatbezug, $F(1,36) = 36.62$, $p < .001$, $\eta^2 = .50$ (vgl. Abbildung 29). Die Pbn gaben an, stärker auf die Fragen nach dem begangenen Scheinverbrechen reagiert zu haben als auf die Fragen nach der anderen Tat (Tatbedingung Ring: Relevant-Ring: $M = 4.32$, $SD = 0.80$,

Relevant-Kette: M = 3.60, SD = 0.90; Tatbedingung Kette: Relevant-Ring: M = 3.27, SD = 0.92, Relevant-Kette: M = 4.28, SD = 1.02).

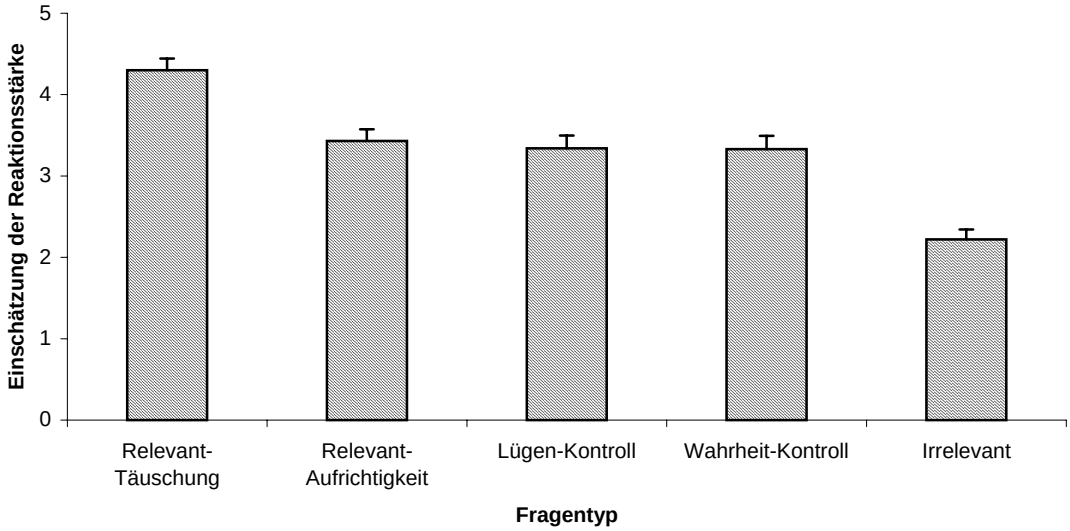


Abbildung 28. DLT – subjektive Einschätzung der Reaktionsstärke: Vergleich der Mittelwerte (Säulen) und Standardfehler (vertikale Linien) der fünf Fragentypen.

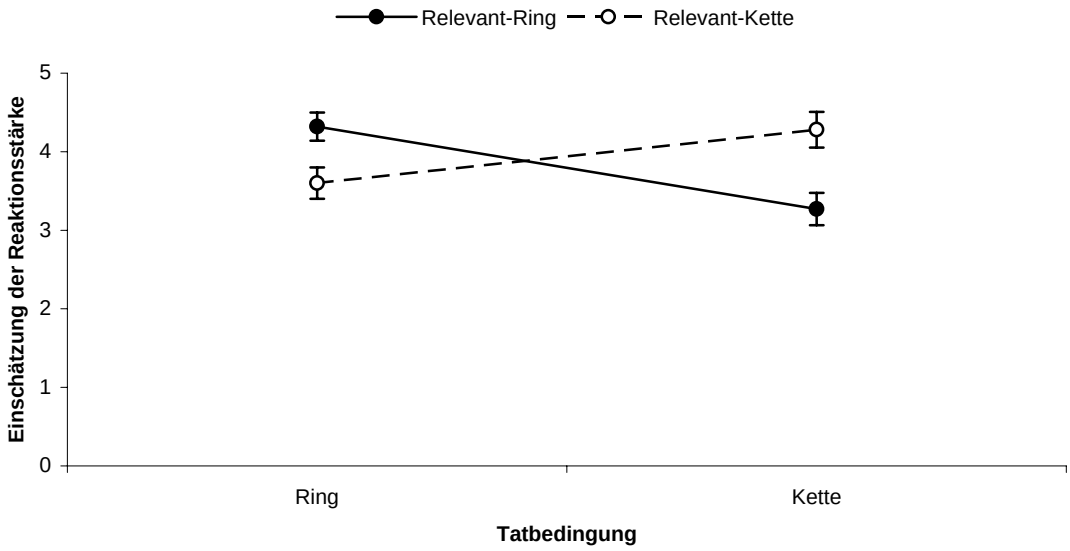


Abbildung 29. DLT – subjektive Einschätzung der Reaktionsstärke: Vergleich der Mittelwerte (Punkte) und Standardfehler (vertikale Linien) der Fragentypen Relevant-Ring und -Kette, getrennt für die beiden Tatbedingungen.

Für den **GAT** ergab die 2 × 2 × 4- (Tatbedingung × EL × Itemtyp-) ANOVA der Reaktionsratings einen signifikanten Haupteffekt des Itemtyps, $F(3,108) = 43.15$, $p < .001$, $\varepsilon = .525$, $\eta^2 = .55$, und eine Interaktion der beiden Gruppenfaktoren, $F(1,36) = 5.31$, $p = .027$, $\eta^2 = .13$. Der Haupteffekt ist in Tabelle 25 und in Abbildung 30 dargestellt.

Tabelle 25. GAT – subjektive Einschätzung der Reaktionsstärke: Mittelwerte (*M*) und Standardabweichungen (*SD*) der vier Itemtypen

	Itemtyp			
	Relevant-Täuschung	Relevant-Aufrichtigkeit	Irrelevant-Vergleich	Irrelevant-1
M	4.15_a	3.71_b	2.97_c	3.16_d
SD	0.86	0.80	0.85	0.88

Anmerkung. Mittelwerte mit unterschiedlichem Subskript unterscheiden sich auf dem 5%-Signifikanzniveau.

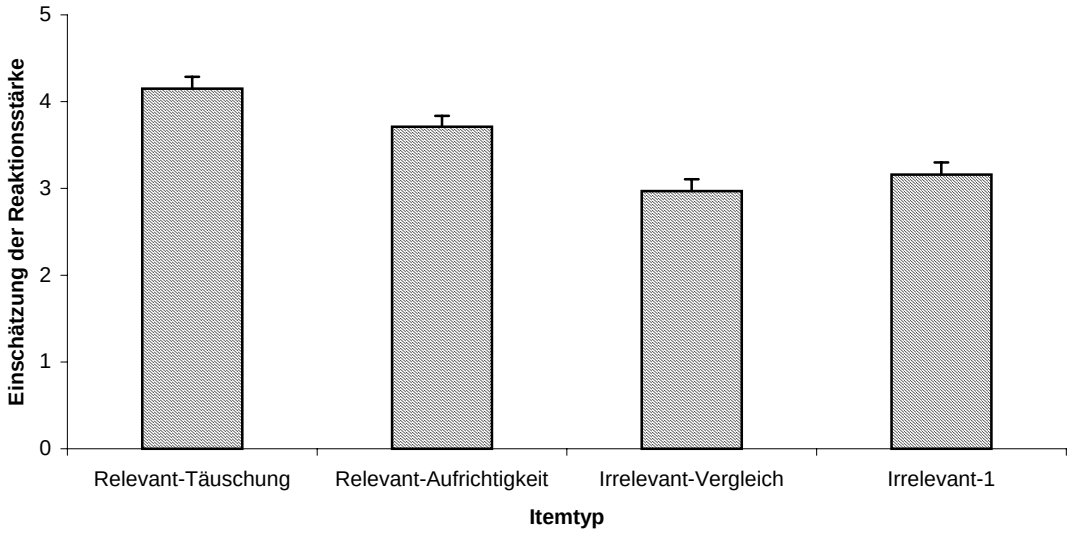


Abbildung 30. GAT – subjektive Einschätzung der Reaktionsstärke: Vergleich der Mittelwerte (Säulen) und Standardfehler (vertikale Linien) der vier Itemtypen.

Die **Paarvergleiche** zeigten, daß sich alle Itemtypen signifikant voneinander unterschieden (*t*s[39] > 3.61, *p*s < .001). Die Pbn berichteten im Durchschnitt die stärksten Reaktionen auf die Relevant-Täuschung-Items, gefolgt von Relevant-Aufrichtigkeit, Irrelevant-1 und schließlich den Irrelevant-Vergleich-Items. Die Interaktion Tatbedingung × EL (vgl. Abbildung 31) war darauf zurückzuführen, daß unter der Bedingung Kette elektrodermal Labile (*M* = 3.87, *SD* = 0.74) insgesamt höhere Reaktionsstärken angaben als Stabile (*M* = 3.08, *SD* = 0.54; *t*[18] = 2.73, *p* = .014). Unter der Bedingung Ring lag ein umgekehrtes Verhältnis vor (Stabile: *M* = 3.63, *SD* = 0.65; Labile: *M* = 3.42, *SD* = 0.79), wobei dieser Mittelwertsunterschied nicht signifikant ausfiel (*t* < 1).

Auch beim GAT resultierte aus der 2 × 2 × 2-ANOVA (Tatbedingung × EL × Tatbezug der relevanten Items) eine **disordinale Wechselwirkung** Tatbedingung × Tatbezug,

$F(1,36) = 19.20$, $p < .001$, $\eta^2 = .35$ (siehe Abbildung 32). Nach Einschätzung der Pbn hatten sie stärker auf Details des eigenen Scheinverbrechens reagiert als auf die Items der Tat, die sie nicht durchgeführt hatten (Tatbedingung Ring: Relevant-Ring: $M = 4.24$, $SD = 0.91$, Relevant-Kette: $M = 3.63$, $SD = 0.75$; Tatbedingung Kette: Relevant-Ring: $M = 3.79$, $SD = 0.85$, Relevant-Kette: $M = 4.07$, $SD = 0.82$).

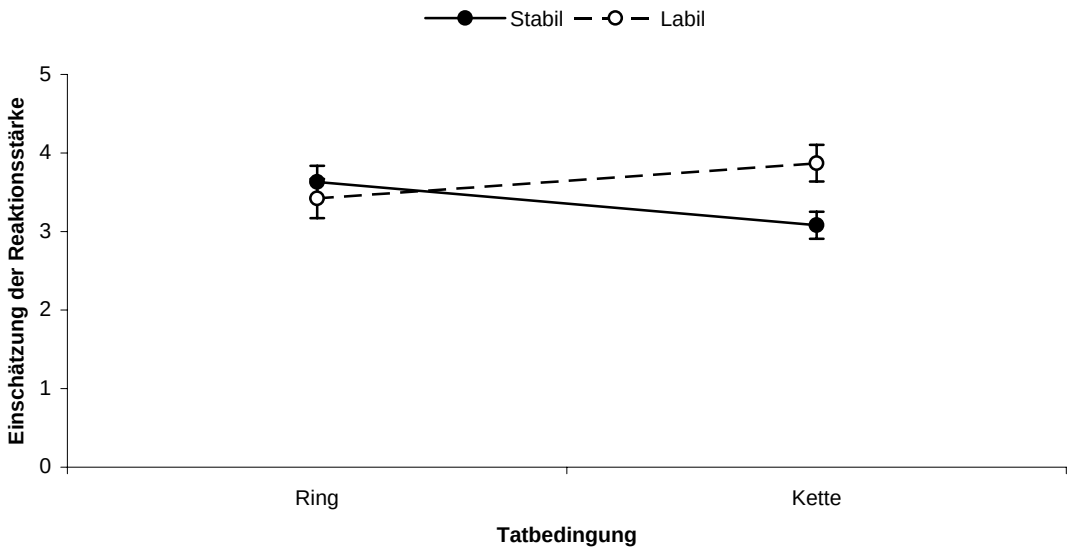


Abbildung 31. GAT – subjektive Einschätzung der Reaktionsstärke: Vergleich der Mittelwerte (Punkte) und Standardfehler (vertikale Linien) der elektrodermal stabilen und labilen Pbn, getrennt für die beiden Tatbedingungen.

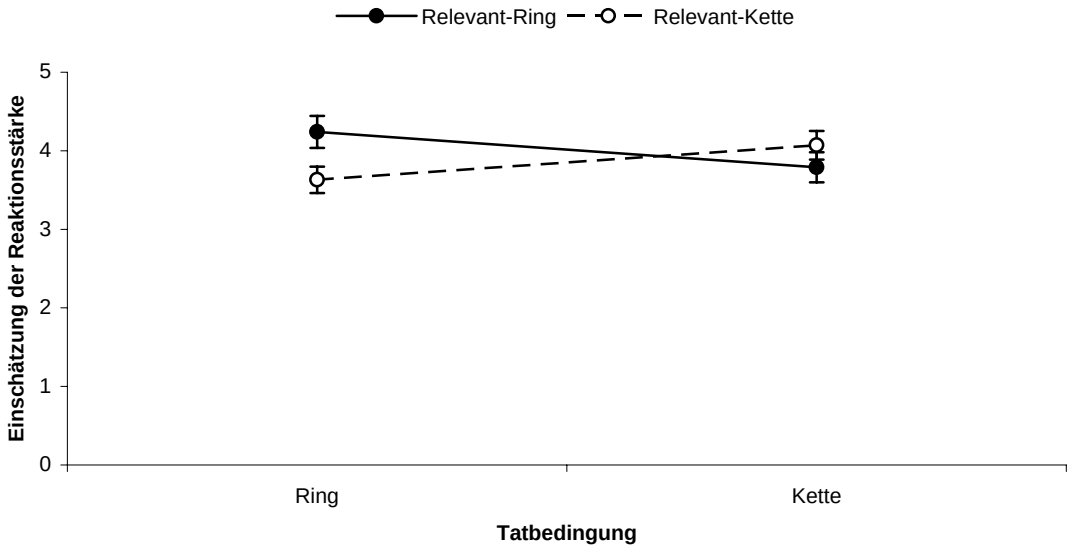


Abbildung 32. GAT – subjektive Einschätzung der Reaktionsstärke: Vergleich der Mittelwerte (Punkte) und Standardfehler (vertikale Linien) der Itemtypen Relevant-Ring und -Kette, getrennt für die beiden Tatbedingungen.

5.3.2 Subjektive Einschätzung der Bedeutsamkeit

Im Hinblick auf die Ratings der Relevanz der Fragen für das Ergebnis des **DLT** fand man bei der 2 × 2 × 5- (Tatbedingung × EL × Fragentyp-) ANOVA einen signifikanten Haupteffekt des Fragentyps, $F(4,144) = 26.20$, $p < .001$, $\varepsilon = .622$, $\eta^2 = .42$, und eine 3-fache Interaktion zwischen den beiden Gruppenfaktoren und dem Meßwiederholungsfaktor, $F(4,144) = 4.41$, $p = .010$, $\varepsilon = .622$, $\eta^2 = .11$. In Tabelle 26 sind die Mittelwerte und Standardabweichungen der fünf Fragentypen insgesamt und für die vier Gruppen getrennt abgedruckt.

Tabelle 26. DLT – subjektive Einschätzung der Bedeutsamkeit: Mittelwerte (*M*) und Standardabweichungen (*SD*) der Fragentypen, gesamt und getrennt für die vier Gruppen

		Fragentyp				
EL		Relevant-Täuschung	Relevant-Aufrichtigkeit	Lügen-Kontroll	Wahrheit-Kontroll	Irrelevant
				Tatbedingung Ring		
Stabil	*M*	5.90	5.40	4.17	4.50	3.57
	SD	1.01	0.87	1.30	1.10	1.46
Labil	*M*	6.10	5.30	3.87	3.97	2.83
	SD	1.10	1.44	1.34	1.35	1.74
				Tatbedingung Kette		
Stabil	*M*	5.80	5.17	3.90	3.87	2.77
	SD	1.19	1.21	1.85	1.79	1.39
Labil	*M*	5.37	4.73	4.83	4.67	5.00
	SD	1.09	1.25	1.29	1.40	1.62
Gesamt	*M*	5.79_a	5.15_b	4.19_c	4.25_c	3.54_d
	SD	1.09	1.19	1.46	1.42	1.75

Anmerkung. Die Gesamtmittelwerte der Fragentypen mit unterschiedlichem Subskript unterscheiden sich auf dem 5%-Signifikanzniveau.

In Abbildung 33 ist der Haupteffekt des Fragentyps veranschaulicht. Die paarweisen t-Tests zeigten, daß mit Ausnahme der beiden Arten von Kontrollfragen ($t < 1$) alle anderen **Mittelwertsvergleiche** signifikant ausfielen ($ts[39] > 2.13$, $ps < .04$). Im Durchschnitt schätzten die Pbn die Relevant-Täuschung-Fragen am wichtigsten für die Beurteilung ihrer Glaubwürdigkeit ein, gefolgt von der Bedingung Relevant-Aufrichtigkeit, den beiden Kontrollfragentypen und den irrelevanten Fragen (vgl. auch die Gesamtmittelwerte in Tabelle 26).

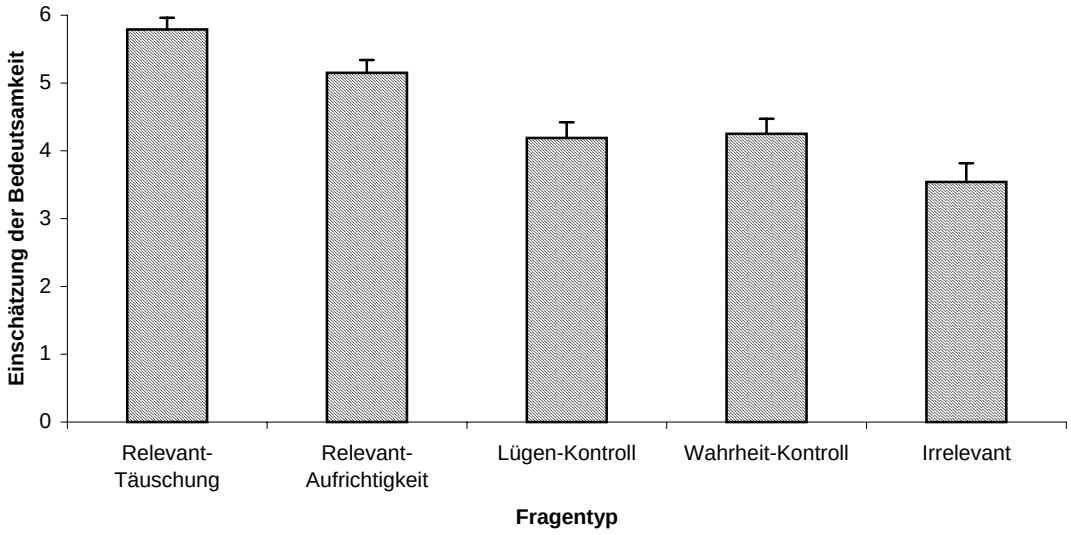


Abbildung 33. DLT – subjektive Einschätzung der Bedeutsamkeit: Vergleich der Mittelwerte (Säulen) und Standardfehler (vertikale Linien) der fünf Fragentypen.

Der Haupteffekt des Fragentyps muß allerdings wegen der signifikanten **3-fachen Interaktion relativiert** werden. Aus den Abbildungen 34 und 35 geht hervor, daß sich der entsprechende Verlauf der Bedeutsamkeitsratings über die Fragentypen hinweg im wesentlichen nur unter der Tatbedingung Ring und bei den elektrodermal stabilen Pbn des Treatments Kette widerspiegelte, wohingegen die elektrodermal Labilen der Kette-Bedingung ein abweichendes Muster erkennen ließen. Bei dieser Gruppe fiel insbesondere die hohe Bedeutsamkeitseinstufung der irrelevanten Fragen auf, die sogar oberhalb der Relevant-Aufrichtigkeit-Fragen lag. Im Rahmen einer reduzierten $2 \times 2 \times 4$-(Tatbedingung × EL × Fragentyp-) ANOVA ohne die Bedingung Irrelevant war die 3-fache Interaktion nicht mehr signifikant, $F(3,108) = 3.17$, $p = .060$, $\varepsilon = .536$, $\eta^2 = .08$, während die Wechselwirkung in allen anderen reduzierten $2 \times 2 \times 4$-ANOVAs unter Ausschluß der einzelnen Fragentypen stets das Signifikanzkriterium erreichte. Diese Befunde deuteten darauf hin, daß die Interaktion der Initialanalyse ($2 \times 2 \times 5$) v. a. auf die hohe Einschätzung der Relevanz der irrelevanten Fragen durch die elektrodermal Labilen der Tatbedingung Kette zurückzuführen war.

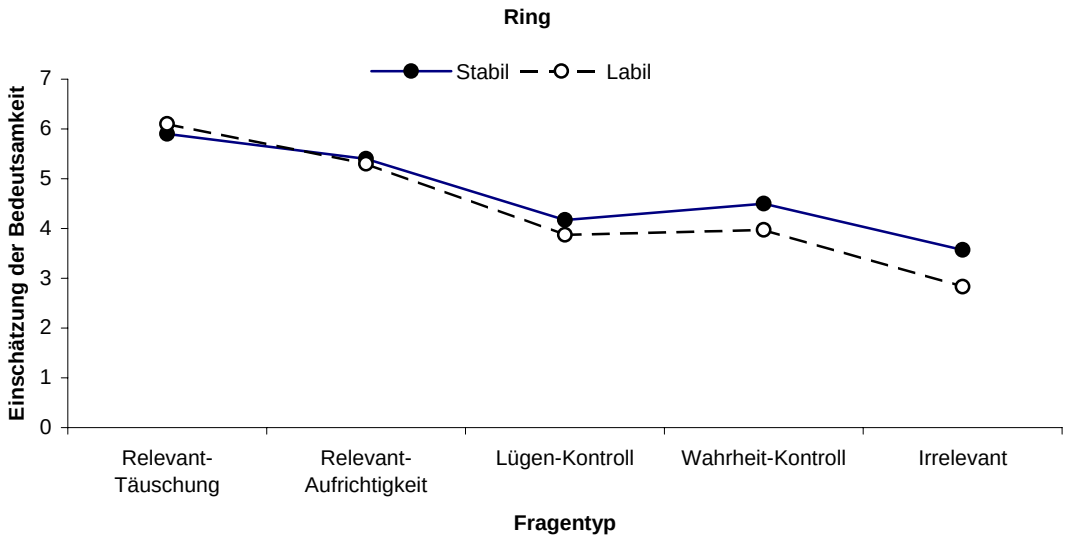


Abbildung 34. DLT – subjektive Einschätzung der Bedeutsamkeit: Vergleich der Mittelwerte der fünf Fragentypen, getrennt für elektrodermal stabile und labile Pbn der Tatbedingung Ring.

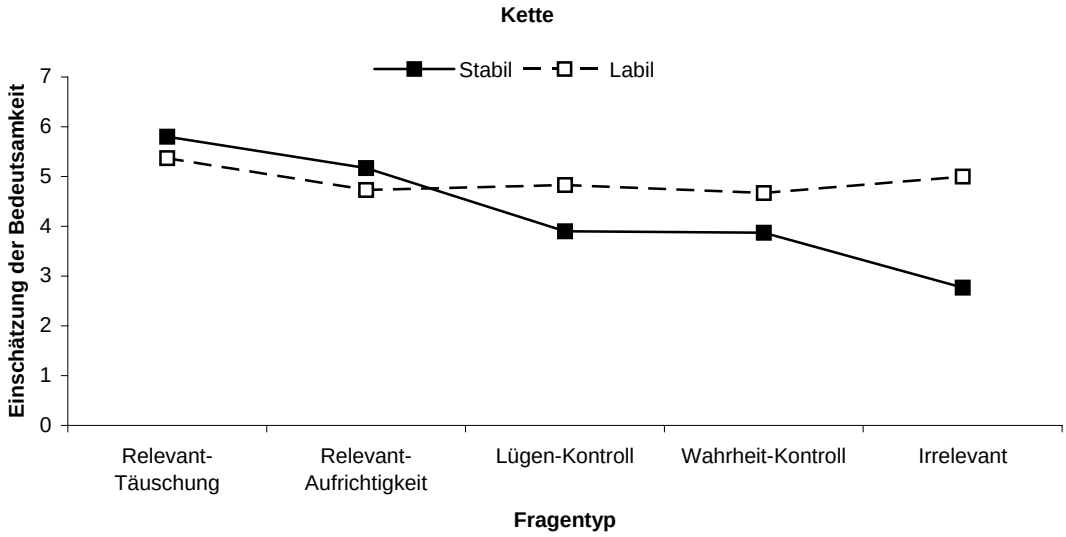


Abbildung 35. DLT – subjektive Einschätzung der Bedeutsamkeit: Vergleich der Mittelwerte der fünf Fragentypen, getrennt für elektrodermal stabile und labile Pbn der Tatbedingung Kette.

Die 2 × 2 × 2-ANOVA (Tatbedingung × EL × Tatbezug der relevanten Fragen) ergab wiederum nur die erwartete **Interaktion** zwischen Tatbedingung und Tatbezug, $F(1,36) = 19.15$, $p < .001$, $\eta^2 = .35$ (vgl. Abbildung 36; Tatbedingung Ring: Relevant-Ring: $M = 6.00$, $SD = 1.03$, Relevant-Kette: $M = 5.35$, $SD = 1.16$; Tatbedingung Kette: Relevant-Ring: $M = 4.95$, $SD = 1.22$, Relevant-Kette: $M = 5.58$, $SD = 1.13$).

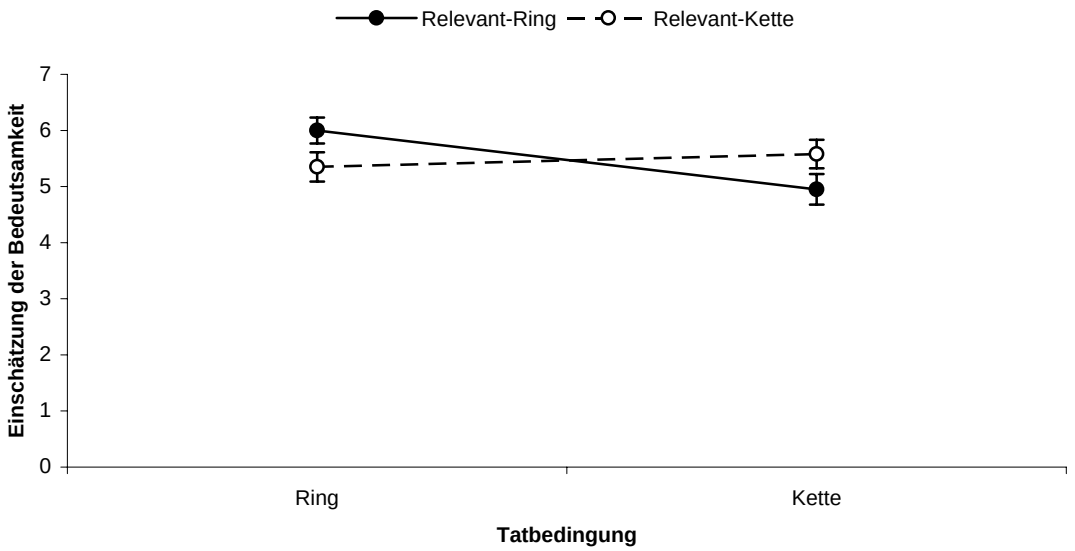


Abbildung 36. DLT – subjektive Einschätzung der Bedeutsamkeit: Vergleich der Mittelwerte (Punkte) und Standardfehler (vertikale Linien) der Fragentypen Relevant-Ring und -Kette, getrennt für die beiden Tatbedingungen.

Das einzige signifikante Ergebnis der 2 × 2 × 4- (Tatbedingung × EL × Itemtyp-) ANOVA zum Bedeutsamkeitsrating des **GAT** war der Haupteffekt des Itemtyps, $F(3,108) = 44.33$, $p < .001$, $\varepsilon = .732$, $\eta^2 = .55$. Bis auf den Unterschied zwischen den beiden Arten von irrelevanten Items ($t < 1$) fielen alle paarigen Mittelwertsvergleiche signifikant aus (vgl. Tabelle 27 und Abbildung 37). Die Relevant-Täuschung-Items wurden wichtiger eingestuft als ihre Pendants der Bedingung Relevant-Aufrichtigkeit ($t[39] = 3.06$, $p = .004$), und die beiden tatbezogenen Itemtypen waren wiederum jeweils subjektiv bedeutsamer als die zwei Arten von irrelevanten Items ($ts[39] > 5.83$, $ps < .001$).

Tabelle 27. GAT – subjektive Einschätzung der Bedeutsamkeit: Mittelwerte (*M*) und Standardabweichungen (*SD*) der vier Itemtypen

	Itemtyp			
	Relevant-Täuschung	Relevant-Aufrichtigkeit	Irrelevant-Vergleich	Irrelevant-1
M	5.37_a	4.65_b	3.30_c	3.32_c
SD	1.32	1.54	1.16	1.26

Anmerkung. Mittelwerte mit unterschiedlichem Subskript unterscheiden sich auf dem 5%-Signifikanzniveau.

Die genauere Untersuchung der Effekte der relevanten Items mittels der 2 × 2 × 2-ANOVA erbrachte eine signifikante **disordinale Interaktion** zwischen den Faktoren

Tatbedingung und Tatbezug, $F(1,36) = 9.50$, $p = .004$, $\eta^2 = .21$. Wie in Abbildung 38 ersichtlich, wurden die Items des begangenen Scheindiebstahl jeweils als wichtiger für die Glaubwürdigkeitsbeurteilung eingeschätzt als die Details der nicht verübten Tat (Tatbedingung Ring: Relevant-Ring: $M = 5.73$, $SD = 1.37$, Relevant-Kette: $M = 4.68$, $SD = 1.80$; Tatbedingung Kette: Relevant-Ring: $M = 4.61$, $SD = 1.23$, Relevant-Kette: $M = 5.01$, $SD = 1.18$).

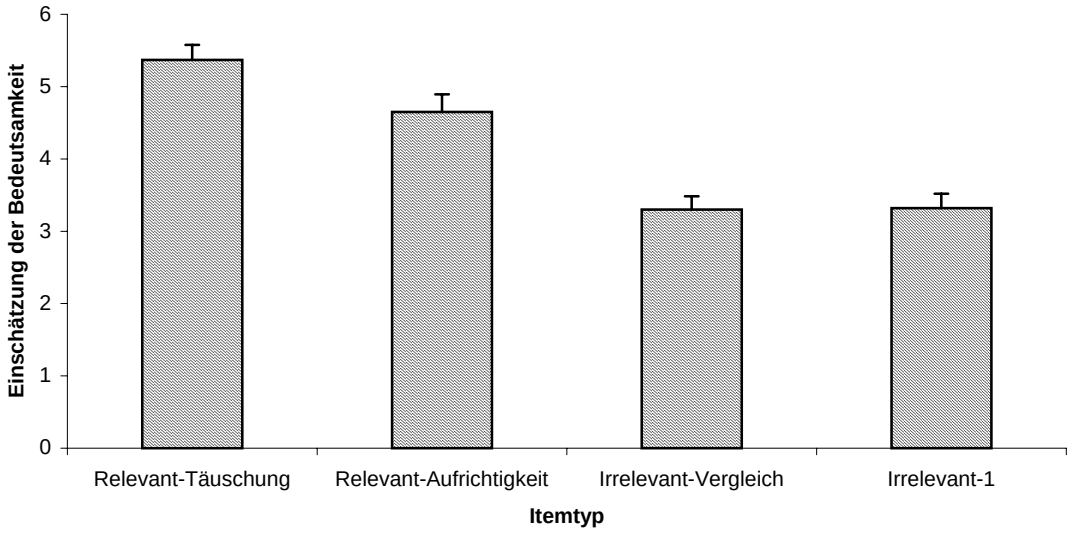


Abbildung 37. GAT – subjektive Einschätzung der Bedeutsamkeit: Vergleich der Mittelwerte (Säulen) und Standardfehler (vertikale Linien) der vier Itemtypen.

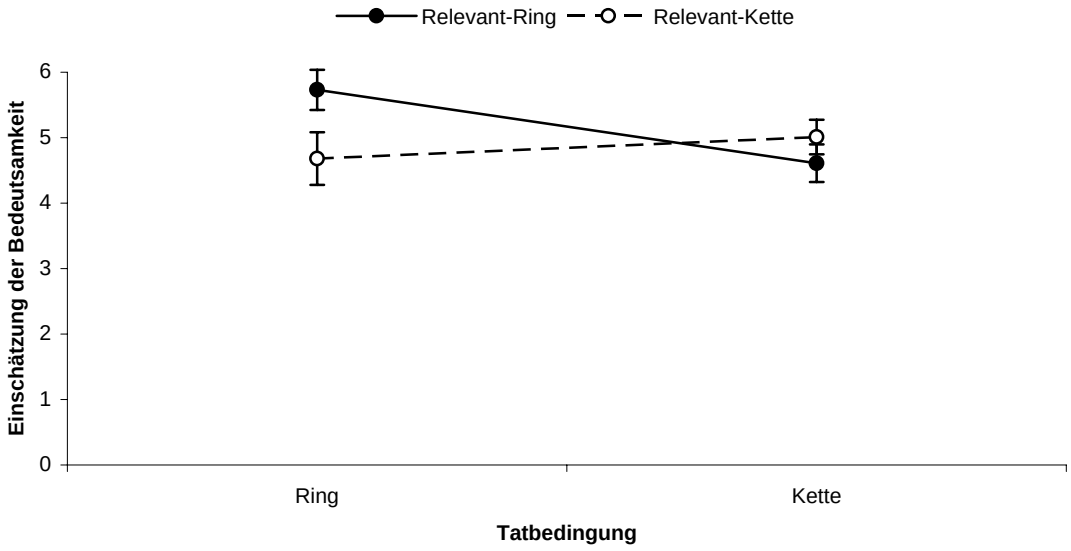


Abbildung 38. GAT – subjektive Einschätzung der Bedeutsamkeit: Vergleich der Mittelwerte (Punkte) und Standardfehler (vertikale Linien) der Itemtypen Relevant-Ring und -Kette, getrennt für die beiden Tatbedingungen.

5.4 Kontrollvariablen

5.4.1 Postexperimentelle Ratings und Nachbefragung

Die auf den siebenstufigen Skalen abgegebenen Einschätzungen der **Täuschungsmotivation** und der **Treffsicherheit** der „Lügendetektion“ wurden varianzanalytisch ausgewertet. Die Häufigkeiten der dichotomen Antworten („ja“ vs. „nein“) auf die Fragen nach eventuellen **Manipulationsversuchen** während der Untersuchung sowie nach **differentiellen Empfindungen** beim Lügen und Wahrheitsagen untersuchte man inferenzstatistisch mit Hilfe von Chi-Quadrat-Tests.

5.4.1.1 Täuschungsmotivation

Um etwaige Unterschiede zwischen den Versuchsgruppen hinsichtlich der **Täuschungsmotivation** zu überprüfen, wurden die entsprechenden Ratings einer 3-fachen ANOVA mit den Gruppenfaktoren Testart (DLT vs. GAT), Tatbedingung (Ring vs. Kette) und Elektrodermale Labilität (Stabil vs. Labil) unterzogen. Es zeigten sich keine statistisch bedeutsamen Haupt- oder Wechselwirkungseffekte. Die von den Vpn abgegebenen Scores betrugen im Durchschnitt $M = 4.80$ ($SD = 1.37$). Dies entsprach annähernd der Skalenausprägung: *(5) Ich war deutlich motiviert, dem Untersucher die Lügendetektion zu erschweren.*

5.4.1.2 Subjektive Treffsicherheit

Der Zusammenhang zwischen der Versuchsgruppenzugehörigkeit und den Ratings der **Treffsicherheit** wurde ebenfalls mittels einer 3-fachen ANOVA analysiert (s. o.). Auch hier resultierten weder signifikante Haupteffekte noch Interaktionen. Der Gesamtmittelwert belief sich auf $M = 4.54$ ($SD = 1.46$) und lag somit zwischen den beiden Skalenausprägungen: *Meiner Einschätzung nach ist es dem Untersucher (4) wiederholt bzw. (5) öfters gelungen, wahrheitsgemäße Antworten und Lügen zu unterscheiden.*

5.4.1.3 Manipulationsversuche

In der Nachbefragung berichteten 66 von 80 Pbn (ca. 83%), sie hätten während der Untersuchung eine **Strategie, Taktik oder Technik** angewendet, um dem Untersucher die „Lügendetektion“ zu erschweren. Zur Überprüfung, ob sich deren Häufigkeit even-

tuell ungleichmäßig über die Versuchsgruppen verteilte, wurden zweiseitige Vier-Felder-Chi-Quadrat-Tests mit (Yates-) Kontinuitätskorrektur getrennt für die Variablen Testart, Tatbedingung und EL gerechnet. Es zeigten sich keine signifikanten Unterschiede zwischen den jeweiligen Gruppen hinsichtlich der Häufigkeit selbstberichteter Manipulationsversuche (stets $\chi^2[1, N = 80] < 1$).

Auf Nachfrage gaben die betreffenden Pbn sowohl körperliche als auch mentale Beeinflussungsversuche an. Viele schilderten, sie hätten sich bemüht, bei den unterschiedlichen Testfragen bzw. Items möglichst ähnliche physiologische Zustände bzw. Reaktionen herbeizuführen (Unterdrücken oder Evozieren körperlicher Veränderungen, z. B. indem sie versuchten, beim Lügen ruhig zu bleiben oder beim Wahrheitsagen Erregung zu provozieren). Als Strategien wurden unter anderem eine Kontrolle der Atmung, Entspannungstechniken oder ein selbstinduziertes Erschrecken genannt. Häufig seien auch gedankliche Ablenkungsmanöver durchgeführt worden, wie z. B. nicht auf die (v. a. relevanten) Stimuli achten, an andere Dinge denken oder etwa die Autosuggestion, keinen bzw. einen anderen Gegenstand entwendet zu haben.

5.4.1.4 Differentielle Empfindungen

Die Frage, ob sie beim Lügen und Wahrheitsagen unterschiedliche **Empfindungen, Gefühle, körperliche Reaktionen** o. ä. verspürt hatten, bejahten etwa 79% der Vpn. Um zu klären, inwiefern sich die Gruppen DLT vs. GAT, Ring vs. Kette und Stabil vs. Labil diesbezüglich unterschieden, wurden wiederum drei Chi-Quadrat-Tests (s. o.) durchgeführt. Auch hier fand man keine statistisch bedeutsamen Zusammenhänge (jeweils $\chi^2[1, N = 80] < 1$).

In der Regel wurde von **körperlichen Reaktionsunterschieden** berichtet, die v. a. im Vergleich zwischen den wahrheitswidrig verneinten, tatbezogenen Reizen (Relevant-Täuschung) und den anderen Fragen- bzw. Itemtypen aufgetreten seien. Darunter waren relativ detaillierte Schilderungen über Veränderungen der Herz-, Atem- oder Schwitzaktivität (meist eine höhere Herz- bzw. Pulsfrequenz, Stocken bzw. Veränderungen der Atmung oder Schwitzen in der Hand). Aber auch weniger spezifische Aussagen, wie etwa eine „stärkere Anspannung“, „Wärmegefühle“ oder ein „Kribbeln“ beim Lügen, wurden getroffen.

5.4.1.5 Versuchserleben

Innerhalb der offenen Nachbefragung, die auf das **gesamte Versuchserleben** abzielte, wurde das Experiment von den meisten Pbn positiv bewertet. Sofern Beanstandungen geäußert wurden, bezogen sie sich überwiegend auf die per Instruktion vorgeschriebene, möglichst unbewegte Sitzposition, die auf Dauer als unbequem empfunden wurde und z. B. zu Taubheitsgefühlen in den Händen oder zu Nackenbeschwerden geführt habe. In diesem Zusammenhang hätten sich einige Vpn kurze Bewegungspausen oder eine zusätzliche Kopfstütze gewünscht.

5.4.2 Gedächtnistests

Insgesamt konnten die Pbn in etwa gleich viele Einzelheiten des begangenen und nicht begangenen Scheinverbrechens **erinnern** (eigene Tat: $M = 5.14$, $SD = 0.78$; andere Tat: $M = 5.05$, $SD = 0.78$; $t[79] = 1.07$, $p = .29$) bzw. **wiedererkennen** (eigene Tat: $M = 5.33$, $SD = 0.76$; andere Tat: $M = 5.26$, $SD = 0.74$; $t < 1$). Berücksichtigte man jedoch zusätzlich die Testart, dann ergaben sich Diskrepanzen. Beim **DLT** lagen wiederum keine signifikanten Unterschiede vor (Erinnern: eigene Tat: $M = 5.10$, $SD = 0.84$; andere Tat: $M = 5.20$, $SD = 0.76$; $t < 1$; Wiedererkennen: eigene Tat: $M = 5.30$, $SD = 0.85$; andere Tat: $M = 5.43$, $SD = 0.75$; $t[39] = 1.53$, $p = .13$). Im Gegensatz dazu konnten jene Pbn, die den **GAT** absolviert hatten, mehr Details der begangenen Tat erinnern und wiedererkennen als Einzelheiten des anderen Scheinverbrechens (Erinnern: $M = 5.18$, $SD = 0.71$ vs. $M = 4.90$, $SD = 0.78$; $t[39] = 2.72$, $p = .010$; Wiedererkennen: $M = 5.35$, $SD = 0.66$ vs. $M = 5.10$, $SD = 0.71$; $t[39] = 2.69$, $p = .011$). Die Mittelwertsunterschiede fielen zwar eher klein aus, die Effekte waren jedoch relativ stark und statistisch bedeutsam (Effektgrößen der o. g. signifikanten t-Tests für abhängige Stichproben: Erinnern: $d = 0.61$; Wiedererkennen: $d = 0.60$; vgl. Bortz & Döring, 1995, S. 569f.).

Um zu überprüfen, ob diese Abweichungen in einem **Zusammenhang mit den Reaktionsdifferenzen** zwischen den Bedingungen Relevant-Täuschung und Relevant-Aufrichtigkeit standen, wurden sie miteinander korreliert. Da beim GAT das Wiedererkennen der Items entscheidend ist, beschränkte man die Analyse auf die Ergebnisse des entsprechenden Gedächtnistests. Für jeden Pb der GAT-Gruppe bestimmte man den Differenzwert: Anzahl der wiedererkannten Relevant-Täuschung-Items minus Anzahl der wiedererkannten Relevant-Aufrichtigkeit-Items. Ein positiver Wert indizierte, daß der Pb eher die tatbezogenen Details des von ihm begangenen Scheinverbrechens identifizieren konnte. Die Differenzwerte korrelierte man jeweils mit den mittleren Reak-

tionsunterschieden zwischen den beiden relevanten Itemtypen ($M_{\text{Relevant-Täuschung}}$ - $M_{\text{Relevant-Aufrichtigkeit}}$; positive Werte bedeuteten im Durchschnitt höhere Reaktionen auf die begangene Tat). Da die Differenzwerte der Anzahl wiedererkannter Items nicht annähernd normalverteilt waren (zweiseitiger Kolmogorov-Smirnov-Anpassungstest auf Normalverteilung: $Z = 2.46$, $p < .001$), wurden non-parametrische Rangkorrelationen berechnet (Kendalls Tau unter Berücksichtigung von Rangbindungen; vgl. auch Bortz, Lienert & Boehnke, 1990, S. 422ff.). Diese Prozedur erfolgte für die elektrodermalen (logarithmierte Amplituden des SCRs mit Latenz von 1 – 3 Sekunden nach Ein- und Ausblendung), kardiovaskulären (ΔHR-Mittelwerte über die Sekunden 1 – 8 nach Ein- und Ausblendung) und erlebnisdeskriptiven Daten (Reaktionsstärke- und Bedeutsamkeitsratings). Die einzige signifikante Korrelation trat bei der phasischen HR nach Einblendung der Items auf ($\tau = .26$, $p = .022$; einseitiger Test). Statt des positiven wäre jedoch ein negativer monotoner Zusammenhang für die HR-Reaktionen zu erwarten gewesen, da der ΔHR-Mittelwert der Bedingung Relevant-Täuschung unter dem von Relevant-Aufrichtigkeit lag (siehe Abschnitt 5.2.3). Für die restlichen Reaktionen bzw. Ratings resultierten nur betragsmäßig kleine, nicht signifikante Korrelationen ($|\tau| < .15$, $p > .136$; einseitige Tests). Dies deutete darauf hin, daß die gefundenen Reaktionsunterschiede zwischen den beiden relevanten Itemtypen des GAT nicht unmittelbar darauf zurückzuführen waren, daß die Details des begangenen Scheinverbrechens (Relevant-Täuschung) besser wiedererkannt wurden als die wahrheitsgemäß beantworteten relevanten Items (Relevant-Aufrichtigkeit).

6. Diskussion

Eine **wesentliche Zielsetzung der vorliegenden Studie** bestand darin, im Rahmen eines Scheinverbrechen-Experiments mit standardisierter Durchführung der psychophysiologischen Aussagebeurteilung und objektiver Auswertung der Testverfahren quantitative Reaktionsunterschiede zwischen den einzelnen Stimulusarten des DLT bzw. GAT nachzuweisen und zu untersuchen. Es zeigten sich entsprechende Effekte der Fragen- bzw. Itemtypen. Dies galt sowohl für die elektrodermalen und kardiovaskulären als auch für die subjektiven Daten. Im folgenden werden zunächst die Ergebnisse der EDA diskutiert, da sie die wichtigste körperliche Variable der psychophysiologischen Aussagebeurteilung darstellt. Danach stehen die Befunde zur HR und zu den Ratings im Zentrum der Analyse. Anschließend wird auf die Frage nach den Einflüssen der elektrodermalen Labilität näher eingegangen.

6.1 Elektrodermale Reaktionsunterschiede zwischen den Fragen- bzw. Itemtypen

6.1.1 SCRs nach Einblendung der Stimuli

Im Hinblick auf die **SCRs**[8] **nach Einblendung der Reize** fand man für **beide Testverfahren** (DLT und GAT) Haupteffekte des Fragen- bzw. Itemtyps. Die genaueren Analysen zu den relevanten Stimuli ergaben, daß die Pbn insgesamt stärker auf diejenigen Reize reagierten, die sich auf das begangene Scheinverbrechen bezogen (Relevant-Täuschung), als auf jene Stimuli, die die nicht durchgeführte Tat thematisierten (Relevant-Aufrichtigkeit). Diese Befunde wurden zusätzlich durch die disordinalen Interaktionen der Faktoren Tatbedingung (Ring vs. Kette) und Tatbezug der relevanten Fragen bzw. Items (Relevant-Ring vs. Relevant-Kette) untermauert.

Beim **DLT** zeigte sich darüber hinaus, daß die beiden Arten von relevanten Fragen jeweils stärkere Reaktionen evozierten als die Kontroll- und irrelevanten Fragen. Entgegen den Erwartungen unterschieden sich die beiden Kontrollfragentypen nicht signifikant. Die SCR-Magnitude der Wahrheit-Kontrollfragen war sogar etwas größer als die der Lügen-Kontrollfragen. Abgesehen davon reagierten die Pbn auf die irrelevanten Stimuli stärker als auf die Kontrollfragen, wobei jedoch nur die Differenz zum Treatment Lügen-Kontroll statistisch bedeutsam war.

[8] Sofern nicht anderweitig spezifiziert, sind damit im folgenden die logarithmierten Amplitudenwerte der SCRs mit einer Latenzzeit von 1 – 3 Sekunden gemeint.

Beim **GAT** unterschieden sich die Mittelwerte der Bedingungen Relevant-Täuschung und Irrelevant-1 nicht signifikant. Auf beide Reiztypen wurde jedoch stärker reagiert als auf die Relevant-Aufrichtigkeit- und Irrelevant-Vergleich-Items. Letztere lösten erwartungsgemäß die schwächsten SCRs aus.

Diese **Reaktionsunterschiede auf die Darbietung der Fragen bzw. Items** lassen sich im Sinne einer differentiellen Bedeutsamkeit oder Signifikanz der Stimuli interpretieren. Auf der Basis empirisch gestützter Überlegungen kann man argumentieren, daß bedeutsame Reize, die etwa Aufgabenrelevanz besitzen, die einen ausgeprägten Selbstbezug aufweisen und/oder persönliche Interessen tangieren, stärkere SCRs auslösen als weniger signifikante Reize (z. B. Bernstein, Taylor & Weinstein, 1975, S. 166f.; Siddle et al., 1979, S. 525; Wingard & Maltzman, 1980, S. 157). Diese Interpretation steht im Einklang mit den Ansichten von Steller (1987, S. 140ff., 1997, S. 95ff.; vgl. auch Steller & Dahle, 1997, S. 310ff.), der die psychophysiologische Aussagebeurteilung als einen Spezialfall der sog. „vergleichenden psychophysiologischen Bedeutsamkeitsdiagnostik“ definiert, wobei die physiologischen Parameter intraindividuelle Unterschiede in dem subjektiv bewerteten Bedeutungsgehalt der Stimuli indizieren sollen.

Betrachtet man die SCR als einen Indikator der **Orientierungsreaktion**, so ist die o. g. Interpretation kompatibel mit jenen Theorien der forensischen Psychophysiologie, die die Reaktionsdifferenzen auf unterschiedlich starke ORs zurückführen (z. B. Dawson, 1980, S. 16; Lykken, 1974, S. 728; Raskin, 1979, S. 591; vgl. auch die Abschnitte 2.10.1 und 2.10.2). Demnach würden die tatbezogenen Stimuli des DLT bzw. GAT eine besonders ausgeprägte Signifikanz bzw. einen höheren Signalwert aufweisen und somit stärkere ORs und damit SCRs evozieren als die entsprechenden Vergleichsreize (Kontrollfragen bzw. Irrelevant-Vergleich-Items). Dies dürfte v. a. für die wahrheitswidrig verneinten relevanten Fragen bzw. Items gelten, da sich diese auf das von den Pbn durchgeführte Scheinverbrechen beziehen und einen potentiell größeren Selbstbezug aufweisen. In diesem Sinne konstatierten bereits Steller und Dahle (1997), daß die relevanten Stimuli „durch die Tatsache, daß der Betreffende eine in Frage stehende Straftat tatsächlich begangen hat, für diesen einen anderen (gewöhnlich höheren) subjektiven Bedeutungsgehalt aufweisen, als dies bei Nicht-Begehung der Fall wäre“ (S. 311). Eine solche Erklärung deckt sich auch weitgehend mit den Ergebnissen der Bedeutsamkeitsratings, auf die im Abschnitt 6.3 näher eingegangen wird.

Ein entscheidendes Bestimmungsmerkmal der OR ist ihre **Habituation** (vgl. Barry, 1984, S. 111). Diese hatte man zwar nicht in die Versuchsplanung der vorliegenden Untersuchung einbezogen, sie konnte jedoch post hoc untersucht werden, indem man die 30 (DLT) bzw. 36 (GAT) Trials zu jeweils sechs Trialblöcken zusammenfaßte und

die pro Vp und Block gemittelten SCR-Magnituden einfaktoriellen ANOVAs mit Meßwiederholung und anschließenden Trendanalysen (Bortz, 1993, S. 253ff.) unterzog. Beim DLT beinhaltete jeder Trialblock fünf aufeinanderfolgende Fragen (von jedem Typ eine; vgl. Abschnitt 4.4.3, Tabelle 12). Beim GAT entsprach ein Trialblock jeweils einer Multiple-Choice-Frage mit den zugehörigen sechs Items (vgl. Abschnitt 4.4.3, Tabelle 13). Für beide Befragungstechniken resultierten signifikante Blockeffekte (vgl. Anhang F: Abbildung 47 und 48). Die Abnahme der Reaktionsstärke folgte annähernd einer negativen Exponentialfunktion (mit der stärksten Reduktion von Block 1 zu Block 2) und erfüllte somit eine wichtige theoretische Voraussetzung für die Habituationskurve der OR (zusammenfassend Barry, 1996, S. 480f.; Vossel & Zimmer, 1989a, S. 112, 1989b, S. 142f.). Die Trendanalysen mit orthogonalen Polynomen ergaben jeweils für den DLT und den GAT neben einer signifikanten linearen auch eine quadratische Komponente (vgl. Anhang F), was unter Berücksichtigung der Blockmittelwerte ebenfalls dafür sprach, daß die Abnahme der SCR-Magnitude bei den ersten Trialblöcken stärker war als bei den späteren.

In diesem Zusammenhang muß man jedoch beachten, daß der **Rekurs auf die Annahmen und Theorien zur OR-Auslösung recht vage** ist. Die Befragungstechniken der psychophysiologischen Aussagebeurteilung unterscheiden sich fundamental von den klassischen experimentellen „Mikroparadigmen“ (vgl. Vossel, 1990, S. 227) zur Untersuchung der OR, die in der Regel eine mehrmalige Darbietung eines (meist relativ einfachen) Stimulus realisieren, gefolgt von einer Reizänderung bzw. Auslassung und eventuell anschließenden Wiederholungen des ursprünglichen Stimulus („repetition-change/omission paradigm“, Vossel & Zimmer, 1998, S. 160). Insofern sind die in der vorliegenden Untersuchung gefundenen Habituationseffekte besonders bemerkenswert, zumal in anderen Experimenten mit ähnlichem Reizmaterial (auditiv bzw. schriftlich dargebotene Fragen, wie z. B. in den DDP-Studien von Furedy et al., 1988, S. 686, und Rill, 1997, S. 125) gar keine bzw. keine exponentielle Abnahme der SCRs über die Trialblöcke nachzuweisen war. Diese Befunde wurden unter anderem auf die hohe Variabilität des Reizmaterials zurückgeführt, die einer Generalisierung der Habituation entgegenwirke (Furedy et al., 1994, S. 21). Dennoch muß man bedenken, daß das Phänomen der Habituation unspezifisch ist und nicht nur im Zusammenhang mit der OR auftritt. So bezieht sich etwa die sog. Zwei-Prozeß-Theorie der Habituation („dual-process theory“, Groves & Thompson, 1970, S. 420f., 1973, S. 175f.) auf die neurophysiologischen Auswirkungen wiederholter Reizung im allgemeinen. Außerdem wird der exponentielle Verlauf nicht von allen Autoren als notwendiges Charakteristikum einer OR-Habituation erachtet (z. B. Simons, 1989, S. 125; Turpin, 1989, S. 136). Folglich indizieren die beobachteten Reaktionsverläufe nicht unbedingt eine OR.

Die Deutung der Befunde im Sinne der **OR stößt auch an Grenzen**, wenn man die Kontroverse hinsichtlich der sie auslösenden bzw. verstärkenden Faktoren Reizneuheit vs. -signifikanz und deren interaktive vs. additive Wirkung auf die Reaktionsstärke berücksichtigt (vgl. Ben-Shakhar, 1994, S. 402f.). So könnte man etwa beim GAT argumentieren, daß die relevanten Items aufgrund des erkannten Tatbezugs signifikanter sind, die irrelevanten Items jedoch wegen ihrer fehlenden Beziehung zum vorherigen Scheinverbrechen eine größere Neuartigkeit aufweisen. Beide Faktoren dürften zu einer Intensivierung der OR beitragen, wobei ihr genaueres Zusammenwirken unklar bleibt. Unter diesen Gesichtspunkten würden die stärkeren SCRs auf die tatbezogenen Stimuli eher die Rolle der Reizbedeutung unterstreichen. Ähnliche Überlegungen lassen sich auch für die relevanten Fragen und Kontrollfragen des DLT anstellen. Dabei ist jedoch zu bedenken, daß die beiden Konzepte Neuheit und Signifikanz selbst innerhalb der OR-Forschung äußerst uneinheitlich definiert und operationalisiert werden (Baltissen & Sartory, 1998, S. 22). Die Übertragbarkeit der entsprechend inkonsistenten Ergebnisse auf die psychophysiologische Aussagebeurteilung ist fraglich.

Man kann die Reaktionsunterschiede zwischen den Fragen- bzw. Itemtypen auch aus der Perspektive des **informationsverarbeitungstheoretischen Erklärungsansatzes** von Waid und Mitarbeitern (Waid & Orne, 1981; Waid et al., 1978; Waid, Orne & Orne, 1981; vgl. Abschnitt 2.10.2) betrachten. Demzufolge dürften die Pbn besonders bedeutsamen Reizen mehr Aufmerksamkeit schenken und sie kognitiv „tiefer“, d. h. elaborierter verarbeiten, was mit stärkeren elektrodermalen Reaktionen und einer besseren Behaltensleistung der Stimuli einhergehen sollte. Zumindest die Reaktionsunterschiede beim **GAT** zwischen den beiden relevanten Itemtypen stehen damit in Einklang. Die entsprechenden Pbn zeigten eine höhere SCR-Magnitude auf die Items ihres Scheinverbrechens und konnten diese anschließend besser erinnern und wiedererkennen als die Details der nicht verübten Tat. Allerdings deuteten die korrelativen Post-hoc-Analysen darauf hin, daß die körperlichen Reaktionsunterschiede in keiner nennenswerten Beziehung dazu standen. Abgesehen von anderen grundlegenden Mängeln der Theorie (z. B. unbestätigte Kausalannahmen, vgl. Abschnitt 2.10.2) muß man außerdem feststellen, daß sich die hier erzielten Ergebnisse zum DLT damit nicht hinreichend erklären lassen, da trotz einer signifikanten elektrodermalen Differenzierung zwischen Relevant-Täuschung und Relevant-Aufrichtigkeit keine entsprechenden Unterschiede in den Gedächtnistests gefunden wurden.

Die höhere SCR-Magnitude unter der Bedingung **Relevant-Täuschung im Vergleich zu Relevant-Aufrichtigkeit** ist aber durchaus mit den bisherigen Befunden zum „Differentiation-of-Deception“-Paradigma (DDP, vgl. Abschnitt 2.9.3) vereinbar, und ebenso mit den Ergebnissen von Studien, die den Einfluß der Antwortart auf die Treff-

sicherheit der psychophysiologischen Aussagebeurteilung untersucht haben (vgl. Abschnitt 2.10.1). Dabei zeigte sich, daß der Wahrheitsgehalt der Antworten einen Effekt auf die Reaktionsstärke ausübt. Zur Erklärung wird häufig auf **konflikttheoretische Annahmen** zurückgegriffen, wonach eine Täuschungsabsicht die Person in einen Konflikt bringt, entweder zu lügen oder die Wahrheit zu sagen (z. B. Davis, 1961, S. 162f.; Furedy & Ben-Shakhar, 1991, S. 169; Furedy et al., 1988, S. 687; Heslegrave, 1982, S. 323). Dieser Zustand soll mit einer erhöhten emotionalen und körperlichen Erregung einhergehen. Einen Hinweis darauf, daß Konflikte einen Einfluß auf elektrodermale Reaktionen haben, erbrachte bereits Berlyne (1961), indem er die Konfliktintensität variierte und die Auswirkungen auf die EDA erfaßte. Seine Untersuchung basierte auf der Annahme, daß Konflikte ORs auslösen bzw. verstärken können (vgl. auch Lynn, 1966, S. 12f.). Entsprechend dieser Hypothese gingen intensivere Konflikte auch mit erhöhten Hautwiderstandsreaktionen einher. Darüber hinaus zeigte sich in einem seiner Experimente, daß der Reaktionsunterschied zwischen den Bedingungen „low-conflict“ und „high-conflict“ (Berlyne, 1961, S. 479) auf die Darbietung jener Stimuli zurückzuführen war, die als Auslöser der divergenten Handlungstendenzen galten. Indes traten bei den Reaktionen auf die Stimulusausblendung (dort imperativer Reiz für die geforderten motorischen Reaktionen) bzw. auf die Handlung selbst keine signifikanten Differenzen mehr auf. Dies wäre eine zusätzliche Erklärung dafür, daß sich in der vorliegenden Studie die Unterschiede zwischen den wahrheitsgemäß vs. wahrheitswidrig beantworteten relevanten Fragen bzw. Items bereits in den Reaktionen auf deren Darbietung manifestierte. Allerdings sind die Reaktionsunterschiede zwischen Relevant-Täuschung und -Aufrichtigkeit nicht zwangsläufig ausschließlich auf den intendierten Wahrheitsgehalt der Antworten zurückzuführen. Die DDP-Studien und viele der bisherigen Untersuchungen zur psychophysiologischen Aussagebeurteilung mit Variation der Antwortart legen zwar nahe, daß eine Täuschungsabsicht bzw. -handlung potentiell zu einer Reaktionssteigerung führt, doch wahrheitswidrige Antworten sind keine notwendige Bedingung für erhöhte Reaktionen auf tatbezogene Stimuli (vgl. Gustafson & Orne, 1965b, S. 12f.; Reid & Inbau, 1977, S. 150ff.). Möglicherweise spielen auch Bedeutsamkeitsunterschiede eine Rolle, die unabhängig vom Wahrheitsgehalt der Antworten sind, wie z. B. der stärkere Selbstbezug der relevanten Fragen bzw. Items, die sich auf die durchgeführte Tat beziehen (s. o.). Ferner sei nochmals betont, daß auch die Ergebnisse des DDP nicht direkt auf die hier verwendeten Befragungstechniken übertragbar sind.

So zeigten sich etwa entgegen den Erwartungen des DDP keine signifikanten Reaktionsunterschiede zwischen den (instruiert) wahrheitsgemäß vs. wahrheitswidrig beantworteten **Kontrollfragen des DLT**. Die SCR-Magnitude von Wahrheit-Kontroll lag sogar leicht höher als die von Lügen-Kontroll. Dies deutete darauf hin, daß der Wahr-

heitsgehalt bei den Kontrollfragen keinen bedeutsamen Effekt auf die Reaktionsstärke ausübte. Der Mittelwert der Lügen-Kontrollfragen reichte auch nicht an die Bedingung Relevant-Aufrichtigkeit heran. Im Gegenteil, relativ zu allen anderen Fragentypen evozierten sie die schwächsten Reaktionen. Selbst der Durchschnitt der irrelevanten Fragen lag signifikant darüber. Diese Befunde stützen die von Lykken (1998, S. 138ff.) angesprochene Kritik, daß die instruierten Falschantworten beim DLT nicht ohne weiteres mit intentionalen Täuschungen zu vergleichen sind und daß die Lügen-Kontrollfragen nicht mit den tatbezogenen Fragen konkurrieren können und somit keine angemessene Vergleichsbedingung darstellen. Dem kann man jedoch entgegenhalten, daß zumindest bei den Reaktionen auf die Ausblendung die SCR-Magnitude der Lügen-Kontrollfragen höher lag als die der Bedingung Relevant-Aufrichtigkeit, wenngleich dieser Unterschied nicht statistisch bedeutsam war (s. u.).

Ebenfalls unerwartet resultierten relativ starke Reaktionen auf die **irrelevanten Fragen des DLT**. Diese sind möglicherweise dadurch zu erklären, daß sie die einzigen Fragen des Testverfahrens waren, die bejaht wurden, was eventuell zu einer erhöhten Bedeutsamkeit der Stimuli und damit zu einer Reaktionssteigerung beigetragen hat.

Bei den SCRs nach Einblendung unterschieden sich die **Irrelevant-1-Items des GAT** nicht signifikant von den Relevant-Täuschung-Items. Beide Magnituden lagen über den entsprechenden Werten der Bedingungen Relevant-Aufrichtigkeit und Irrelevant-Vergleich. Die starken Reaktionen auf die ersten irrelevanten Items pro Multiple-Choice-Frage sind eventuell auf die Neuheit der angesprochenen Thematik zurückzuführen sowie auf das komplexere Reizmaterial, da in den entsprechenden Trials nacheinander sowohl die Einleitungsfrage als auch das erste Item dargeboten wurden.

Hypothesenkonform traten die schwächsten Reaktionen des GAT bei den **Irrelevant-Vergleich-Items** auf. Im Einklang mit der Feststellung von Elaad und Ben-Shakhar (1989, S. 450), wonach vorhandenes Tatwissen eine hinreichende Voraussetzung für einen positiven GAT-Befund darstellt, war die SCR-Magnitude von Relevant-Aufrichtigkeit höher als unter der Bedingung Irrelevant-Vergleich. Im Sinne der OR-Theorie zur psychophysiologischen Aussagebeurteilung könnte man daraus schließen, daß (unabhängig von der Täterschaft bzw. vom Wahrheitsgehalt der Antworten) bereits der erkannte Tatbezug der relevanten Items diesen einen erhöhten Signalwert verleiht und somit zu stärkeren Reaktionen führt als die zum Vergleich herangezogenen irrelevanten Items.

Unter **physiologischen Gesichtspunkten** lassen sich die gefundenen elektrodermalen Reaktionsunterschiede im Sinne einer differentiellen sympathischen Erregung auffas-

sen, zumal es gerechtfertigt erscheint „von EDA-Maßen – insbesondere von Amplitudenmaßen – relativ direkt auf die Aktivität des sympathischen Nervensystems zu schließen“ (Vossel, 1990, S. 49). D. h., stärkere SCRs auf bestimmte Fragen- bzw. Itemtypen indizieren eine erhöhte Aktivierung des peripherphysiologischen Teils des Sympathikus, der die ekkrinen Schweißdrüsen innerviert. Diese Feststellung wird später noch relevant, wenn die elektrodermalen und kardiovaskulären Daten in Bezug zueinander gesetzt werden, da die periphere vegetative Modulation der Herzaktivität im Gegensatz zur Hautleitfähigkeit sowohl sympathischen als auch parasympathischen Einflüssen unterliegt.

6.1.2 SCRs nach Ausblendung der Stimuli

In den **SCRs nach Ausblendung** zeigte sich sowohl beim DLT als auch beim GAT eine **geringere Differenzierung** zwischen den unterschiedlichen Stimulusarten. Zwar resultierten wiederum für beide Befragungstechniken statistisch bedeutsame Haupteffekte des Fragen- bzw. Itemtyps[9]. Die Effekte waren jedoch schwächer als bei den SCRs nach Einblendung der Reize (gemäß Bortz & Döring, 1995, S. 571, aus η^2 bestimmte Effektgröße *f* [SCRs nach Ein- vs. Ausblendung]: DLT: 0.82 vs. 0.27, GAT: 0.87 vs. 0.35). Abgesehen davon ergaben die Einzelvergleiche, daß nur der Mittelwert der Bedingung Relevant-Täuschung signifikant höher lag als die SCR-Magnituden der anderen Fragen- bzw. Itemtypen, während sich letztere nicht überzufällig voneinander unterschieden. Gleichzeitig folgte daraus, daß die Pbn nach Ausblendung der Stimuli, die sich auf das begangene Scheinverbrechen bezogen (Relevant-Täuschung), stärker reagierten als bei den entsprechenden Reizen der nicht verübten Tat (Relevant-Aufrichtigkeit). Diese Interpretation wurde wiederum durch die 3-fachen ANOVAs und die daraus resultierenden signifikanten Interaktionen Tatbedingung × Tatbezug der relevanten Fragen bzw. Items gestützt.

Diese Befunde **konfligieren** teilweise mit der Analogstudie von Dawson (1980, S. 14). Dawson konnte für die elektrodermalen Reaktionen schuldiger und unschuldiger Pbn auf die um 8 Sekunden verzögerten Antworten keine überzufälligen Unterschiede zwischen den relevanten Fragen und Kontrollfragen des KFT nachweisen. Und auch in der DDP-Studie von Furedy et al. (1988, S. 686) zeigte sich zwar in den SCRs auf die Fragen eine reliable Differenzierung zwischen Täuschung und Aufrichtigkeit, nicht jedoch in den SCRs auf die späteren Antworten.

[9] Was den DLT anbelangt, ist jedoch einschränkend anzumerken, daß die Analyse der untransformierten (nicht logarithmierten) Amplitudenwerte keinen signifikanten Haupteffekt des Fragentyps erbrachte.

Die Ergebnisse der vorliegenden Untersuchung sind aber **konkordant** mit anderen DDP-Experimenten, die auch in den SCRs auf die verzögerten Antworten bzw. auf den imperativen Reiz der Antwortgabe DDP-Phänomene fanden (Furedy et al., 1991, S. 95f.; Gödert et al., 2001, S. 67; vgl. dazu ferner Rill, 1997, S. 122). Allerdings fiel dort die Differenzierung jeweils deutlich schwächer aus als bei den SCRs auf die Fragendarbietung. In ähnlicher Weise waren hier die Magnitudendifferenzen zwischen Relevant-Täuschung und Relevant-Aufrichtigkeit bei den SCRs nach Stimuluseinblendung (DLT: 0.0541 log µS; GAT: 0.0433 log µS) größer als bei den SCRs nach Ausblendung (DLT: 0.0252 log µS; GAT: 0.0252 log µS). Diese Resultate decken sich gleichermaßen mit dem Kartentest-Experiment von Furedy und Ben-Shakhar (1991, S. 167), die für die relevanten und irrelevanten Items des TWT ebenfalls nachweisen konnten, daß die elektrodermalen Reaktionsunterschiede auf die Präsentation der Stimuli stärker waren als auf die verzögerten Antworten.

Die **Diskrepanzen zwischen den Reaktionen auf die Fragen und Antworten** interpretierte Dawson (1980, S. 16) im Sinne der OR-Theorie der psychophysiologischen Aussagebeurteilung. Demzufolge sei die differentielle Signifikanz der Reize für die Reaktionsunterschiede zwischen den relevanten und Kontrollfragen des KFT verantwortlich, während die Antworten darauf und somit deren Wahrheitsgehalt eine nebensächliche Rolle spielen würden. Furedy et al. (1988, S. 687) und Rill (1997, S. 123f.) gingen davon aus, daß man das elektrodermale DDP-Phänomen im wesentlichen auf die Täuschungsabsicht – im Sinne einer Handlungsvorbereitung – und weniger auf das Lügen selbst zurückführen könne. Gleichsam konstatierten Furedy und Ben-Shakhar (1991, S. 169), daß der verbale Akt des Lügens per se weder eine notwendige noch eine hinreichende Bedingung für stärkere SCRs auf die relevanten Items des TWT darstelle. Statt dessen verwiesen sie ebenfalls auf die intentionalen Prozesse der Täuschung als potentiell ausschlaggebende Faktoren für die Reaktionsunterschiede.

Die o. g. Erklärungen lassen sich analog auf die **elektrodermalen Ergebnisse der vorliegenden Studie** übertragen. Folglich führt beim DLT und GAT bereits die Darbietung der Stimuli zu differentiellen SCRs auf die Fragen- bzw. Itemtypen, und die Täuschungsabsicht trägt eventuell zum Reaktionsunterschied zwischen den Bedingungen Relevant-Täuschung und Relevant-Aufrichtigkeit bei. Die körperlichen Begleiterscheinungen der Antworten, die sich auch in den SCRs auf die Ausblendung manifestieren sollen, zeigen hingehen schwächere Effekte. Allerdings sei an dieser Stelle bereits darauf hingewiesen, daß für die HR-Daten ein umgekehrtes Muster resultierte, mit einer stärkeren Differenzierung zwischen den Fragen- bzw. Itemtypen bei den Reaktionen nach Ausblendung. Auf diesen Befund wird später noch näher eingegangen.

Als einen alternativen, mehr physiologisch orientierten Erklärungsansatz boten Furedy et al. (1988, S. 687) das Konzept der **Reaktionsinterferenz** an. Demnach werde ähnlich wie während einer relativen Refraktärphase die Stärke einer SCR in Abhängigkeit von der Intensität einer zuvor erfolgten Reaktion reduziert (vgl. Abschnitt 3.2.3). Je höher also beispielsweise die SCR auf eine Frage ausfällt, desto größer sollte die anschließende Hemmung der SCR auf die verzögerte Antwort sein. Dadurch könnten nicht nur die Reaktionen auf die Antworten insgesamt vermindert, sondern zusätzlich potentielle Reaktionsunterschiede zwischen den Reiztypen nivelliert werden. Gemäß diesen Annahmen waren in den o. g. Experimenten von Dawson (1980), Furedy und Ben-Shakhar (1991), Furedy et al. (1991) sowie Gödert et al. (2001) bei den SCRs auf die Antworten bzw. imperativen Reize der Antwortgabe sowohl die Unterschiede zwischen den Reaktionen als auch deren Absolutbeträge relativ niedrig. Die Möglichkeit einer Reaktionsinterferenz zwischen den SCRs nach Ein- und Ausblendung ist in der vorliegenden Untersuchung ebenfalls nicht ohne weiteres von der Hand zu weisen. Denn beim DLT und GAT lagen die Gesamtmagnituden der SCRs auf die Ausblendung deutlich unterhalb der entsprechenden Mittelwerte nach Einblendung (DLT: Einblendung: M = 0.1614 log µS, Ausblendung: M = 0.1336 log µS; GAT: Einblendung: M = 0.1439 log µS, Ausblendung: M = 0.0889 log µS).

Das Konzept der Reaktionsinterferenz ist jedoch sehr vage, zumal keine fundierte theoretische Erklärung dafür gegeben wird. Nach Furedy und Ben-Shakhar (1991, S. 169) kann man die Reaktionsinterferenz im Zusammenhang mit der **Ausgangswertabhängigkeit** der EDA sehen (vgl. dazu Boucsein, 1988, S. 233ff.). Diese wurde bereits von Dawson (1980, S. 15f.) als mögliche Begründung für die schwachen SRRs und fehlenden Reaktionsunterschiede nach den Antworten beim KFT herangezogen. Demzufolge soll in Anlehnung an das sog. „Ausgangswert-Gesetz“ (Wilder, 1931, S. 317) die Stärke einer elektrodermalen Reaktion in einem umgekehrt proportionalen Verhältnis stehen zum vorherigen Niveauwert, auf dem sie ansetzt. Dawson gab zu bedenken, daß die elektrodermalen Reaktionen auf die Fragen der „Delayed Answer“-Bedingung meistens noch nicht vollständig abgeklungen waren, als die Reaktionen auf die verzögerten Antworten begannen. Dadurch hätten letztere aufgrund des höheren Ausgangswertes schwächer ausfallen können. Mittlerweile wird jedoch die Ausgangswertabhängigkeit der EDA stark angezweifelt, zumal es sich bei den diesbezüglich gefundenen Zusammenhängen zwischen tonischen und phasischen Größen um ein potentielles statistisches Artefakt handelt (zusammenfassend Vossel, 1990, S. 58ff.).

6.2 Kardiovaskuläre Reaktionsunterschiede zwischen den Fragen- bzw. Itemtypen

6.2.1 HR-Reaktionen nach Einblendung der Stimuli

Betrachtet man beim **DLT und GAT** die HR-Verläufe der Fragen- bzw. Itemtypen in den Poststimulus-Sekunden 1 – 8 nach Einblendung, so fällt v. a. auf, daß die Reaktionen der Bedingungen Relevant-Täuschung und Relevant-Aufrichtigkeit unterhalb der Reaktionen auf die restlichen Stimulusarten lagen. D. h., auf die tatbezogenen Stimuli wurde im Durchschnitt mit schwächeren Akzelerationen bzw. stärkeren Dezelerationen reagiert. Darüber hinaus verlief die ΔHR von Relevant-Täuschung jeweils unter der entsprechenden Kurve von Relevant-Aufrichtigkeit. Für beide Testverfahren resultierten außerdem signifikante Haupteffekte des Fragen- bzw. Itemtyps, und der Gesamtmittelwert von Relevant-Täuschung war jeweils kleiner als der von Relevant-Aufrichtigkeit. Die Einzelvergleiche der über die 8 Sekunden gemittelten ΔHR-Werte ergaben jedoch, daß sich die beiden Bedingungen nicht signifikant unterschieden. Dies galt sowohl für den DLT als auch den GAT. Die entsprechenden ANOVAs zu den Effekten der tatbezogenen Stimuli erbrachten ebenfalls nur marginal signifikante Interaktionen der Faktoren Tatbedingung und Tatbezug der relevanten Fragen bzw. Items. Dies deutete darauf hin, daß sich zwischen den relevanten Stimuli, die sich auf die begangene vs. nicht begangene Tat bezogen, keine statistisch bedeutsamen Unterschiede in den HR-Reaktionen nach Einblendung zeigten.

Im Hinblick auf die instruierten Lügen- und Wahrheit-Kontrollfragen sowie die irrelevanten Fragen des **DLT** manifestierten sich keine signifikanten Differenzen zwischen den Mittelwerten über die Sekunden. Die drei Bedingungen unterschieden sich lediglich in Relation zu den beiden relevanten Fragentypen.

Auch der paarweise Vergleich der beiden irrelevanten Itemtypen des **GAT** fiel nicht statistisch bedeutsam aus. Darüber hinaus erreichte der Mittelwertsunterschied zwischen Irrelevant-1 und Relevant-Aufrichtigkeit nicht das Signifikanzniveau, wohingegen der Mittelwert von Relevant-Aufrichtigkeit signifikant unterhalb des Durchschnitts von Irrelevant-Vergleich lag.

Die **mangelnde Differenzierung** zwischen den Bedingungen Relevant-Täuschung und Relevant-Aufrichtigkeit läßt sich mit den HR-Daten der DDP-Studie von Gödert et al. (2001, S. 67ff.) vergleichen. Dort resultierte im Zeitfenster nach Einblendung der Fragen ebenfalls kein kardiovaskuläres DDP-Phänomen (weder in den sekundenbezogenen Verläufen noch in den Reaktionsstärkeparametern). Damit kann man auch den nicht

signifikanten Unterschied zwischen den Lügen- und Wahrheit-Kontrollfragen des DLT in der vorliegenden Studie in Einklang bringen.

Die **niedrigere ΔHR auf die tatbezogenen Stimuli** und die negativen Gesamtmittelwerte insbesondere von Relevant-Täuschung sind vereinbar mit jenen numerischen Auswertungsmethoden der psychophysiologischen Aussagebeurteilung, die eine Verlangsamung der Puls- bzw. Herzschlagfrequenz als Reaktionskriterium heranziehen (z. B. Bradley & Janisse, 1981, S. 310; Raskin & Hare, 1978, S. 129). Außerdem korrespondieren sie mit Analogstudien zum KFT, in denen Schuldige auf die relevanten Fragen im Durchschnitt mit einer ausgeprägteren Dezeleration reagierten (Podlesny & Raskin, 1978, S. 354; Raskin & Hare, 1978, S. 133). Allerdings trat dort die differentielle dezelerative Komponente erst nach dem durchschnittlichen Zeitpunkt der Antwortgabe auf. Sie wurde im Sinne einer Aufmerksamkeitsreaktion interpretiert. Darin spiegle sich möglicherweise der Versuch schuldiger Pbn wider, externe Hinweise und Informationen darüber zu erhalten, inwiefern der Untersucher ihre Täuschungen erkenne (Raskin & Hare, 1978, S. 135). Podlesny und Raskin (1978, S. 355) fanden jedoch keine entsprechenden Reaktionsunterschiede zwischen den relevanten und irrelevanten Items des TWT. Im Gegensatz dazu zeigten einige Pbn im Kartentest-Experiment von Cutrow et al. (1972, S. 584) eine stärkere Dezeleration auf die relevanten Items, was als Anzeichen für eine Orientierungsreaktion gewertet wurde.

Diese Interpretation basiert auf der umstrittenen Annahme, daß eine HR-Dezeleration unter bestimmten Randbedingungen einen **Indikator der OR** darstellt (vgl. Graham & Clifton, 1966, S. 316). Nach Barry (1984, S. 118, 1996, S. 474) hingegen kennzeichnet eine initiale Abnahme der HR nach Reizbeginn allenfalls die basale Registrierung eines überschwelligen Stimulus nach dem Alles-oder-Nichts-Prinzip, ohne daß dieser Prozeß bereits eine bewußte Reizverarbeitung umfaßt. Demzufolge ist die Reaktion weitgehend unabhängig von physikalischen oder psychologischen Reizcharakteristika, wie Intensität, Neuheit oder Bedeutsamkeit. Darüber hinaus findet man für die initiale Dezeleration keine konsistenten empirischen Belege im Hinblick auf die theoretisch erwarteten Kennzeichen einer OR (s. o.), wie z. B. einen exponentiellen Habituationsverlauf (Vossel & Zimmer, 1989a, S. 121).

Allgemein kann man feststellen, daß die phasische HR häufig im Zusammenhang mit Prozessen der **Aufmerksamkeitsregulation** untersucht wird (Vossel & Zimmer, 1998, S. 78). Eine klassische Versuchsanordnung hierfür ist das sog. „S_1-S_2-Paradigma“. Dabei markiert der erste Stimulus (Ankündigungsreiz S_1) den Beginn einer Vorbereitungsphase von mehreren Sekunden. Diese endet mit dem zweiten Stimulus (S_2), der gleichzeitig als imperativer Reiz für eine Reaktion dient (meist motorischer bzw.

diskriminatorischer Art mit der Instruktion, so schnell und genau wie möglich zu reagieren). Bei konstanten Interstimulusintervallen, die ungefähr im Bereich der hier verwendeten Zeitspanne zwischen Ein- und Ausblendung liegen, zeigt sich in der Regel ein charakteristischer triphasiger Verlauf der HR (Bohlin & Kjellberg, 1979, S. 169; Simons, 1988, S. 227). Dessen Form ähnelt sehr stark der für den DLT und GAT gefundenen mittleren Reaktionsmorphologie nach Einblendung der Fragen bzw. Items (vgl. Grand Averages in den Abbildungen 14 und 15, Abschnitt 5.2.1). D. h., die phasische HR beginnt mit einer kurzen Dezeleration auf S_1, geht dann in eine akzelerative Komponente über und mündet schließlich in einer erneuten Abnahme, die ihr Maximum ungefähr bei S_2 bzw. zu Beginn der geforderten (motorischen) Reaktion erreicht (vgl. z. B. Koers, Gaillard & Mulder, 1997, S. 270; Otten, Gaillard & Wientjes, 1995, S. 96f.; Wölk, Velden, Zimmermann & Krug, 1989, S. 399f.). Gödert et al. (2001, S. 67f.) fanden in ihrem DDP-Experiment ebenfalls entsprechende Reaktionsverläufe.

Zwei mögliche Erklärungen für die **initiale (ereignisbezogene) Dezeleration** nach S_1 wurden bereits angesprochen (OR-Indikator vs. Reizregistrierung). Auch beim GAT und (in geringerem Ausmaß) beim DLT konnte man in der ersten Poststimulus-Sekunde nach Einblendung eine kurzfristige Abnahme der HR beobachten. Diese Komponente erwies sich aber als wenig relevant, weil sie keine signifikante Differenzierung zwischen den Fragen- bzw. Itemtypen zeigte. Die post hoc für diese Sekunde gerechneten einfaktoriellen ANOVAs mit Meßwiederholung auf dem Fragen- bzw. Itemtyp erbrachten keine entsprechenden Haupteffekte (DLT: $F[4,156] = 1.21$, $p = .311$, $\varepsilon = .857$; GAT: $F < 1$). Insofern sprechen diese Befunde gegen die OR-Interpretation. Denn aufgrund der a priori anzunehmenden differentiellen Bedeutsamkeit der Fragen- bzw. Itemtypen wäre mit Reaktionsunterschieden zu rechnen gewesen.

Der Status der **akzelerativen Komponente** ist strittig. Bohlin und Kjellberg (1979, S. 182ff.) stellten vier alternative Erklärungsansätze gegenüber. Dabei wurde die Akzeleration im S_1-S_2-Paradigma interpretiert als: (1) Reaktion auf Merkmale von S_1, die keine Signalbedeutung besitzen, (2) Manifestation der durch S_1 ausgelösten Informationsverarbeitungsprozesse, (3) Reaktion auf die Signalbedeutung von S_1, (4) Antizipation und Vorbereitung der Reaktion auf S_2. Aufgrund mangelnder empirischer Evidenzen war jedoch keine abschließende Bewertung der sich teilweise widersprechenden Annahmen möglich. In Anlehnung an Punkt (2) und (3) gingen Koers et al. (1997, S. 270) davon aus, daß die akzelerative Komponente die Verarbeitung des Informationsgehalts und der Signaleigenschaften von S_1 widerspiegele, und sie fanden experimentelle Belege dafür.

In bezug auf die vorliegende Studie ist jedoch der Erklärungswert aller o. g. Ansätze sehr begrenzt. Sie können nicht ohne weiteres plausibel begründen, warum auf die potentiell bedeutsameren tatbezogenen Fragen und Items mit einem schwächeren HR-Anstieg reagiert wurde. Unter den genannten Voraussetzungen könnte man allenfalls spekulieren, daß die Pbn eine mentale Beschäftigung mit den relevanten Stimuli *vermieden* und versuchten, sich davon *abzulenken*. Diese Vermutung wird durch die Nachbefragung gestützt, in der viele Vpn gedankliche Ablenkungsmanöver, wie z. B. nicht auf die betreffenden Reize zu achten, als Strategie gegen eine erfolgreiche „Lügendetektion“ angaben (vgl. Abschnitt 5.4.1.3). Dagegen würde eine solche These den elektrodermalen Daten und der kognitiven Theorie der Arbeitsgruppe um Waid (siehe Waid & Orne, 1981; Waid et al., 1978; Waid, Orne & Orne, 1981) widersprechen, wonach die stärkeren SCRs nach Darbietung der tatbezogenen Stimuli auf eine „tiefere“, elaboriertere Verarbeitung zurückzuführen sind. Somit bleiben im Hinblick auf die HR-Akzeleration und die hier beobachteten Reaktionsunterschiede erhebliche Interpretationsprobleme, denn: „heart rate acceleration has been related to cognitive elaboration“ (Graham & Hackley, 1991, S. 259).

Die zweite dezelerative Komponente, die man auch als **antizipatorische Dezeleration** bezeichnet, wird wiederum relativ konsistent im Zusammenhang mit Aufmerksamkeitsprozessen und der Orientierungsaktivität diskutiert (vgl. Übersicht von Zimmer, Vossel & Fröhlich, 1989, S. 252ff.). In einem S_1-S_2-Paradigma mit komplexen Reizen, das bereits während der Vorbereitungsphase recht hohe Anforderungen an die Informationsverarbeitung stellte (v. a. Gedächtnisprozesse und deren Koordination mit Wahrnehmung und Reaktionsvorbereitung), konnten Zimmer et al. (1989) die antizipatorische Dezeleration reliabel nachweisen und einen Zusammenhang zur Reaktionszeit auf S_2 finden. Bei stärkerer HR-Abnahme waren die Reaktionszeiten kürzer. Darüber hinaus zeigte sich ein marginal signifikanter Komplexitätseffekt mit einer tendenziell ausgeprägteren Dezeleration bei komplexeren Reizen. Die Autoren interpretierten diese Reaktion als einen Indikator der Aufmerksamkeitsregulation. Mit Verweis auf die Arbeiten der Laceys (zusammenfassend Lacey & Lacey, 1974) vermuteten bereits Chase, Graham und Graham (1977, S. 647) hinter der zweiten dezelerativen Komponente einen Aufmerksamkeitsprozeß, der primär zur Erhöhung der Wahrnehmungssensitivität diene. Auch nach Walter und Porges (1976, S. 569) manifestiere sich darin external gerichtete Aufmerksamkeit. Als weitere Erklärungsansätze diskutierten Bohlin und Kjellberg (1979, S. 189ff.) Prozesse der Reaktionsvorbereitung und Erwartungsbildung (vgl. ebenso Coles & Duncan-Johnson, 1975, S. 425f.).

Wölk und Velden (1989, S. 378) schrieben der antizipatorischen Dezeleration sogar eine instrumentelle Bedeutung bei der **Aufmerksamkeitssteuerung** zu. In ihrer Revi-

sion der ursprünglich von Lacey (1967, S. 26ff.) formulierten „Barorezeptorhypothese“ postulierten sie, daß die HR-Reduktion über die barorezeptorischen Afferenzen eine Desynchronisation der hirnelektrischen Aktivität und somit eine Anpassung sensumotorischer Funktionen des Zentralnervensystems an die antizipierten Anforderungen bewirkt (siehe auch Wölk et al., 1989, S. 401). Mit anderen Worten, die Herzfrequenzverlangsamung dürfte zu einer Verbesserung der Reizaufnahme und -verarbeitung führen (Velden, 1994, S. 71).

Im Gegensatz dazu vermutete Simons (1988, S. 260f.), daß die Dezeleration vor S_2 weniger mit Aspekten der Wahrnehmung und Reaktionsvorbereitung zu tun hat, als vielmehr mit affektiven bzw. **emotionalen Prozessen**. Otten et al. (1995, S. 100) konnten jedoch diese Hypothese anhand ihrer Untersuchung nicht bestätigen.

Interessanterweise ergab die **eigene Studie** nur in den mittleren HR-Verläufen der relevanten Fragen bzw. Items ausgeprägte dezelerative Komponenten (vgl. Abschnitt 5.2.3, DLT: Abbildung 16, Poststimulus-Sekunden 6 – 8; GAT: Abbildung 19, Poststimulus-Sekunden 7 – 8) [10]. Dieser Befund ist kongruent mit den o. g. Ergebnissen der Scheinverbrechen-Experimente von Podlesny und Raskin (1979) sowie Raskin und Hare (1978), in denen Schuldige mit einer deutlichen Dezeleration auf die relevanten Fragen des DLT reagierten (dort jedoch nach Antwortgabe). Wie bereits erwähnt, wurde dies als eine auf die Verarbeitung externer Stimuli ausgerichtete Aufmerksamkeitsreaktion interpretiert. Auch angesichts der geschilderten Ergebnisse zum S_1-S_2-Paradigma kann man argumentieren, daß die Dezeleration eine erhöhte Aufmerksamkeit bei tatbezogenen Stimuli speziell in Vorbereitung auf die Antwort indiziert. Eine solche Sichtweise wäre ebenso mit den elektrodermalen Daten vereinbar, die stärkere SCR-Magnituden für die relevanten Fragen bzw. Items erbrachten.

Einschränkend sei jedoch vermerkt, daß trotz konzeptioneller Ähnlichkeiten zwischen dem S_1-S_2-Paradigma und der eigenen Methodik einige gravierende Unterschiede bestehen. Diese betreffen zum einen das Reizmaterial. Im vorliegenden Experiment wurden komplexe sprachliche Stimuli längerfristig dargeboten. Die schriftliche Darbietung der Fragen und Items erstreckte sich bis zum imperativen Reiz der Antwortgabe (Ausblendung). Das Zeitintervall zwischen Ein- und Ausblendung war nicht konstant. Außerdem handelte es sich um keine Reaktionszeitaufgabe mit der Instruktion, so schnell wie möglich zu antworten. Dennoch erscheinen gewisse Analogieschlüsse zulässig, zumal im Experiment von Zimmer et al. (1989) ebenfalls recht komplexe

[10] Dabei ist zu beachten, daß zumindest bei kurzen Reizdarbietungen (im Bereich von 8 – 9 Sekunden) bereits die HR-Reaktion auf die Ausblendung die echtzeitskalierte Schlagfrequenz in der achten Poststimulus-Sekunde nach Einblendung beeinflussen konnte.

Reize verwendet wurden und die Aufgabe hohe Anforderungen an die Informationsverarbeitung stellte. Und selbst bei noch anspruchsvolleren Problemstellungen (z. B. Items von Intelligenztests), die unter anderem eine erhöhte Aufmerksamkeit auf externe Stimulation erfordern, hat man ähnliche Reaktionsverläufe gefunden (vgl. McCanne & Lyons, 1990).

Sowohl die initiale als auch die antizipatorische Dezeleration sollen im wesentlichen auf den zentralnervös modulierten, inhibitorischen Einflüssen des **Parasympathikus** auf die Herzschlagfrequenz basieren (Velden, 1994, S. 71; Wölk et al., 1989, S. 401; Zimmer et al., 1989, S. 254f.). Teilweise wird die akzelerative Komponente gleichfalls auf vagale Effekte zurückgeführt (Koers et al., 1997, S. 272), zumal phasische Herzfrequenzbeschleunigungen im Humanbereich häufig durch Hemmungen der Vagusaktivität bedingt seien (Jennings, Nebes & Yovetich, 1990, S. 88). Heslegrave (1982) nahm innerhalb einer psychophysiologischen Untersuchung zum Phänomen der Täuschung eine kombinierte Messung der SCR, HR und T-Wellen-Amplitude (TWA) vor. Während eine Veränderung der HR (Beschleunigung oder Verlangsamung) sowohl die sympathische als auch die parasympathische Regulation der Herztätigkeit wiedergibt (vgl. Abschnitt 3.2.2), gilt die Dämpfung der TWA im EKG als ein relativ sensibler und valider Indikator der Sympathikuswirkung auf das Herz (zusammenfassend Furedy, 1985, S. 239ff.). Heslegrave (1982) interpretierte seine Befunde dahingehend, daß der Täuschungsprozeß auf psychologischer Ebene einen Konflikt und auf physiologischer Ebene eine Aktivitätssteigerung des Parasympathikus sowie eine Verhaltenshemmung impliziere: „Arousal during deception could be characterized as *inhibitory* arousal associated with excitation of the parasympathetic system“ (S. 323). Da die Arbeit nur in Form eines Abstracts publiziert wurde, ist nicht unmittelbar ersichtlich, um welches Paradigma es sich handelte. Nach Angaben von Furedy et al. (1988, S. 684, 1991, S. 91) dürfte die Untersuchung analog zum DDP konzipiert gewesen sein. Ebenfalls mit Verweis auf Heslegrave (1982) berichteten Ben-Shakhar und Furedy (1990), daß die Differenzierung zwischen Täuschung und Aufrichtigkeit in der HR-Dezeleration ohne Anzeichen von differentiellen sympathischen Aktivitätsveränderungen in der TWA beobachtet worden sei. Die Autoren deuteten diesen Befund im Sinne einer parasympathisch vermittelten „Versteck“-Reaktion: „This pattern of results suggests that deception involves parasympathetic withdrawal (‘hiding’) rather than a sympathetic activation (fight-or-flight response)“ (Ben-Shakhar & Furedy, 1990, S. 142).

Auch in einer unveröffentlichten Studie von Heslegrave und Furedy (1984, zitiert nach Furedy, 1985, S. 252) sei unter der **Bedingung Täuschung eine stärkere HR-Dezeleration** beobachtet worden als unter der Bedingung Aufrichtigkeit. Erneut fand man keine Unterschiede hinsichtlich der TWA-Dämpfung. Unter eher ethologischen

Gesichtspunkten wurde die Aktivierung des Parasympathikus (PNS) bei gleichzeitig konstanter Aktivität des Sympathikus (SNS) in Analogie zur Reaktion des „Erstarrens“ interpretiert: „In terms of the PNS-SNS distinction, deception would seem to involve the vegetative, freezing reaction (PNS) rather than the fight-or-flight (SNS) reaction“ (Furedy, 1985, S. 252). Jedoch finden sich auch dort keine expliziten Angaben über das experimentelle Design von Heslegrave und Furedy (1984). Insofern lassen sich allenfalls vage Schlußfolgerungen treffen.

Aus psychophysiologischer Sicht sind die o. g. Annahmen durchaus richtungsweisend. Allgemein wird eine vorübergehende, parasympathisch vermittelte Herzfrequenzverlangsamung in Zusammenhang gebracht mit einer inhibitorischen Verhaltenskontrolle (vgl. Jennings, van der Molen & Brock, 1997, S. 154ff.; van der Veen, van der Molen & Jennings, 2000, S. 611f.), wie sie etwa für Konflikte, „Verstecken“ oder „Erstarren“ typisch sein dürfte. Die Befunde der **vorliegenden Studie** deuten ebenfalls darauf hin, daß man die beim DLT bzw. GAT beobachteten stärkeren Dezelerationen auf die relevanten Fragen und Items im wesentlichen auf eine gesteigerte vagale Erregung zurückführen kann. Gleichermaßen könnten auch die entsprechend schwächeren Akzelerationen bei den tatbezogenen Reizen auf einer erhöhten parasympathischen Aktivität beruhen. Darüber hinaus lag die ΔHR unter der Bedingung Relevant-Täuschung niedriger als unter der Bedingung Relevant-Aufrichtigkeit, was durchaus mit dem o. g. DDP-Ergebnis von Heslegrave (1982) und seiner konflikttheoretischen Interpretation in Einklang zu bringen wäre. Wie bereits erwähnt, fiel aber der Vergleich der Gesamtmittelwerte der beiden Bedingungen über die 8 Sekunden sowohl beim DLT als auch beim GAT nicht signifikant aus. Außerdem erbrachten die genaueren Analysen zu den Effekten der relevanten Stimuli keine signifikanten Interaktionen der beiden Faktoren Tatbedingung und Tatbezug der relevanten Fragen bzw. Items.

Die recht spekulative These, wonach die **Reaktionsunterschiede v. a. parasympathisch vermittelt** sind, läßt sich anhand der vorliegenden Daten nicht ohne weiteres überprüfen. Dazu müßte man neben der HR zusätzlich einen potentiell „reinen“ Sympathikusindikator heranziehen, wie etwa die TWA-Dämpfung. Dieser Parameter wurde jedoch nicht in die Versuchsplanung einbezogen. Ferner muß man berücksichtigen, daß seine Validität im Sinne eines Indikators der Sympathikusaktivität umstritten ist (vgl. Bunnell, 1980, S. 596; Schwartz & Weiss, 1983, S. 700) und daß die TWA bislang überwiegend im Kontext eher tonischer Veränderungen der kardiovaskulären Aktivität untersucht wurde (z. B. Furedy, Szabo & Péronnet, 1996; Myrtek, Hilgenberg, Brügner & Müller, 1997). Folglich würde eine nachträgliche Auswertung der TWA, die aufgrund des vorliegenden EKG-Rohsignals durchaus denkbar wäre, zunächst eine umfangreiche methodenkritische Analyse des Datenmaterials voraussetzen.

6.2.2 HR-Reaktionen nach Ausblendung der Stimuli

In den Grand Averages der HR-Reaktionen nach Ausblendung der Stimuli auf dem Bildschirm zeigte sich für den **DLT und GAT** jeweils in der ersten Poststimulus-Sekunde eine dezelerative Komponente (vgl. Abschnitt 5.2.1: Abbildung 14 und 15). Inwiefern es sich dabei um Reaktionen auf den imperativen Reiz der Antwortgabe und/oder Nachwirkungen der vorhergehenden HR-Abnahme handelte, bleibt zunächst unklar. Diesbezüglich ist zu beachten, daß die initiale Dezeleration nach Ausblendung lediglich in den mittleren HR-Verläufen der relevanten Fragen bzw. Items zu beobachten war (vgl. Abschnitt 5.2.4: Abbildung 22 und 25). Und nur bei diesen Stimuli hatte sich bereits vor der Ausblendung eine dezelerative Komponente manifestiert. Ähnlich wie dort war nach der Ausblendung die HR-Abnahme für die Bedingung Relevant-Täuschung stärker ausgeprägt als für Relevant-Aufrichtigkeit. Im Anschluß an die initiale Dezeleration folgte in den Grand Averages jeweils ein Anstieg der Schlagfrequenz in Relation zur Prästimulus-Baseline. Diese akzelerative Komponente dominierte den Verlauf innerhalb des achtsekündigen Zeitfensters, das zur weiteren Auswertung herangezogen wurde.

Eine **ähnliche Reaktionsmorphologie** beobachteten van Olst, Heemstra und ten Kortenaar (1979, S. 536ff.) in einem Reaktionszeit-Experiment, in dem wiederholt zwei unterschiedliche Töne jeweils 5 Sekunden lang dargeboten wurden. Einer der beiden Töne erhielt Signalcharakter durch die Instruktion an die Vpn, auf seine Ausblendung möglichst rasch motorisch zu reagieren. Die dezelerative Komponente, die ca. 2 Sekunden vor Ende des Signalreizes begann und bis kurz danach andauerte, wurde als Anzeichen der Reaktionsvorbereitung erachtet. Diese Sichtweise deckt sich auch partiell mit dem Interpretationsansatz von Bohlin und Kjellberg (1979) zur antizipatorischen Dezeleration im S_1-S_2-Paradigma (s. o.). Der anfänglichen Abnahme der Schlagfrequenz folgte eine längerfristige Akzeleration, die van Olst et al. als Begleiterscheinung der motorischen Reaktion interpretierten. Auch nach Jennings et al. (1997, S. 163) soll die Herzfrequenzbeschleunigung im Anschluß an einen imperativen Reiz mit der Durchführung der geforderten Handlung assoziiert sein. Analog dazu wäre die in der vorliegenden Studie beobachtete akzelerative Komponente nach Ausblendung als Korrelat der Antwort aufzufassen.

Betrachtet man die **Reaktionen über die 8 Sekunden** hinweg, so fällt beim DLT und GAT auf, daß die phasische HR der Bedingung Relevant-Täuschung klar unterhalb der Verläufe der restlichen Fragen- bzw. Itemtypen lag. Darüber hinaus handelte es sich um die einzige Bedingung, deren Gesamtmittelwerte jeweils negativ ausfielen. Im Vergleich zu den Reaktionen nach Einblendung deutete sich außerdem eine **ausgeprägtere**

Differenzierung zwischen den unterschiedlichen Stimulusarten an. Für beide Testverfahren waren die Haupteffekte des Fragen- bzw. Itemtyps nach Ausblendung stärker als nach Einblendung (aus η^2 bestimmte Effektgröße *f*: DLT: 0.5 vs. 0.45, GAT: 0.58 vs. 0.42). Im Gegensatz zu den Reaktionen nach Einblendung erbrachten auch die Einzelvergleiche der Fragen- bzw. Itemtypen signifikante Mittelwertsunterschiede zwischen Relevant-Täuschung und Relevant-Aufrichtigkeit. Ferner ergaben die ANOVAs zu den Effekten der tatbezogenen Stimuli signifikante Wechselwirkungen der Faktoren Tatbedingung und Tatbezug der relevanten Fragen bzw. Items.

Neben der statistisch bedeutsamen Differenz zwischen den beiden tatbezogenen Fragentypen zeigten die weiteren **Einzelvergleiche zum DLT**, daß der Gesamtdurchschnitt von Relevant-Täuschung signifikant unterhalb der restlichen Fragentypen lag. Der Mittelwert von Relevant-Aufrichtigkeit war zwar auch kleiner als die entsprechenden Parameter der nicht tatbezogenen Fragen, aber nur der Vergleich zu den Lügen-Kontrollfragen erreichte das Signifikanzkriterium. Der Mittelwert der Bedingung Lügen-Kontroll lag insgesamt am höchsten, differierte aber nicht statistisch bedeutsam von Wahrheit-Kontroll und den irrelevanten Fragen.

Die **paarweisen Vergleiche beim GAT** erbrachten ebenfalls einen signifikant niedrigeren Mittelwert von Relevant-Täuschung in Relation zu den restlichen Itemtypen. Darüber hinaus erwies sich die durchschnittliche ΔHR der Bedingung Relevant-Aufrichtigkeit als statistisch bedeutsam kleiner als die von Irrelevant-Vergleich. Dazwischen lag der Mittelwert der Irrelevant-1-Items, ohne daß er sich überzufällig von Relevant-Aufrichtigkeit und Irrelevant-Vergleich unterschied.

Gödert et al. (2001, S. 67ff.) fanden in ihrer DDP-Studie eine **kardiovaskuläre Differenzierung von Täuschung und Aufrichtigkeit** in einer akzelerativen Komponente nach Ausblendung der Fragen auf dem Bildschirm, ohne daß sich vorher in den HR-Reaktionen nach Einblendung signifikante Effekte des Wahrheitsgehalts manifestiert hatten. Nach der Ausblendung, die dort ebenfalls als imperativer Reiz der Antwortgabe diente, war der Anstieg der Schlagfrequenz unter der Bedingung Aufrichtigkeit höher als unter der Bedingung Täuschung. Diese Befunde sind in mehrfacher Hinsicht mit den vorliegenden Ergebnissen zum DLT und GAT vergleichbar. Zum einen lag auch hier die ΔHR von Relevant-Täuschung unterhalb der von Relevant-Aufrichtigkeit. Zum anderen war der Mittelwertsunterschied zwischen den beiden Bedingungen nur bei den Reaktionen nach Ausblendung signifikant, nicht jedoch nach Einblendung. Diese Befunde sind ebenfalls partiell in Einklang zu bringen mit den o. g. Analogstudien von Podlesny und Raskin (1979) und Raskin und Hare (1978), die erst nach dem durch-

schnittlichen Antwortzeitpunkt HR-Reaktionsunterschiede zwischen den relevanten Fragen und Kontrollfragen des KFT fanden.

Mit Verweis auf die hypothetische **Unterscheidung zwischen Täuschungsabsicht und -handlung** diskutierten Gödert et al. (2001, S. 71f.) ihre Ergebnisse dahingehend, daß das kardiovaskuläre DDP-Phänomen möglicherweise ein psychophysiologisches Korrelat der Täuschungshandlung darstelle, da es erst nach der Ausblendung auftrat, also im zeitlichen Bereich der Antwortgabe. Diese Interpretation ist kompatibel mit den o. g. Annahmen, daß die akzelerative Komponente eine Begleiterscheinung der Antwort darstellt und die HR-Reaktionsunterschiede zwischen Täuschung und Aufrichtigkeit v. a. vagal vermittelt werden. Da die Latenz der parasympathischen Innervation des Herzens weniger als eine Sekunde beträgt (vgl. Abschnitt 3.2.2), hätte das DDP-Phänomen früher auftreten müssen, sofern man es auf intentionale Prozesse zurückführen wollte. Denn es war anzunehmen, daß die Entscheidung, zu lügen oder die Wahrheit zu sagen, bereits im Zeitfenster nach Einblendung und nicht erst nach dem imperativen Reiz der Antwortgabe getroffen wurde. Die Vermutung, daß die schwächere HR-Akzeleration auf parasympathischen Einflüssen basierte, begründeten Gödert et al. (2001, S. 72) damit, daß eine erhöhte vagale Erregung einen inhibitorischen Effekt auf die herzfrequenzbeschleunigende Wirkung des Sympathikus haben kann. Andererseits könnte man ebenso gut argumentieren, daß die akzelerative Komponente auf eine Reduktion der vagalen Hemmung zurückzuführen ist (s. o.). Somit käme auch eine *geringere* Deaktivierung des Parasympathikus als Ursache für den niedrigeren HR-Anstieg unter der Bedingung Täuschung in Betracht.

Übertragen auf die vorliegende Studie würde dies bedeuten, daß die HR-Reaktionsunterschiede zwischen Relevant-Täuschung und Relevant-Aufrichtigkeit zumindest teilweise auf dem Wahrheitsgehalt der Antworten und auf einer differentiellen parasympathischen Erregung beruhen. Solche Schlußfolgerungen sind aber nur unter Vorbehalt möglich, zumal sie nicht ohne weiteres die Reaktionsunterschiede zwischen den relevanten und nicht tatbezogenen Stimuli erklären können. Außerdem sind die Implikationen des DDP für die Befragungstechniken der psychophysiologischen Aussagebeurteilung sehr begrenzt. So zeigte sich etwa – entgegen den Erwartungen des DDP – keine signifikante Differenzierung zwischen den Wahrheit- und Lügen-Kontrollfragen des DLT. Dieser Befund ging aber wiederum konform mit den SCR-Daten, in denen ebenfalls keine entsprechenden Unterschiede gefunden wurden.

Darüber hinaus konnte man weitere deutliche **Parallelen zwischen den kardiovaskulären und elektrodermalen Reaktionen** erkennen, allerdings unter „umgekehrten Vorzeichen“. D. h., bei den SCRs bestand die relevante Reaktionsrichtung in einer

höheren Amplitude und bei der HR in einem geringeren Anstieg bzw. einer stärkeren Dezeleration. Die elektrodermalen Reaktionsunterschiede wurden auf eine differentielle sympathische Erregung zurückgeführt, andererseits sollten die unterschiedlichen HR-Verläufe v. a. durch parasympathische Einflüsse bedingt sein. Ferner war die Differenzierung zwischen den Fragen- bzw. Itemtypen bei den SCRs nach Einblendung stärker, bei der HR hingegen nach Ausblendung. In den jeweiligen Zeitfenstern zeigte sich aber sowohl für die elektrodermalen als auch die kardiovaskulären Daten, daß die Reaktionen auf die Lügen-Kontrollfragen des DLT bzw. die Irrelevant-Vergleich-Items des GAT nicht an die Reaktionen der tatbezogenen Stimuli heranreichten. Im Vergleich dazu gingen sogar die Reaktionen auf die irrelevanten Fragen bzw. Irrelevant-1-Items eher in Richtung der relevanten Stimuli. Dies deutet darauf hin, daß auch im Hinblick auf die HR die Lügen-Kontrollfragen des DLT keine angemessene Vergleichsbedingung für die wahrheitsgemäß beantworteten tatbezogene Fragen darstellen und daß beim GAT bereits der erkannte Tatbezug der Relevant-Aufrichtigkeit-Items Reaktionsunterschiede zu Irrelevant-Vergleich bewirkt.

6.3 Subjektive Reaktions- und Bedeutsamkeitsunterschiede zwischen den Fragen- bzw. Itemtypen

Die Ergebnisse der erlebnisdeskriptiven abhängigen Variablen sprechen zugunsten der Annahme, daß sich die Reaktionsunterschiede im subjektiven Erleben widerspiegeln. Beim DLT und GAT fand man in den **Ratings der Reaktionsstärke** signifikante Haupteffekte des Fragen- bzw. Itemtyps. Für beide Testverfahren berichteten die Pbn intensivere körperliche Veränderungen auf die Stimuli der Bedingung Relevant-Täuschung in Relation zu Relevant-Aufrichtigkeit. Ferner erbrachten die genaueren Analysen der Effekte der tatbezogenen Reize signifikante disordinale Interaktionen zwischen der Tatbedingung und dem Tatbezug der relevanten Fragen bzw. Items. D. h., sowohl in der Ring- als auch in der Kette-Gruppe gaben die Pbn an, stärker auf die Stimuli des von ihnen begangenen Scheinverbrechens reagiert zu haben, als auf jene Reize, die sich auf die nicht begangene Tat bezogen.

In den Paarvergleichen zum **DLT** lag der Mittelwert von Relevant-Täuschung signifikant höher als die entsprechenden Parameter der restlichen Fragentypen. Die Bedingungen Relevant-Aufrichtigkeit, Lügen-Kontroll und Wahrheit-Kontroll unterschieden sich nicht überzufällig. Lediglich der Durchschnitt der irrelevanten Fragen lag in Relation dazu deutlich niedriger. Insgesamt wurde die Reaktionsstärke nicht besonders hoch eingestuft. Die Mittelwerte der Fragentypen schwankten ungefähr zwischen den Ausprä-

gungen *(2) kaum reagiert* (Irrelevant: *M* = 2.22) und *(4) mäßig reagiert* (Relevant-Täuschung: *M* = 4.30).

Beim **GAT** unterschieden sich alle vier Itemtypen signifikant. Für Relevant-Täuschung wurde im Durchschnitt eine höhere Reaktionsintensität berichtet als für Relevant-Aufrichtigkeit, gefolgt von den Irrelevant-1-Items und schließlich den Irrelevant-Vergleich-Items. Letztgenannte hatten nach Einschätzung der Pbn die schwächsten körperlichen Veränderungen evoziert. Auch hier waren die Mittelwerte recht niedrig. Sie lagen im Bereich von ca. *(3) eher schwach reagiert* (Irrelevant-Vergleich: *M* = 2.97) bis *(4) mäßig reagiert* (Relevant-Täuschung: *M* = 4.15).

Bradley und Janisse (1981, S. 310) sowie Horowitz et al. (1997, S. 111) erzielten in ihren Analogstudien für die untersuchten direkten Befragungstechniken insofern **vergleichbare Ergebnisse**, als dort die Selbsteinschätzungen ebenfalls zwischen den unterschiedlichen Fragentypen differenzierten. Schuldige berichteten stärkere Reaktionen auf die tatbezogenen Fragen und Unschuldige auf die Vergleichsfragen. Zumindest letzteres widerspricht jedoch partiell den vorliegenden DLT-Ergebnissen, die keine entsprechenden Unterschiede zwischen Relevant-Aufrichtigkeit und den beiden Arten von Kontrollfragen zeigten. Darüber hinaus fanden die o. g. Autoren deutliche Übereinstimmungen zwischen den physiologischen Reaktionsprofilen und den Reaktionssratings. Derartige Befunde sind auch unter theoretischen Gesichtspunkten interessant. So ging etwa Steller (1987, S. 137) in seinem integrativen Systemmodell der psychophysiologischen Aussagebeurteilung davon aus, daß die interozeptive Wahrnehmung autonomer Erregung und sogar die bloße Antizipation körperlicher Veränderungen bei bestimmten Stimuli die physiologischen Reaktionen im Sinne einer positiven Rückkopplungsschleife verstärken bzw. auslösen können.

Die subjektiven Daten der vorliegenden Studie zeigen ebenfalls **Parallelen, aber auch Diskrepanzen zu den körperlichen Parametern**. Die Gleichartigkeit der Befunde betrifft v. a. die Reaktionsunterschiede zwischen Relevant-Täuschung und Relevant-Aufrichtigkeit, die mangelnden Unterschiede zwischen den Lügen- und Wahrheit-Kontrollfragen des DLT sowie die relativ schwachen Reaktionen auf die Irrelevant-Vergleich-Items des GAT. Abweichungen lassen sich insbesondere bei den irrelevanten Fragen des DLT und den Irrelevant-1-Items des GAT konstatieren. Während auf die irrelevanten Fragen des DLT ähnlich bzw. sogar mitunter etwas stärker reagiert wurde als auf die Kontrollfragen, stuften die Pbn die entsprechenden Reaktionen am niedrigsten ein. Und obwohl die Reaktionsintensität der Irrelevant-1-Items des GAT speziell in den SCRs nach Ein- und Ausblendung und der HR nach Ausblendung an die relevanten Items heranreichte, wurde sie als geringer eingestuft.

Dies deutet darauf hin, daß eventuell **Antworttendenzen**, wie etwa die soziale Erwünschtheit, einen Einfluß auf die Selbsteinschätzungen hatten (vgl. Heidenreich, 1984, S. 415f.; Petermann & Noack, 1984, S. 452). Unter diesen Voraussetzungen müßte man davon ausgehen, daß die Vpn Hypothesen hinsichtlich des vermeintlich von ihnen erwarteten Reaktionsprofils generierten und sich diese Erwartungen in ihren Einschätzungen niederschlugen. Demnach taxierten die Vpn ihre körperlichen Veränderungen nach Darbietung der irrelevanten Fragen des DLT bzw. der Irrelevant-1-Items des GAT eventuell deswegen relativ niedrig, weil die Logik der Verfahren, wie sie teilweise auch in den Instruktionen kolportiert wurde, eher schwache Reaktionen auf neutrale, d. h. nicht tatbezogene, wahrheitsgemäß beantwortete Stimuli nahelegte. Dennoch konnten die betreffenden Reize starke körperliche Reaktionen auslösen. Möglicherweise weil die irrelevanten Fragen die einzigen waren, die beim DLT bejaht wurden, bzw. die Irrelevant-1-Items des GAT unmittelbar im Anschluß an die Multiple-Choice-Fragen dargeboten wurden. In diesem Zusammenhang ist auch zu bedenken, daß die Pbn ihre Einschätzungen retrospektiv treffen mußten, was etwa aufgrund von Gedächtniseffekten ebenfalls zu Verfälschungen führen konnte.

Bei der Diskussion der elektrodermalen und kardiovaskulären Reaktionsunterschiede wurde bereits mehrfach auf Annahmen zur Orientierungsaktivität verwiesen. Diese betrafen v. a. das Konzept der OR und ihre Abhängigkeit von der Reizbedeutung. Ein grundlegendes Problem der Signifikanzhypothese zur OR-Auslösung besteht darin, daß die **Bedeutsamkeit der Stimuli** oft nicht unabhängig von den Reaktionen definiert bzw. operationalisiert wird (O'Gorman, 1979, S. 257f.). Diesem Kritikpunkt wurde im Rahmen der eigenen Studie insofern Rechnung getragen, als man die subjektiv eingeschätzte Relevanz der Fragen bzw. Items für das Testergebnis und für die Glaubwürdigkeitsbeurteilung zusätzlich durch Ratings erfaßte. Außerdem wollte man damit überprüfen, inwiefern die von den DLT-Instruktionen intendierte manipulative Erhöhung der Bedeutsamkeit der Lügen-Kontrollfragen gelungen war.

Die **subjektive Beurteilung der Wichtigkeit** der unterschiedlichen Stimulustypen erbrachte sowohl für den DLT als auch für den GAT signifikante Haupteffekte. Die Fragen bzw. Items der Bedingung Relevant-Täuschung wurden bedeutsamer eingestuft als die der Bedingung Relevant-Aufrichtigkeit. Im Einklang damit ergaben die Analysen der Effekte der tatbezogenen Stimuli signifikante disordinale Wechselwirkungen Tatbedingung × Tatbezug der relevanten Fragen bzw. Items.

Die Mittelwerte der beiden relevanten Fragentypen des **DLT** lagen jeweils höher als die der Lügen-Kontroll-, Wahrheit-Kontroll- und irrelevanten Fragen. Der paarweise Vergleich der beiden Arten von Kontrollfragen erbrachte keinen signifikanten Unterschied.

Lediglich die Bedeutung der irrelevanten Fragen wurde niedriger eingeschätzt. Ihr Mittelwert (*M* = 3.54) lag in etwa zwischen den Ausprägungen *(3) eher wenig* und *(4) mäßig wichtig*. Im Vergleich dazu entsprach der Durchschnitt von Relevant-Täuschung (*M* = 5.79) annähernd der Stufe *(6) sehr wichtig*.

Beim **GAT** wurden die wahrheitswidrig und wahrheitsgemäß beantworteten relevanten Items ebenfalls als bedeutsamer eingestuft als die irrelevanten Itemtypen. Letztere unterschieden sich nicht überzufällig. Ihre beiden Mittelwerte lagen im Bereich von *(3) eher wenig wichtig*. Der Durchschnitt von Relevant-Täuschung (*M* = 5.37) hingegen entsprach etwa der Ausprägung *(5) eher wichtig*.

Mehrere Befunde sind **besonders diskussionswürdig**. Einerseits entsprechen die Bedeutsamkeitsratings über weite Strecken der Logik der Verfahren. Die relevanten Fragen bzw. Items (v. a. Relevant-Täuschung) wurden als wichtiger für das Testergebnis erachtet als die nicht tatbezogenen. Und auch die Signifikanz der irrelevanten Fragen des DLT bzw. Irrelevant-1-Items des GAT wurde als niedrig eingestuft, was durchaus ihrem objektiven Stellenwert entspricht, zumal die Reaktionen darauf normalerweise nicht in die Auswertung der Testergebnisse eingehen. Es fällt aber andererseits auf, daß die Lügen- und Wahrheit-Kontrollfragen des **DLT** zwar subjektiv wichtiger waren als die irrelevanten Fragen, sie blieben aber deutlich unterhalb der Bedingung Relevant-Aufrichtigkeit. Darüber hinaus zeigten sich keine Unterschiede zwischen den beiden Kontrollfragentypen. Dies spricht dafür, daß die Versuche, ihre subjektiv eingeschätzte Bedeutsamkeit durch standardisierte Instruktionen manipulativ zu erhöhen, nicht hinreichend effektiv waren. Möglicherweise ist darin einer der Gründe zu sehen, weshalb die Lügen-Kontrollfragen entgegen den Erwartungen des DLT keine stärkeren Reaktionen auslösten als die Relevant-Aufrichtigkeit-Fragen. Die subjektiven Daten stehen nämlich im Einklang mit den körperlichen Variablen, die ebenfalls schwache Reaktionen auf die Lügen-Kontrollfragen und keine Unterschiede zu den Wahrheit-Kontrollfragen ergaben. Gleichzeitig waren beim **GAT** die Relevant-Aufrichtigkeit-Items nach Ansicht der Pbn bedeutender als die Irrelevant-Vergleich-Items, was sich wiederum weitgehend mit den körperlichen Reaktionsverhältnissen deckte.

Auch der **Unterschied zwischen den beiden Arten von tatbezogenen Stimuli** birgt gewisse Implikationen für die Interpretation der physiologischen Daten. Jene relevanten Fragen bzw. Items, die sich auf das begangene Scheinverbrechen bezogen, wurden als wichtiger für das Testergebnis bewertet. Dies kann als Indiz dafür gelten, daß man die körperlichen Reaktionsunterschiede zwischen Relevant-Täuschung und Relevant-Aufrichtigkeit nicht nur auf die Variation des Wahrheitsgehalts zurückführen kann. Statt dessen scheinen Faktoren, die unabhängig von den Antworten sind, wie etwa der sub-

jektiv eingeschätzte Stellenwert der Fragen bzw. Items für die Glaubwürdigkeitsbeurteilung, zur Differenzierung beizutragen.

Wie bereits angedeutet, hat die unterschiedlich eingestufte Wichtigkeit der Fragen- bzw. Itemtypen für das Testergebnis einen gewissen **Erklärungswert** im Hinblick auf die OR-Theorie der psychophysiologischen Aussagebeurteilung. Dabei werden die Reaktionsunterschiede damit begründet, daß subjektiv bedeutsamere Reize stärkere ORs auslösen. Allerdings erfaßt das hier verwendete Rating zur Relevanz der Stimuli für das Testergebnis wohl nur einen Teilaspekt des viel breiter gefaßten Konzepts der Reizsignifikanz im Kontext der OR (vgl. z. B. van Olst et al., 1979, S. 521ff.). Außerdem sind auch bei diesem Selbstbeurteilungsverfahren typische Antworttendenzen, wie z. B. soziale Erwünschtheit, nicht ohne weiteres auszuschließen.

6.4 Einflüsse der elektrodermalen Labilität

Die **Klassifikation** in elektrodermal labile und stabile Pbn wurde per Mediansplit der interindividuellen Häufigkeiten elektrodermaler Spontanfluktuationen (NSRs) während einer reizfreien Ruhemessung getroffen. Der für die $N = 80$ Vpn resultierende Median von 29 liegt in einem Wertebereich, der als typisch für Messungen unter ähnlichen Rahmenbedingungen gilt (z. B. männliche Stichprobe, fünfminütige Ruhemessung, Amplitudenkriterium von 0.02 µS; vgl. Vossel & Zimmer, 1990, S. 71f.). In bezug auf die Gruppeneinteilung ist das vorliegende Experiment also durchaus mit anderen Studien vergleichbar.

Generell erwartet man in Untersuchungen zur SCR deutliche **Labilitätseffekte**, wobei Labile insgesamt stärker reagieren sollen als Stabile. Die differentielle elektrodermale Reaktivität auf die wiederholte Darbietung einfacher Reize ist bereits vielfach dokumentiert (z. B. Bull & Gale, 1973, S. 303f.; Johnson, 1963, S. 420; Schell et al., 1988, S. 624). Als mögliche Erklärung werden gemeinhin Unterschiede in der zentralen Informationsverarbeitung angenommen. Labile zeichnen sich demnach durch eine stärkere Orientierungsaktivität aus (Vossel, 1990, S. 229). Sie widmen den Stimuli mehr Aufmerksamkeit bzw. Verarbeitungskapazität („information processing capacity“, Schell et al., 1988, S. 630) und halten diese Zuwendung über einen längeren Zeitraum aufrecht, so daß die Reize im stärkeren Maße und über eine höhere Anzahl von Darbietungen hinweg ORs auslösen können. Vornehmlich auf bedeutungsvolle Stimuli zeigen Labile stärkere SCRs und eine langsamere Habituation. Ausgehend von der „Signifikanzhypothese“ der OR (Bernstein, 1979, S. 263) kann man argumentieren, daß Labile dazu tendieren, wiederholt dargebotene Reize über einen längeren Zeitraum als potenti-

ell bedeutsam einzuschätzen. Folglich sollten sie auch noch auf spätere Reizdarbietungen mit stärkeren ORs reagieren

Im eigenen Experiment hatten elektrodermal Labile beim **DLT und GAT** höhere SCR-Magnituden als Stabile. Dies galt gleichermaßen für die Reaktionen nach Ein- und Ausblendung der Fragen bzw. Items. Die Labilitätseffekte waren jedoch nur beim GAT konsistent signifikant. Beim DLT fielen die Unterschiede nicht statistisch bedeutsam aus (SCRs nach Einblendung), oder sie verfehlten knapp das Signifikanzniveau (SCRs nach Ausblendung).

Diese Diskrepanzen sind nicht unbedingt auf die Befragungstechniken zurückzuführen, sondern eventuell auf **Merkmale der Teilstichproben**. Betrachtet man die NSR-Häufigkeiten elektrodermal stabiler und labiler Pbn in den beiden Testbedingungen getrennt voneinander, so fällt auf, daß die Mittelwerte beim GAT weiter auseinanderliegen als beim DLT (vgl. Anhang B: Tabelle 28). Außerdem sind unter der Bedingung DLT die Verteilungen beider Labilitätsgruppen in Richtung des Medians geneigt (Stabile: leichte Rechtsschiefe; Labile: deutliche Linksschiefe; GAT: jeweils leichte Linksschiefe). Dies weist darauf hin, daß beim DLT elektrodermal Labile und Stabile weniger klar voneinander abgegrenzt waren. D. h., es gab dort mehr Personen, deren NSR-Anzahl im Bereich des Grenzwertes (Median) lag. Dadurch wurden Unterschiede zwischen den Gruppen nivelliert, was möglicherweise dazu geführt hat, daß beim DLT die Labilitätseffekte auf die SCRs nicht das Signifikanzniveau erreichten. Dennoch unterschieden sich die Stabilen und Labilen unter dieser Testbedingung gemäß den Erwartungen in ihren SCR-Magnituden (s. o.). Somit mußte man nicht davon ausgehen, daß die Gruppentrennung gänzlich ineffektiv war.

Konkret impliziert ein Einfluß der EL auf die Trefferquoten der psychophysiologischen Aussagebeurteilung (vgl. Abschnitt 3.4) eine **Wechselwirkung zwischen dem Personenmerkmal und dem Fragen- bzw. Itemtyp** (vgl. Abschnitt 3.5.2). Der einzige richtungsweisende Effekt, der auf entsprechende Unterschiede in den Reaktionsprofilen hindeutete, resultierte beim GAT. Dort fand man in den logarithmierten SCRs mit einer Latenz von 1 – 3 Sekunden nach Einblendung eine signifikante Wechselwirkung EL × Itemtyp. Diese basierte jedoch im wesentlichen auf den Irrelevant-1-Items, die bei der Glaubwürdigkeitsbeurteilung normalerweise unberücksichtigt bleiben. Darüber hinaus konnte die Interaktion nicht über die unterschiedlichen Auswertungen hinweg repliziert werden. Für die nicht transformierten SCRs bzw. bei der Amplitudenbestimmung im breiten Latenzzeitfenster (1 – 10 Sekunden) erreichte sie nicht das Signifikanzkriterium. Ansonsten manifestierten sich in den elektrodermalen, kardiovaskulären und subjektiven Daten nur einfache Interaktionen zwischen der EL und der Tatbedingung (vgl.

DLT: HR nach Ein- und Ausblendung; GAT: Reaktionsstärkeratings) oder komplexe Wechselwirkungen der beiden Gruppenfaktoren mit dem Fragen- bzw. Itemtyp (DLT: SCRs nach Ausblendung, Bedeutsamkeitsratings). Diese waren entweder weitgehend auf die irrelevanten Fragen zurückzuführen (DLT: Bedeutsamkeitsratings) oder nicht über verschiedene Auswertungen hinweg stabil (DLT: SCRs nach Ausblendung, logarithmierte vs. nicht logarithmierte Daten); bzw. sie waren nicht ohne weiteres zu erklären (wie etwa die o. g. dreifaktoriellen Interaktionen) oder im Zusammenhang mit den Reaktionsunterschieden auf die Fragen- bzw. Itemtypen belanglos (vgl. die Interaktionen EL × Tatbedingung).

Insgesamt ist somit festzustellen, daß die EL keine relevanten Einflüsse auf die Reaktionsunterschiede zwischen den Fragen- bzw. Itemtypen des DLT und GAT ausübte. Dies galt gleichermaßen für die elektrodermalen, kardiovaskulären und subjektiven Variablen. Der Mangel an eindeutigen Interaktionen widerspricht der Annahme, daß ein Zusammenhang zwischen der EL und der kriterienbezogenen Validität der psychophysiologischen Aussagebeurteilung besteht (Ben-Shakhar & Furedy, 1990, S. 78; Berning, 1992, S. 147). Die Ergebnisse sind jedoch kongruent mit Untersuchungen, die keine entsprechenden Effekte fanden (Steller, 1987, S. 74; Waid, Wilson & Orne, 1981, S. 1122). Diese Befunde bezogen sich jedoch ausschließlich auf den TWT, wohingegen andere Studien speziell für den KFT, aber mitunter auch für den TWT korrelative Beziehungen zwischen der EL und den Trefferquoten nachweisen konnten (Horneman & O'Gorman, 1987, S. 329; Waid & Orne, 1981, S. 88f.; Waid, Wilson & Orne, 1981, S. 1122).

Die **Gründe für die Abweichungen** zu früheren Arbeiten sind v. a. im Hinblick auf die direkten Befragungstechniken zunächst unklar. Eine mögliche Erklärung bestünde darin, daß konträr zu den konventionellen Methoden (speziell dem KFT) die EL bei den neueren Befragungstechniken der forensischen Psychophysiologie eine untergeordnete Rolle spielt. Abgesehen davon sind auch andere Erklärungen denkbar. Auf die Eventualität einer ineffektiven Operationalisierung der EL beim DLT wurde bereits hingewiesen. Dabei ist auch zu berücksichtigen, daß die eigene Arbeit im Gegensatz zu anderen Studien (Gödert et al., 2001, S. 67ff.; Schell et al., 1988, S. 624) keine eindeutigen Hinweise auf konsistente HR-Reaktionsunterschiede zwischen elektrodermal Labilen und Stabilen erbrachte. Dies galt gleichermaßen für den GAT, obwohl dort die signifikanten Labilitätseffekte auf die SCRs zugunsten einer effektiven Trennung von Labilen und Stabilen sprachen. Die beiden o. g. Studien wurden allerdings nicht im Zusammenhang mit Befragungstechniken der psychophysiologischen Aussagebeurteilung durchgeführt, insofern sind abweichende Ergebnisse nicht unbedingt verwunderlich. Und das DDP-Experiment von Gödert et al. ergab trotz (teils auch marginal) signifikanter Labi-

litätseffekte in den dezelerativen HR-Komponenten keinen moderierenden Einfluß der EL auf die Stärke des kardiovaskulären DDP-Phänomens. Ebenso fand man dort in den elektrodermalen und subjektiven Variablen keine direkten Wechselwirkungen der EL mit dem Wahrheitsgehalt. Diese Befunde sind durchaus mit dem Mangel an eindeutigen Interaktionen zwischen der EL und dem Fragen- bzw. Itemtyp in der vorliegenden Untersuchung vergleichbar, wenngleich eine Gegenüberstellung der jeweiligen Ansätze nur unter Vorbehalt gestattet ist.

In der o. g. **DDP-Studie** wurden außerdem Labilitätseffekte in den Kontrollvariablen der Nachbefragung gefunden (vgl. dazu Rill, 1997, S. 133ff.). Signifikant mehr Stabile als Labile äußerten, Manipulationsversuche unternommen zu haben, um die vermeintliche „Lügendetektion“ zu erschweren. Ferner berichteten mehr Labile als Stabile von verschiedenartigen Empfindungen, Gefühlen oder körperlichen Reaktionen beim Lügen und Wahrheitsagen. Diese Befunde wurden mit Verweis auf mögliche Persönlichkeitskorrelate der EL und potentielle Unterschiede in der Wahrnehmungssensitivität diskutiert (siehe auch Crider, 1993, S. 176ff.; Munro, Dawson, Schell & Sakai, 1987, S. 255f.). Obwohl die Nachbefragung der vorliegenden Studie methodisch stark daran angelehnt war, konnten die Ergebnisse nicht repliziert werden. Inwiefern dafür Diskrepanzen zwischen den Paradigmen oder Probleme der Gruppenaufteilung verantwortlich sind, bleibt offen. Man muß aber auch beachten, daß v. a. die hypothetischen Persönlichkeitsunterschiede elektrodermal labiler und stabiler Personen äußerst vage und empirisch wenig fundiert sind. Im Hinblick auf eine Reihe von Merkmalen (z. B. Ängstlichkeit, Neurotizismus und Extraversion) konnten bislang keine Persönlichkeitskorrelate der EL konsistent nachgewiesen werden; dies betrifft speziell psychiatrisch unauffällige Stichproben (Vossel, 1990, S. 196).

Beziehungen der EL zu komplexeren Persönlichkeitsmerkmalen gelten auch als wenig plausibel, da letztere wohl multifaktoriell bedingt sind und eine Vielzahl von Verhaltensdipositionen, Überzeugungen bzw. Einstellungen subsumieren (vgl. auch die methodischen Bedenken, die O’Gorman, 1983, S. 441ff., gegenüber potentiellen Persönlichkeitskorrelaten der Orientierungsaktivität vorbringt). Doch selbst wenn man wie etwa Crider (1993, 177ff.) davon ausgeht, daß ein Zusammenhang besteht zwischen der EL und der Persönlichkeitsdimension Verträglichkeit vs. Antagonismus („agreeableness vs. antagonism“, vgl. Costa & McCrea, 1992a, b, c) und somit im weitesten Sinne mit Soziopathie bzw. Psychopathie (McCrea & Costa, 1987, S. 88), so ergeben sich daraus nicht notwendigerweise Effekte auf die Trefferquoten der psychophysiologischen Aussagebeurteilung oder auf die Reaktionsunterschiede zwischen relevanten vs. nicht tatbezogenen bzw. wahrheitsgemäß vs. wahrheitswidrig beantworteten Stimuli. Barland und Raskin (1975b, S. 224), Raskin und Hare (1978, S. 134) sowie Patrick und Iacono

(1989, S. 353) fanden unter Verwendung des KFT keine niedrigeren Trefferquoten für Psychopathen im Vergleich zu Pbn ohne entsprechende Diagnose. Balloun und Holmes (1979, S. 321) konnten ebenfalls keinen Einfluß der Psychopathie auf die Validität des TWT nachweisen. „Psychopathen“ und „Nicht-Psychopathen“ dürften sich demnach kaum in der Differenz der physiologischen Reaktionsstärke auf relevante Fragen und Kontrollfragen bzw. relevante und irrelevante Items unterscheiden.

6.5 Schlußfolgerungen und Ausblick

Im folgenden werden die im Abschnitt 3.5 formulierten Forschungsfragen nochmals gezielt angesprochen und die zugehörigen Befunde integrativ diskutiert. In bezug auf die **zentrale Problemstellung** der vorliegenden Untersuchung gilt es zunächst hervorzuheben, daß man für die neueren Befragungstechniken DLT und GAT anhand eines Scheinverbrechen-Experiments mit standardisierter Durchführung und objektiver Auswertung der psychophysiologischen Aussagebeurteilung quantitative Reaktionsunterschiede zwischen den Fragen- bzw. Itemtypen nachweisen konnte.

Bemerkenswert ist, daß die Effekte unter **simulierten Tat- und Testbedingungen** auftraten. Dies spricht zugunsten der Realisierung des experimentellen Paradigmas. Auch die postexperimentellen Ratings und die Nachbefragung legten nahe, daß es gelungen war, eine relativ hohe Täuschungsmotivation und subjektive Treffsicherheit zu induzieren. So gaben die Pbn im Durchschnitt an, deutlich motiviert gewesen zu sein, die Lügendetektion zu erschweren. Dies deckt sich mit den Angaben der Nachbefragungen, in denen die Teilnehmer mehrheitlich von Manipulationsversuchen berichteten. Außerdem schätzten die Vpn, daß es dem Untersucher wiederholt bis öfters gelungen sei, wahrheitsgemäße und wahrheitswidrige Antworten zu differenzieren, was wiederum mit der Häufigkeit von selbstberichteten Reaktionsunterschieden beim Lügen und Wahrheitsagen zu vereinbaren war. Wie jedoch bereits im Abschnitt 2.9.2 dargestellt, eignen sich Laborstudien aufgrund der erheblichen Abweichungen zu Realfällen nicht direkt zur Abschätzung der kriterienbezogenen Validität, sondern allenfalls zur Untersuchung der Grundlagen der psychophysiologischen Aussagebeurteilung. Darum wurden hier im Gegensatz zur konventionellen Methodik keine primär individualdiagnostischen Zielsetzungen verfolgt, sondern die Reaktionen und ihre Bedingungsfaktoren direkt analysiert.

Die **Reaktionsunterschiede zwischen Relevant-Täuschung und Relevant-Aufrichtigkeit** sowie die Interaktionen zwischen der Tatbedingung und dem Tatbezug der relevanten Fragen bzw. Items zeigten, daß die Pbn insgesamt auf jene Stimuli stär-

ker reagierten[11], die das begangene Scheinverbrechen thematisierten. Dies ist mit der allgemeinen Prämisse der „Lügendetektion“ kompatibel, wonach tatbezogene Reize stärkere Reaktionen evozieren, wenn man die Tat verübt hat und sie wahrheitswidrig abstreitet, als wenn man diesbezüglich unschuldig ist und den Tatvorwurf wahrheitsgemäß verneint (vgl. Lykken, 1998, S. 120). Aufgrund der intraindividuellen Bedingungsvariation der beiden Fragen- bzw. Itemtypen konnte diese Hypothese anhand der vorliegenden Studie erstmals stringent für den DLT und GAT bestätigt werden. Inwiefern man die Reaktionsunterschiede im Sinne des DDP singulär auf die Variation von Täuschung und Aufrichtigkeit zurückführen kann, bleibt fraglich. Die Bedeutsamkeitsratings sprechen für die Annahme, daß auch andere Faktoren, die unabhängig vom Wahrheitsgehalt sind, wie etwa die subjektiv eingeschätzte Relevanz der Stimuli für das Testergebnis, dabei eine Rolle spielen. Darüber hinaus dürften Reize, die sich auf eine durchgeführte (Straf-)Tat beziehen, eine subjektiv größere Signifikanz aufweisen als solche, die eine unterlassene Handlung betreffen (Steller & Dahle, 1997, S. 311). Konflikttheoretische Überlegungen und die bisherige Forschung zum Einfluß der Antwortart (vgl. Abschnitt 2.10.1) legen aber die Vermutung nahe, daß die Differenzierung zwischen Relevant-Täuschung und -Aufrichtigkeit zumindest partiell auf dem Wahrheitsgehalt der Antworten beruht.

Im Gegensatz dazu zeigte die entsprechende **Variation bei den Kontrollfragen des DLT** keine gravierenden Effekte. Die Bedingungen Lügen- vs. Wahrheit-Kontroll unterschieden sich weder in den physiologischen noch in den subjektiven Variablen signifikant. Dies paßt einerseits zu dem Befund, daß der Wahrheitsgehalt der Antworten für die Reaktionen auf die Kontrollfragen des KFT nebensächlich ist (Honts et al., 1992, S. 270), stützt aber andererseits ebenso die Kritik am DLT, daß instruierte Falschaussagen nicht unbedingt mit intentionalen Täuschungen vergleichbar sind (vgl. Iacono, 2000, S. 781; Lykken, 1998, S. 137ff.). Unter Umständen basiert die potentiell erhöhte Bedeutsamkeit der Kontrollfragen weniger auf der Notwendigkeit, daß die Pbn darauf lügen müssen, als vielmehr auf den Suggestionen und Manipulationen, die normalerweise im Rahmen des Vortest-Interviews durchgeführt werden (siehe Abschnitt 2.4.1).

Darin ist auch eine der möglichen Ursachen zu sehen, weshalb in der vorliegenden Studie **relativ schwache Reaktionen auf die Lügen-Kontrollfragen** auftraten. Zwar fand man mitunter etwas intensivere Reaktionen auf Lügen-Kontroll in Relation zu Relevant-Aufrichtigkeit (vgl. SCRs nach Ausblendung), wobei der Unterschied statistisch nicht signifikant ausfiel. Im allgemeinen wurde aber unter der Bedingung Relevant-Aufrichtigkeit deutlich stärker reagiert. Dies widerspricht einer Grundannahme der

[11] Im Hinblick auf die Herzschlagfrequenz bedeutet dies eine niedrigere ΔHR.

Kontrollfragentechniken, wonach die Kontrollfragen intensivere körperliche Veränderungen auslösen sollen als wahrheitsgemäß verneinte relevante Fragen (Lykken, 1998, S. 122). Darüber hinaus konfligiert dieser Befund mit den bisherigen DLT-Studien, die auch für Unschuldige relativ hohe Trefferquoten erbrachten (vgl. Abschnitt 2.5). Demnach wären nämlich stärkere Reaktionen auf Lügen-Kontroll im Vergleich zu Relevant-Aufrichtigkeit zu erwarten gewesen. Neben der fundamentalen Kritik, daß Kontrollfragen keine angemessenen Vergleichsreize für die tatbezogenen Fragen darstellen (Furedy, 1996a, S. 100f.), sind auch andere Gründe diskutabel, die eher die spezifische Methodik der vorliegenden Untersuchung betreffen. Die Tatsache, daß alle Pbn hinsichtlich eines Scheinverbrechens „schuldig“ waren, mag insgesamt zu einer erhöhten Bedeutsamkeit der tatbezogenen Fragen und somit von Relevant-Aufrichtigkeit beigetragen haben. Möglicherweise war auch das hohe Maß an Standardisierung für die schwachen Reaktionen auf die Kontrollfragen verantwortlich. Abweichend vom KFT wurde ihre Formulierung nicht individuell an die Pbn angepaßt, und anders als beim konventionellen DLT fand kein Vortest-Interview statt, sondern die Instruktionen wurden schriftlich erteilt. Außerdem gab es im Gegensatz zum üblichen Testablauf keine Pausen zwischen den Fragendurchgängen, in denen der Untersucher durch weitere Gespräche mit dem Pb suggestiv die Signifikanz der Lügen-Kontrollfragen betonen und aufrecht erhalten soll (vgl. Honts & Raskin, 1988, S. 58; Horowitz et al., 1997, S. 111).

Bradley und Black (1998, S. 698) führten in ihrer Analogstudie zum KFT ebenfalls **kein Vortest-Interview** durch. Darüber hinaus verglichen sie anhand einer interindividuellen Bedingungsvariation Standard-Kontrollfragen, die frühere, dem Scheinverbrechen ähnliche Vergehen thematisierten (z. B. Diebstähle), mit speziell auf die studentische Stichprobe zugeschnittenen Kontrollfragen, die etwa Abschreiben oder andere Betrugsversuche bei Klausuren zum Inhalt hatten. Unschuldige reagierten relativ schwach auf die Standard-Kontrollfragen, was die Autoren teils auf das fehlende Vortest-Interview zurückführten. Daraus ergibt sich als mögliche Implikation, daß ein Vortest-Interview und/oder eine individualisierte Fragenformulierung nötig sind, um die Bedeutsamkeit bzw. Bedrohlichkeit der Kontrollfragen zu steigern. In ähnlicher Weise hat bereits Raskin (1979, S. 590) argumentiert, daß der erhöhte Signalwert der Kontrollfragen v. a. auf den Suggestionen und Manipulationen im Vortest-Interview beruhen soll. Demnach würde die potentielle Standardisierbarkeit, die als ein wesentlicher Vorteil des DLT gilt, an Grenzen stoßen. Und der bereits im Zusammenhang mit dem KFT (Abschnitt 2.4.2) angesprochene Konflikt zwischen einer standardisierten Testkonstruktion und der Notwendigkeit, das Vortest-Interview bzw. die Fragenformulierung individuell an den jeweiligen Einzelfall anpassen zu müssen, wäre auch mit der neuen Befragungstechnik nicht zu lösen.

An dieser Stelle sei aber nochmals betont, daß die **Generalisierbarkeit** der Ergebnisse von Analogstudien im allgemeinen und der vorliegenden Befunde im speziellen sehr eingeschränkt ist. Abgesehen von den bereits erwähnten Abweichungen gegenüber dem konventionellen DLT (z. B. Variationen des Wahrheitsgehalts der Antworten, keine Vortest-Interviews bzw. Pausen zwischen den Fragendurchgängen) wurde die Abfolge der unterschiedlichen Fragentypen in der vorliegenden Arbeit durchmischt, anstatt – wie bei den Kontrollfragentechniken üblich – die Kontrollfragen stets vor den jeweiligen relevanten Fragen zu präsentieren (Raskin, 1979, S. 595). Aufgrund der zu erwartenden Habituationseffekte sollte letztere Vorgehensweise tendenziell zu einer Erhöhung der Reaktionen auf die Kontrollfragen im Vergleich zu den relevanten Fragen beitragen, wohingegen ein Wechsel der Abfolge solche Effekte eher unterdrücken dürfte. Darin wäre eine weitere Erklärung für die relativ schwachen Reaktionen auf die (Lügen- und Wahrheit-) Kontrollfragen zu sehen.

Hinsichtlich der **irrelevanten Stimuli** wurde bereits erwähnt, daß die unerwartet starken Reaktionen auf die wahrheitsgemäß beantworteten neutralen Fragen des DLT eventuell darauf zurückzuführen waren, daß sie als einzige bejaht wurden. Indes war die relativ hohe Reaktionsstärke auf die Irrelevant-1-Items des **GAT** hypothesenkonform. Ebenso erwartungsgemäß wurde dort auf die Irrelevant-Vergleich-Items am schwächsten reagiert, wenngleich der Mittelwertsunterschied zu Relevant-Aufrichtigkeit nicht immer signifikant ausfiel (vgl. SCRs nach Ausblendung). Letzterer Befund erweist sich als besonders interessant, da er – zumindest, was die SCRs nach Ausblendung anbelangt – mit der Testlogik des GAT in Einklang zu bringen ist. Demnach müssen wahrheitsgemäß beantwortete relevante Items selbst unter der Voraussetzung, daß man ihren Tatbezug erkennt, nicht zwangsläufig signifikant stärkere Reaktionen auslösen als die irrelevanten Vergleichsreize. Und falls doch (z. B. SCRs nach Einblendung), dann ist der Reaktionsunterschied von Irrelevant-Vergleich zu Relevant-Aufrichtigkeit kleiner als zu Relevant-Täuschung. Damit könnte man unter anderem die Differenzierung zwischen Schuldigen und Unschuldigen mit Tatwissen beim GAT erklären.

In den Reaktionsunterschieden zwischen den Fragen- bzw. Itemtypen wiesen die **elektrodermalen und kardiovaskulären Daten** auffällige Übereinstimmungen auf, allerdings in entgegengesetzte Richtungen (höhere SCRs vs. niedrigere ΔHR, sympathische vs. mutmaßlich parasympathische Vermittlung, stärkere Differenzierung nach Ein- vs. Ausblendung). Trotz der nicht zu vernachlässigenden Diskrepanzen legt die Gleichartigkeit der elektrodermalen und kardiovaskulären Befunde nahe, daß teils identische psychophysiologische Prozesse darin involviert sind. Ähnlich argumentierte bereits Heslegrave (1982), daß sowohl die stärkere parasympathische Aktivität in bezug auf die HR als auch die sympathisch vermittelten höheren SCRs unter Täuschungsbedingungen

auf den zugrundeliegenden Konflikten basieren würden. Eine mögliche Erklärung für diese zunächst paradox anmutende Hypothese bestünde darin, daß beide Variablen zwar peripherphysiologisch divergent gesteuert werden, gleichzeitig aber den synchronisierenden Effekten komplexer zentralnervöser Modulations- und Mediationsprozesse durch übergeordnete Kontrollinstanzen unterliegen (vgl. etwa Bernstein et al., 1975, S. 167f.; Larsen, Schneiderman & Pasin, 1986, S. 124ff.; Vossel, 1990, S. 50f.; Zimmer et al., 1989, S. 267). Auch Furedy (1985) versuchte den „Widerspruch“ zwischen den SCR- und HR-Daten im Zusammenhang mit den psychophysiologischen Korrelaten der Täuschung damit begreiflich zu machen, daß die elektrische Hautleitfähigkeit nur peripher sympathisch gesteuert werde, während zentralnervös parasympathische Einflüsse dominieren sollen: „SCR may be sympathetic only in the periphery, with PNS influences predominating in the central pathways“ (S. 252). Welche konkreten Prozesse damit gemeint sind, bleibt offen, da sie von Furedy nicht weiter spezifiziert werden.

Die **Frage nach den psychophysiologischen Grundlagen der Reaktionsunterschiede** läßt sich hier nicht definitiv beantworten. Die o. g. Ergebnisse zur Psychophysiologie der Täuschung und die konflikttheoretischen Annahmen, die v. a. im Kontext des DDP diskutiert werden, können nur bedingt auf die vorliegende Studie übertragen werden. Sie geben allenfalls Aufschluß darüber, inwiefern der Wahrheitsgehalt der Antworten zu einer Differenzierung zwischen Relevant-Täuschung und Relevant-Aufrichtigkeit beiträgt. Der innerhalb der psychophysiologischen Aussageforschung dominierende Ansatz, der allgemein zur Interpretation der beobachteten Reaktionsunterschiede zwischen den Fragen bzw. Items herangezogen wird, ist das Konzept der OR und deren Abhängigkeit von der Reizbedeutung. Dieser theoretische Bezugsrahmen kann zwar zur Klärung einiger Phänomene beitragen, birgt aber zusätzliche Ungereimtheiten, auf die an entsprechender Stelle bereits hingewiesen wurde. Eine mögliche Ursache für die Interpretationsprobleme mit der OR liegt in der generell relativ breiten Verwendung des Konzepts (zur Kritik daran vgl. bereits Näätänen, 1979, S. 61ff.; O’Gorman, 1981, S. 788). D. h., Reaktionen, die in unterschiedlichsten Paradigmen zu beobachten sind, werden ohne weiteres darunter subsumiert, was eine einheitliche Theoriebildung und die Integration der Befunde erschwert. Somit ist eigentlich unklar, inwiefern man die Reaktionen der psychophysiologischen Aussagebeurteilung dem OR-Konzept zuordnen kann. Diese konzeptionellen Unklarheiten werden durch die insgesamt recht diffuse Verwendung des Signifikanz-Begriffs innerhalb der OR-Forschung (siehe Baltissen & Sartory, 1998, S. 22) zusätzlich forciert.

Die beobachteten Diskrepanzen zwischen den SCR- und HR-Befunden sind auch im Hinblick auf die Trennung zwischen den **Reaktionen nach Ein- vs. Ausblendung** diskussionswürdig. Die hier erzielten Resultate lassen sich weitgehend mit anderen Studien

zur forensischen Psychophysiologie bzw. zum DDP vergleichen, die ebenfalls den Antwortzeitpunkt verzögerten. Demnach ist die elektrodermale Differenzierung zwischen den Fragen- bzw. Itemtypen im Anschluß an deren Präsentation ausgeprägter als in den Reaktionen auf die Antworten bzw. den imperativen Reiz der Antwortgabe (vgl. Dawson, 1980; Furedy et al., 1988, 1991; Gödert et al., 2001). Indes fand man für die phasische HR stärkere Effekte nach der Ausblendung (vgl. Gödert et al., 2001). Im Gegensatz zu einigen der früheren Untersuchungen resultierten hier jedoch in beiden Zeitfenstern sowohl elektrodermale als auch kardiovaskuläre Reaktionsdifferenzen, wobei die Einzelvergleiche zwischen den Fragen- bzw. Itemtypen durchaus verschiedenartig ausfielen. Diese Ergebnisse sind zwar nicht ohne weiteres unter dem Gesichtspunkt der hypothetischen Unterscheidung zwischen Täuschungsabsicht und -handlung zu interpretieren. Sie können aber als ein Indiz gewertet werden, daß eine Trennung der Reaktionen auf die Darbietung der Fragen bzw. Items und dem imperativen Reiz der Antwortgabe unter methodischen Gesichtspunkten sinnvoll ist und möglicherweise zusätzliche Informationen über die Rolle unterschiedlicher Prozesse bietet (wie etwa Reizwahrnehmung und -verarbeitung vs. Reaktionsvorbereitung und -durchführung).

Die vorliegende Studie hat erneut die Bedeutung der **Erhebung subjektiver Daten** im Zusammenhang mit der psychophysiologischen Aussagebeurteilung unterstrichen. Die Resultate weisen darauf hin, daß sich die körperlichen Reaktionsunterschiede auch im introspektiven Erleben der Pbn manifestieren und daß die Relevanz der Fragen- bzw. Itemtypen des DLT bzw. GAT differentiell bewertet wird. Allerdings muß man diesen Ergebnissen mit Vorsicht begegnen, da mit der bereits angesprochenen Problematik reaktiver Antworttendenzen zu rechnen ist.

Eindeutige Interaktionen des Fragen- bzw. Itemtyps mit der **elektrodermalen Labilität** konnten nicht nachgewiesen werden, d. h., das Personenmerkmal übte keine nennenswerten Einflüsse auf die Reaktionsunterschiede aus. Dies spricht für die Annahme, daß die EL in Zusammenhang mit dem DLT und dem GAT vernachlässigbar ist und möglicherweise auch keine gravierenden Effekte auf deren Treffsicherheit hat. Aufgrund des Pilotcharakters der Studie sind solche Schlußfolgerungen aber nur unter Vorbehalt gestattet. Denn das Experiment sollte einen Beitrag zu den psychophysiologischen Grundlagen der Verfahren erbringen. Ihre kriterienbezogene Validität und deren Einflußfaktoren sind nur durch angemessene Feldforschung näher zu untersuchen.

In diesem Zusammenhang sei nochmals auf das **Problem der ökologischen Validität** hingewiesen, dem Analogstudien generell unterliegen (vgl. Abschnitt 2.9.2). So sind etwa Geldbelohnungen, wie sie auch im Rahmen dieser Untersuchung in Aussicht gestellt wurden, möglicherweise keine angemessene Umsetzung der potentiellen Kon-

sequenzen, die in der Feldanwendung der forensischen Psychophysiologie resultieren können (z. B. drohende Bestrafung nach einem positiven Befund; vgl. Podlesny & Truslow, 1993, S. 796). D. h., insbesondere die emotionalen Faktoren von realen Tests (z. B. Angst) lassen sich in Scheinverbrechen-Experimenten nur schwer simulieren, ohne forschungsethische Grenzen zu überschreiten. Darin mag auch einer der Gründe liegen, warum die gefundenen Reaktionsunterschiede v. a. auf der Basis kognitiver, speziell aufmerksamkeits- und informationsverarbeitungstheoretischer Ansätze zu erklären waren und weniger anhand emotionstheoretischer Modellvorstellungen. Dennoch kann man davon ausgehen, daß solche kognitiven Prozesse auch in Realsituationen eine wichtige Rolle spielen und daß ihr Stellenwert in Relation zu den eher emotionalen Faktoren nicht unterschätzt werden sollte (vgl. Steller, 1987, S. 106ff.). Diese Annahme wird durch die Vielfalt an kognitiven Theorien in der psychophysiologischen Aussageforschung gestützt (siehe Abschnitt 2.10.2).

Als wesentliches **Fazit** bleibt somit festzuhalten, daß man unter simulierten Tat- und Testbedingungen Reaktionsunterschiede zwischen den Fragen- bzw. Itemtypen der neueren Befragungstechniken DLT und GAT aufzeigen und untersuchen kann. Im Hinblick auf die zugrundeliegenden psychologischen und physiologischen Mechanismen bieten sich einige interessante Erklärungsansätze an. Dabei sind Faktoren, wie die differentielle Signifikanz der Reize und die Konflikte unter Täuschungsbedingungen, ebenso hervorzuheben wie die parasympathischen Einflüsse, die insbesondere im Zusammenhang mit der phasischen HR und möglicherweise auch bei den elektrodermalen Reaktionsunterschieden eine wichtige Rolle spielen. Von einer stringenten theoretischen Modellvorstellung ist man jedoch noch deutlich entfernt. Diesbezüglich ist weitere Forschung nötig. Hierzu erweist sich die laborexperimentelle Analyse der psychophysiologischen Aussagebeurteilung als ebenso zweckmäßig wie die entsprechende Feldforschung. Somit belegt die eigene Arbeit wie bereits eine Reihe anderer Studien, daß Scheinverbrechen-Experimente einen wichtigen Beitrag zur näheren Untersuchung der Methoden der forensischen Psychophysiologie leisten können.

Abschließend soll ein **Ausblick** auf potentielle zukünftige Forschungsrichtungen in diesem Bereich gegeben werden:

Was den **DLT** anbelangt, ist v. a. die Problematik der instruierten Lügen-Kontrollfragen von Interesse. Die vorliegenden Ergebnisse deuten darauf hin, daß auch die Kontrollfragen des DLT keine angemessene Kontrollbedingung für die wahrheitsgemäß beantworteten relevanten Fragen bieten. Die Gründe für die Diskrepanzen zu den früheren Studien mit dieser Befragungstechnik sind zunächst unklar, da sich das eigene Paradigma hinsichtlich mehrerer Aspekte von der herkömmlichen Vorgehensweise unterscheidet.

Dies betrifft insbesondere die Vortest-Prozedur und die Testdurchführung. D. h., es bleibt fraglich, ob es sich um ein grundsätzliches Manko der instruierten Lügen-Kontrollfragen handelt oder ob etwa das Fehlen eines Vortest-Interviews bzw. das hohe Maß an Standardisierung für die relativ schwachen Reaktionen auf diesen Fragentyp verantwortlich ist. Eine Möglichkeit, die Effekte solcher Faktoren auf die Reaktionen laborexperimentell zu untersuchen, besteht darin, sie systematisch zu variieren, anstatt sie wie hier konstant zu halten. Konkret könnte man sich beispielsweise vorstellen, in zwei unabhängigen Bedingungen die Art der Vortest-Prozedur (konventionelles Interview vs. schriftliche Instruktionen) zu manipulieren und deren Einflüsse auf die Reaktionsunterschiede zwischen den relevanten Fragen und Kontrollfragen zu analysieren. Ebenso wäre die Gegenüberstellung einer standardisierten, computergesteuerten Testdurchführung mit einer persönlichen Befragung durch den Untersucher denkbar. Derartige Studien können Hinweise geben, inwiefern solche Faktoren beim DLT von Belang sind. Allerdings ist die Bewährung dieses Verfahrens im Anwendungsbereich nur durch weitere Feldforschung zu überprüfen.

Den **GAT** betreffend, weist die eigene Studie darauf hin, daß die Differenzierung zwischen Schuldigen und Unschuldigen mit Tatwissen teilweise darauf zurückzuführen ist, daß tatbezogene Items stärkere Reaktionen auslösen, wenn man die Tat verübt hat und wahrheitswidrig abstreitet (Relevant-Täuschung), als wenn man ihren Tatbezug zwar erkennt, diesbezüglich aber unschuldig ist und den Vorwurf wahrheitsgemäß verneint (Relevant-Aufrichtigkeit). Interessant wäre, ob sich diese Reaktionsunterschiede auch unter Realbedingungen nachweisen lassen. Wie bereits im Abschnitt 2.8 angesprochen, ist in Realfällen damit zu rechnen, daß die relevanten Items aufgrund des konkreten Tatverdachts, der drohenden Konsequenzen und einer entsprechend gesteigerten Täuschungsmotivation auch für Unschuldige mit Tatwissen eine außerordentlich hohe Signifikanz aufweisen. Dadurch könnten die Reaktionsunterschiede zwischen Relevant-Täuschung und -Aufrichtigkeit schwinden bzw. verlorengehen, und somit auch die Differenzierung zwischen Schuldigen und Unschuldigen mit Tatwissen. Anhand eines Vergleichs mit dem TWT wäre zu klären, inwiefern die potentiellen Vorteile des GAT auch unter Feldbedingungen zu beobachten sind. Einschränkend ist jedoch zu bemerken, daß die geforderte Feldforschung dadurch erschwert wird, daß in der angewandten psychophysiologischen Aussagebeurteilung der TWT bislang nur selten und der GAT gar nicht zum Einsatz kamen. Folglich liegen diesbezüglich kaum Erfahrungswerte vor.

Im Hinblick auf die **physiologischen Grundlagen** der beobachteten Reaktionsunterschiede beim DLT und GAT liegt ein besonderes Augenmerk auf der Berücksichtigung zusätzlicher körperlicher Variablen und der weiteren Klärung des Zusammenspiels von Sympathikus und Parasympathikus. Neben der T-Wellen-Amplitude, deren Probleme

bereits erörtert wurden, kommt die Auswertung der ebenfalls im Rahmen dieser Studie erhobenen peripheren Durchblutung und der Atembewegungen in Betracht (vgl. die Abschnitte 4.5.3 und 4.5.4). Aufgrund der besonderen Methodik der vorliegenden Untersuchung und ihres Pilotcharakters setzt die Parametrisierung zwar eine gesonderte Aufbereitung des Datenmaterials voraus, dennoch sollen hier einige Möglichkeiten in Erwägung gezogen werden.

Die Differenz zwischen (systolischem) Maximal- und (diastolischem) Minimalwert einer Herzperiode des am Finger abgeleiteten Plethysmogramms bezeichnet man als **Fingerpulsvolumenamplitude** (FPA). Die FPA wurde bereits in einigen Validitätsstudien zur psychophysiologischen Aussagebeurteilung miterhoben und zeigte überwiegend eine reliable Differenzierung zwischen Schuldigen und Unschuldigen (z. B. Honts & Raskin, 1988; Horowitz et al., 1997; Kircher & Raskin, 1988; O'Toole et al., 1994; Podlesny & Raskin, 1978). Dabei muß man berücksichtigen, daß die Durchblutung der Haut überwiegend durch die Weite der herzfernen Blutgefäße bestimmt ist, deren Vasomotorik ausschließlich über den Sympathikus gesteuert wird (Jänig, 1987, S. 254). D. h., phasische Abnahmen der peripheren Durchblutung und somit der FPA, wie sie auch nach Darbietung psychologischer Reize zu beobachten sind (vgl. Schandry, 1998, S. 174f.), geben die Vasokonstriktion und somit die Aktivitätssteigerung der entsprechenden sympathischen Innervation wieder (Vossel & Zimmer, 1998, S. 75ff.). Ein Vorteil der zusätzlichen Erhebung der FPA bestünde also darin, daß man neben der HR, in der sich sowohl die sympathische als auch die parasympathische Modulation der Herztätigkeit manifestiert, einen weiteren kardiovaskulären Parameter erhält, der jedoch primär die Sympathikusaktivität indiziert.

Ebenso wie die elektrodermalen und kardiovaskulären Variablen zählt die Erfassung der **Atemkurve** seit langem zu den Standards der psychophysiologischen Aussagebeurteilung (Podlesny & Raskin, 1977, S. 793f.; Tent, 1967, S. 214ff.). In diesem Zusammenhang hat sich insbesondere die sog. „Atemkurvenlänge" („respiration line length", RLL) als ein brauchbarer Indikator erwiesen (z. B. Ben-Shakhar & Dolev, 1996; Bradley & Rettinger, 1992; Elaad & Ben-Shakhar, 1997; Honts et al., 1996; Horowitz et al., 1997; Kircher & Raskin, 1988; Timm, 1982). Dabei wird die Länge des kurvilinearen Verlaufs des Pneumogramms innerhalb eines definierten Zeitintervalls gemessen. In dieses kombinierte Maß gehen zwei für die psychophysiologische Aussagebeurteilung wesentliche Veränderungen der Respiration ein, nämlich sowohl die Verlangsamung als auch die Abflachung der Atmung (Verringerung der Frequenz und Amplitude der Atemkurve; siehe auch Elaad et al., 1992, S. 758). Die relevante Reaktionsrichtung besteht somit in einer Abnahme der RLL nach Stimulusdarbietung. Abgesehen davon könnte auch das Ausmaß der „Herzfrequenz-Respirations-Kopplung" (Schandry, 1998,

S. 141) zusätzliche Informationen über die Einflüsse des vegetativen Nervensystems bieten. Die Atmungsaktivität ist zwar in komplexer Weise von zentralnervösen und autonomen (v. a. parasympathischen) Steuerungsprozessen sowie peripheren Feedbackmechanismen abhängig und unterliegt außerdem der willkürlichen Kontrolle (vgl. Lorig & Schwartz, 1990, S. 583f.), aber in Kombination mit der HR lassen sich über die respiratorische Sinusarrhythmie (RSA) Rückschlüsse auf den Anteil der Vagusaktivität an der vegetativen Kontrolle des Herzens ziehen (zusammenfassend Berntson et al., 1997, S. 636f.; Grossman, 1992, S. 141ff.; Porges & Byrne, 1992, S. 98f.).

7. Zusammenfassung

Die **forensische Psychophysiologie** („psychophysiologische Aussagebeurteilung“, „Lügendetektion“) beschäftigt sich mit dem Problem, inwieweit man anhand der körperlichen Reaktionen auf bestimmte Reize (Fragen oder Items) diagnostische Schlußfolgerungen über die Glaubwürdigkeit bzw. Tatkenntnisse einer Person treffen kann. Der Directed Lie Test (DLT) und der Guilty Actions Test (GAT) sind Weiterentwicklungen der beiden wichtigsten Befragungstechniken der forensischen Psychophysiologie. Anders als der konventionelle Kontrollfragentest bietet der DLT eine erhöhte Standardisierbarkeit, und im Vergleich zum Tatwissentest soll der GAT eine bessere Differenzierung zwischen Tätern und Unschuldigen mit Tatwissen ermöglichen.

Ausgehend von der Kritik an der üblichen individualdiagnostischen Zielsetzung von Laborstudien zur psychophysiologischen Aussagebeurteilung, die v. a. auf die Abschätzung der kriterienbezogenen Validität ausgerichtet sind, wurde hier eine alternative Vorgehensweise gewählt. Statt auf die Bestimmung von Trefferquoten zielte die **vorliegende Studie** im wesentlichen darauf ab, mittels eines Scheinverbrechen-Experiments mit standardisierter Durchführung und objektiver Auswertung der Testverfahren quantitative Reaktionsunterschiede zwischen den einzelnen Stimulusarten des DLT bzw. GAT nachzuweisen und zu analysieren.

Die **Effekte der Fragen- bzw. Itemtypen** auf die körperlichen Reaktionen sollten direkt quantifiziert und untersucht werden. Unter anderem interessierte, inwiefern Personen auf relevante Stimuli, die sich auf eine von ihnen durchgeführte Tat bezogen und somit wahrheitswidrig verneint wurden, stärker reagierten als auf entsprechende Vergleichsreize, die ein nicht verübtes Delikt thematisierten und somit wahrheitsgemäß abgestritten wurden. Da eine solche Fragestellung eine intraindividuelle Variation von Täuschung und Aufrichtigkeit bei den relevanten Stimuli impliziert und daher durch einen interindividuellen Vergleich „schuldiger“ vs. „unschuldiger“ Probanden (Pbn) nicht adäquat beantwortet werden kann, variierte man die Täterschaft innerhalb der Personen. Bei den Scheinverbrechen handelte es sich um zwei simulierte Schmuckdiebstähle, die äquivalent gestaltet waren. Jeder Pb sollte nur einen davon durchführen, wurde aber anschließend hinsichtlich beider getestet, wobei entweder der DLT oder der GAT zum Einsatz kamen. Darüber hinaus variierte man den Wahrheitsgehalt der Antworten auf die Kontrollfragen des DLT, um zu überprüfen, ob dies einen Effekt auf die Reaktionsstärke hat. Beide Testverfahren beinhalteten zusätzlich irrelevante Stimuli, die in keiner Beziehung zu den Scheinverbrechen standen.

Neben den physiologischen **abhängigen Variablen**, Hautleitfähigkeitsreaktion (SCR) und phasische Herzschlagfrequenz (HR), wurden auch subjektive Einschätzungen der Reaktionsstärke und Bedeutsamkeit der Fragen bzw. Items erfaßt. Durch eine zeitliche Verzögerung der Antworten wollte man außerdem die körperlichen Reaktionen auf die Darbietung der Reize von den physiologischen Begleiterscheinungen der Antworten trennen. Ferner untersuchte man erstmals für den DLT und den GAT die Einflüsse des Personenmerkmals **elektrodermale Labilität** (EL) auf die Reaktionsunterschiede zwischen den Fragen- bzw. Itemtypen, da frühere Studien Zusammenhänge zwischen der EL und der Treffsicherheit der psychophysiologischen Aussagebeurteilung nahegelegt hatten.

Einhundertzweiundzwanzig männliche Pbn nahmen am **Experiment** teil. Die Daten von $N = 80$ Personen gingen in die Auswertung ein. Die Einteilung in elektrodermal Stabile und Labile erfolgte per Mediandichotomisierung der Häufigkeit elektrodermaler Spontanfluktuationen während einer reizfreien Ruhemessung. Unabhängig davon variierte man die Tatbedingung, indem jeweils die Hälfte der Pbn instruiert wurde, im Rahmen eines simulierten Schmuckdiebstahls entweder einen Ring oder eine Kette zu entwenden. Somit war jede Person hinsichtlich eines der beiden Delikte „schuldig“ und hinsichtlich des anderen „unschuldig“. Während der Tatbegehung wurden alle Pbn auch mit den kritischen Details des jeweils nicht verübten Scheinverbrechens konfrontiert. Danach absolvierten sie den DLT oder GAT.

Der **DLT** bestand aus fünf Fragentypen. Die relevanten Fragenpaare waren inhaltlich parallelisiert und zielten auf das begangene (Relevant-Täuschung) vs. nicht begangene Scheinverbrechen (Relevant-Aufrichtigkeit) ab. Die ebenfalls parallelisierten Kontrollfragenpaare thematisierten kleinere Normverstöße und wurden instruiert wahrheitswidrig oder wahrheitsgemäß verneint (Lügen-Kontroll vs. Wahrheit-Kontroll). Die neutralen Fragen nach dem momentanen Aufenthaltsort (Irrelevant) wurden wahrheitsgemäß bejaht. Der **GAT** bestand aus mehreren Multiple-Choice-Fragen mit jeweils sechs Antwortalternativen, die sich zu vier Itemtypen gruppieren ließen. Die tatbezogenen Items (Relevant-Täuschung vs. Relevant-Aufrichtigkeit) thematisierten Details des begangenen vs. nicht begangenen Scheinverbrechens. Ferner wurde die erste nicht tatbezogene Alternative (Irrelevant-1) pro Multiple-Choice-Frage, die normalerweise bei der Testauswertung unberücksichtig bleibt, von den restlichen irrelevanten Items (Irrelevant-Vergleich) abgegrenzt. Die Pbn sollten alle Items verneinen.

Die **Durchführung** der Tests war standardisiert. Alle Instruktionen wurden schriftlich erteilt. Die Darbietung der Fragen bzw. Items erfolgte computergesteuert sowohl per Bildschirm als auch per Lautsprecher. Die Stimuli wurden für die Dauer von 8 – 10

Sekunden auf einem Monitor eingeblendet und parallel dazu akustisch eingespielt. Die Ausblendung auf dem Bildschirm fungierte als imperativer Reiz für die Antwortgabe. Somit konnte man die elektrodermalen und kardiovaskulären Reaktionen auf die Darbietung der Fragen bzw. Items und die körperlichen Veränderungen nach Ausblendung voneinander abgrenzen und separat analysieren. Während der Befragung befanden sich die Pbn allein in der Meßkabine. Dadurch wurden potentielle Untersuchereffekte zusätzlich reduziert. Im Anschluß an die psychophysiologische Aussagebeurteilung sollten die Pbn retrospektiv ihre Reaktionsstärke auf die einzelnen Fragen bzw. Items und die Relevanz der Stimuli für das Testergebnis auf siebenstufigen Ratingskalen einschätzen.

Sowohl beim DLT als auch beim GAT resultierten Haupteffekte der Fragen- bzw. Itemtypen. Dies galt für die elektrodermalen und kardiovaskulären Reaktionen ebenso wie für die subjektiven Daten. In beiden Testverfahren fand man quantitative Reaktionsunterschiede zwischen den relevanten Stimuli, die sich auf das begangene vs. nicht begangene Scheinverbrechen bezogen. Allgemein wurde auf Relevant-Täuschung mit einer stärkeren SCR-Magnitude und einer niedrigeren phasischen HR reagiert als auf Relevant-Aufrichtigkeit. Entgegen den Erwartungen erreichten die Reaktionen auf die instruierten Lügen-Kontrollfragen des **DLT** nicht das Niveau von Relevant-Aufrichtigkeit. Darüber hinaus zeigten sich keine signifikanten Unterschiede zwischen den Lügen- und Wahrheit-Kontrollfragen. Dies deutete darauf hin, daß die Kontrollfragen des DLT keine angemessene Referenzbedingung für die wahrheitsgemäß beantworteten relevanten Fragen darstellten und daß die Reaktionen auf die Kontrollfragen unabhängig vom Wahrheitsgehalt der Antworten waren. Beim **GAT** evozierten die Irrelevant-Vergleich-Items in der Regel schwächere Reaktionen als die Relevant-Aufrichtigkeit-Items. Dies stützte die Annahme, daß unabhängig von der Täterschaft und einer eventuellen wahrheitswidrigen Verneinung bereits der erkannte Tatbezug der relevanten Items zu einer Reaktionserhöhung führen konnte. Die Reaktionsunterschiede zwischen Relevant-Täuschung und -Aufrichtigkeit und deren Verhältnis zu Irrelevant-Vergleich boten aber eine potentielle Erklärung dafür, daß der GAT eine Differenzierung zwischen Tätern und Unschuldigen mit Tatwissen ermöglicht.

Insgesamt war für die SCRs die Differenzierung zwischen den Fragen- bzw. Itemtypen nach Einblendung der Stimuli größer, bei den HR-Reaktionen hingegen nach Ausblendung. Die Diskrepanzen zwischen den Reaktionen nach Ein- vs. Ausblendung wurden auf der Basis einer Unterscheidung zwischen reizverarbeitenden, reaktionsvorbereitenden und antwortbegleitenden Prozessen interpretiert. Außerdem fand man einige Übereinstimmungen zwischen den physiologischen und den subjektiven Daten. Die Reaktionsstärke und die Relevanz der tatbezogenen Fragen bzw. Items (speziell

Relevant-Täuschung) wurden höher eingeschätzt als bei den anderen Stimuli. Weder für die physiologischen Reaktionen noch für die Ratings konnten bedeutsame Interaktionen zwischen der elektrodermalen Labilität und den Fragen- bzw. Itemtypen nachgewiesen werden, d. h., das Personenmerkmal übte keine nennenswerten moderierenden Einflüsse auf die entsprechenden Reaktionsunterschiede aus.

Die beobachteten Reaktionsunterschiede wurden v. a. unter aufmerksamkeits- und informationsverarbeitungstheoretischen Gesichtspunkten (z. B. Orientierungsreaktion, Einflüsse der Reizsignifikanz, konflikttheoretische Annahmen) und im Hinblick auf ihre sympathische (SCR) vs. parasympathische (HR) Vermittlung **diskutiert**. Die Überschneidungen zwischen den SCR- und HR-Ergebnissen sprachen zugunsten der These, daß über weite Strecken ähnliche psychophysiologische Mechanismen involviert waren. Bezüglich ihrer theoretischen Grundlagen besteht jedoch noch zusätzlicher Klärungsbedarf. Hierzu wurden konkrete Vorschläge für die künftige Forschung in diesem Bereich geäußert.

8. Literaturverzeichnis

Achenbach, H. (1984). Polygraphie pro reo? *Neue Zeitschrift für Strafrecht (NStZ), 4* (8), 350–352.

Amelang, M. & Zielinski, W. (1997). *Psychologische Diagnostik und Intervention* (2. Aufl.). Berlin: Springer.

Amelung, K. (1982). Verwendung eines Lügendetektors. Anmerkung zu BVerfG, Beschl. v. 18. 8. 1981 – 2 BvR 166/81. *Neue Zeitschrift für Strafrecht (NStZ), 2* (1), 38–40.

Amtsgericht Demmin/Zweigstelle Malchin. (1998). Urteil vom 7. September 1998 – 94 Ls 182/98. *JurPC* [Internet-Zeitschrift für Rechtsinformatik]. Verfügbar unter: http://www.jura.uni-sb.de/jurpc/rechtspr/19980176.htm [2001-03-27].

Anastasi, A. & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.

Backster, C. (1962). Methods of strengthening our polygraph technique. *Police, 6* (5), 61–68. (zitiert nach Podlesny & Raskin, 1977, S. 786)

Backster, C. (1974). Anticlimax dampening concept. *Polygraph, 3*, 28–50. (zitiert nach Lykken, 1998, S. 123)

Baddeley, A. (1997). *Human memory: Theory and practice* (rev. ed.). Hove: Psychology Press.

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.

Baltissen, R. & Sartory, G. (1998). Orientierungs-, Defensiv- und Schreckreaktionen in Grundlagenforschung und Anwendung. In F. Rösler (Hrsg.), *Enzyklopädie der Psychologie: Themenbereich C Theorie und Forschung, Serie 1 Biologische Psychologie, Band 5 Ergebnisse und Anwendungen der Psychophysiologie* (S. 1–45). Göttingen: Hogrefe.

Barland, G. H. (1981). *A validation and reliability study of the counterintelligence screening test*. Unpublished manuscript, Fort George Meade, MD: Security Support Battalion, Military Intelligence Group. (zitiert nach OTA, 1983, S. 76)

Barland, G. H. & Raskin, D. C. (1975a). An evaluation of field techniques in detection of deception. *Psychophysiology, 12*, 321–330.

Barland, G. H. & Raskin, D. C. (1975b). Psychopathy and detection of deception in criminal suspects [Abstract]. *Psychophysiology, 12*, 224.

Barland, G. H. & Raskin, D. C. (1976). *Validity and reliability of polygraph examinations of criminal suspects* (U. S. Department of Justice Report No. 76-1, Contract 75-NI-99-0001). Salt Lake City, UT: University of Utah, Department of Psychology. (zitiert nach Blinkhorn, 1988, S. 32, und Berning, 1992, S. 69)

Barry, R. J. (1984). Preliminary processes in O-R elicitation. *Acta Psychologica, 55*, 109–142.

Barry, R. J. (1996). Preliminary process theory: Towards an integrated account of the psychophysiology of cognitive processes. *Acta Neurobiologiae Experimentalis*, *56*, 469–484.

Barry, R. J. & Maltzman, I. (1985). Heart rate deceleration is not an orienting reflex; heart rate acceleration is not a defensive reflex. *The Pavlovian Journal of Biological Science*, *20*, 15–28

Bashore, T. R. & Rapp, P. E. (1993). Are there alternatives to traditional polygraph procedures? *Psychological Bulletin*, *113*, 3–22.

Beardsley, T. (1999). Truth or consequences: A polygraph screening program raises questions about the science of lie detection. *Scientific American, 281* (4), 11–12.

Ben-Shakhar, G. (1977). A further study of the dichotomization theory in detection of information. *Psychophysiology*, *14*, 408–413.

Ben-Shakhar, G. (1980). Habituation of the orienting response to complex sequences of stimuli. *Psychophysiology*, *17*, 524–534.

Ben-Shakhar, G. (1991a). Clinical judgment and decision-making in CQT-polygraphy: A comparison with other pseudoscientific applications in psychology. *Integrative Physiological and Behavioral Science*, *26*, 232–240.

Ben-Shakhar, G. (1991b). Future prospects of psychophysiological detection: Replacing the CQT by the GKT. In J. R. Jennings, P. K. Ackles & M. G. H. Coles (Eds.), *Advances in psychophysiology* (Vol. 4, pp. 193–199). London: Jessica Kingsley Publishers.

Ben-Shakhar, G. (1994). The roles of stimulus novelty and significance in determining the electrodermal orienting response: Interactive versus additive approaches. *Psychophysiology*, *31*, 402–411.

Ben-Shakhar, G., Asher, T., Poznansky-Levy, A., Asherowitz, R. & Lieblich, I. (1989). Stimulus novelty and significance as determinants of electrodermal responsivity: The serial position effect. *Psychophysiology*, *26*, 29–38.

Ben-Shakhar, G. & Dolev, K. (1996). Psychophysiological detection through the guilty knowledge technique: Effects of mental countermeasures. *Journal of Applied Psychology*, *81*, 273–281.

Ben-Shakhar, G. & Furedy, J. J. (1990). *Theories and applications in the detection of deception: A psychophysiological and international perspective*. New York: Springer-Verlag.

Ben-Shakhar, G. & Lieblich, I. (1982). The dichotomization theory for differential autonomic reponsivity reconsidered. *Psychophysiology*, *19*, 277–281.

Ben-Shakhar, G., Lieblich, I. & Bar-Hillel, M. (1982). An evaluation of polygraphers' judgments: A review from a decision theoretic perspective. *Journal of Applied Psychology*, *67*, 701–713.

Ben-Shakhar, G., Lieblich, I. & Kugelmass, S. (1975). Detection of information and GSR habituation: An attempt to derive detection efficiency from two habituation curves. *Psychophysiology*, *12*, 283–288.

Ben-Shakhar, G., Lieblich, I. & Kugelmass, S. (1982). Interactive effects of stimulus probability and significance on the skin conductance response. *Psychophysiology*, *19*, 112–114.

Berlyne, D. E. (1960). *Conflict, arousal, and curiosity*. New York: McGraw-Hill.

Berlyne, D. E. (1961). Conflict and the orientation reaction. *Journal of Experimental Psychology, 62*, 476–483.

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 und Strafrechtsreform (MschrKrim), 76* (4), 242–255.

Bernstein, A. S. (1979). The orienting response as novelty and significance detector: Reply to O'Gorman. *Psychophysiology, 16*, 263–273.

Bernstein, A. S., Taylor, K. W. & Weinstein, E. (1975). The phasic electrodermal response as a differentiated complex reflecting stimulus significance. *Psychophysiology, 12*, 158–169.

Berntson, G. G., Bigger, J. T., Jr, Eckberg, D. L., Grossman, P., Kaufmann, P. G., Malik, M., Nagaraja, H. N., Porges, S. W., Saul, J. P., Stone, P. H. & van der Molen, M. W. (1997). Heart rate variability: Origins, methods, and interpretive caveats. *Psychophysiology, 34*, 623–648.

Bersh, P. J. (1969). A validation study of polygraph examiner judgments. *Journal of Applied Psychology, 53*, 399–403.

Birbaumer, N. & Schmidt, R. F. (1999). *Biologische Psychologie* (4. Aufl.). Berlin: Springer.

Blinkhorn, S. (1988). Lie detection as a psychometric procedure. In A. Gale (Ed.), *The polygraph test: Lies, truth and science* (pp. 29–39). London: Sage Publications.

Bohlin, G. & Kjellberg, A. (1979). Orienting activity in two-stimulus paradigms as reflected in heart rate. In H. D. Kimmel, E. H. van Olst & J. F. Orlebeke (Eds.), *The orienting reflex in humans* (pp. 169–197). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bortz, J. (1993). *Statistik für Sozialwissenschaftler* (4. Aufl.). Berlin: Springer.

Bortz, J. & Döring, N. (1995). *Forschungsmethoden und Evaluation* (2. Aufl.). 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.

Bradley, M. T. & Ainsworth, D. (1984). Alcohol and the psychophysiological detection of deception. *Psychophysiology, 21*, 63–71.

Bradley, M. T. & Black, M. E. (1998). A control question test oriented towards students. *Perceptual and Motor Skills, 87*, 691–700.

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. & Black, M. E. (1996). The control question test in polygraphic examinations with actual controls for truth. *Perceptual and Motor Skills*, *83*, 755–762.

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.

Bronfenbrenner, U. (1981). *Die Ökologie der menschlichen Entwicklung: Natürliche und geplante Experimente*. Stuttgart: Klett-Cotta.

Bull, R. H. C. & Gale, M. A. (1973). The reliability of and interrelationships between various measures of electrodermal activity. *Journal of Experimental Research in Personality*, *6*, 300–306.

Bundesgerichtshof. (1954). Urteil vom 16. Februar 1954 – 1 StR 578/53. *Entscheidungen des Bundesgerichtshofes in Strafsachen (BGHSt)*, *5*, 332–338.

Bundesgerichtshof. (2000). Urteil vom 17. Dezember 1998 – 1 StR 156/98. *Entscheidungen des Bundesgerichtshofes in Strafsachen (BGHSt)*, *44*, 308–328.

Bundesverfassungsgericht. (1981). Beschluß vom 18. August 1981 – 2 BvR 166/81. *Neue Zeitschrift für Strafrecht (NStZ)*, *1* (11), 446–447.

Bunnell, D. E. (1980). T-wave amplitude and the P-Q interval: Relationships to non-invasive indices of myocardial performance. *Psychophysiology*, *17*, 592–597.

Campbell, D. T. & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Chicago, IL: Rand McNally.

Carroll, D. (1988). How accurate is polygraph lie detection? In A. Gale (Ed.), *The polygraph test: Lies, truth and science* (pp. 19–28). London: Sage Publications.

Chase, W. G., Graham, F. K. & Graham, D. T. (1968). Components of HR response in anticipation of reaction time and exercise tasks. *Journal of Experimental Psychology*, *76*, 642–648.

Coles, M. G. H. & Duncan-Johnson, C. C. (1975). Cardiac activity and information processing: The effects of stimulus significance, and detection and response requirements. *Journal of Experimental Psychology*: *Human Perception and Performance*, *1*, 418–428.

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.

Cook, T. D., Campbell, D. T. & Peracchio, L. (1990). Quasi experimentation. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 1, 2nd ed., pp. 491–576). Palo Alto, CA: Consulting Psychologists Press.

Cook, E., III & Turpin, G. (1997). Differentiating orienting, startle, and defense responses: The role of affect and its implications for psychopathology. In P. J. Lang, R. F. Simons & M. T. Balaban (Eds.), *Attention and orienting: Sensory and motivational processes* (pp. 137–164). Mahwah, NJ: Lawrence Erlbaum Associates.

Corteen, R. S. (1969). Skin conductance changes and word recall. *British Journal of Psychology, 60*, 81–84.

Costa, P. T., Jr & McCrae, R. R. (1992a). Four ways five factors are basic. *Personality and Individual Differences, 13*, 653–665.

Costa, P. T., Jr & McCrae, R. R. (1992b). Reply to Eysenck. *Personality and Individual Differences, 13*, 861–865.

Costa, P. T., Jr & McCrae, R. R. (1992c). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Craik, F. I. M. & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*, 671–684.

Craik, F. I. M. & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*, 268–294.

Crider, A. (1993). Electrodermal response lability-stability: Individual difference correlates. In J.-C. Roy, W. Boucsein, D. C. Fowles & J. H. Gruzelier (Eds.), *Progress in electrodermal research* (pp. 173–186). New York: Plenum Press.

Cross, T. P. & Saxe, L. (1992). A critique of the validity of polygraph testing in child sexual abuse cases. *Journal of Child Sexual Abuse, 1* (4), 19–33.

Curio, I. & Scholz, O. B. (1991). Glaubhaftigkeitsbeurteilung von kurzen Zeugenaussagen mittels behavioraler und psychophysiologischer Parameter. *Zeitschrift für Experimentelle und Angewandte Psychologie, 38*, 188–200.

Cutrow, R. J., Parks, A., Lucas, N. & Thomas, K. (1972). The objective use of multiple physiological indices in the detection of deception. *Psychophysiology, 9*, 578–588.

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.

de Boer, R. W., Karemaker, J. M. & Strackee, J. (1985). Description of heart-rate variability data in accordance with a physiological model for the genesis of heartbeats. *Psychophysiology, 22*, 147–155.

Delvo, M. (1981). *Der Lügendetektor im Strafprozeß der U.S.A.* Königstein: Athenäum.

Dickinson, J. R., Jr & Smith, B. D. (1973). Nonspecific activity and habituation of tonic and phasic skin conductance in somatic complainers and controls as a function of auditory stimulus intensity. *Journal of Abnormal Psychology, 82*, 404–413.

Dionisio, D. P., Granholm, E., Hillix, W. A. & Perrine, W. F. (2001). Differentiation of deception using pupillary responses as an index of cognitive processing. *Psychophysiology*, *38*, 205–211.

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. & Ben-Shakhar, G. (1989). Effects of motivation and verbal response type on psychophysiological detection of information. *Psychophysiology*, *26*, 442–451.

Elaad, E. & Ben-Shakhar, G. (1997). Effects of item repetitions and variations on the efficiency of the guilty knowledge test. *Psychophysiology*, *34*, 587–596.

Elaad, E., Ginton, A. & Ben-Shakhar, G. (1994). The effects of prior expectations and outcome knowledge on polygraph examiners' decisions. *Journal of Behavioral Decision Making*, *7*, 279–292.

Elaad, E., Ginton, A. & Jungman, N. (1992). Detection measures in real-life criminal guilty knowledge tests. *Journal of Applied Psychology*, *77*, 757–767.

Endres, J. & Scholz, O. B. (1994). Sexueller Kindesmißbrauch aus psychologischer Sicht – Formen, Vorkommen, Nachweis. *Neue Zeitschrift für Strafrecht (NStZ)*, *14* (10), 466–473.

Engel, B. T. (1960). Stimulus-response and individual-response specifity. *Archives of General Psychiatry*, *2*, 305–313.

Fahrenberg, J. (1969). Die Bedeutung individueller Unterschiede für die Methodik der Aktivierungsforschung. In W. Schönpflug (Hrsg.), *Methoden der Aktivierungsforschung* (S. 95–121). Bern: Huber.

Fahrenberg, J. (1983). Psychophysiologische Methodik. In K.-J. Groffmann & L. Michel (Hrsg.), *Enzyklopädie der Psychologie: Themenbereich B Methodologie und Methoden, Serie II Psychologische Diagnostik, Band 4 Verhaltensdiagnostik* (S. 1–192). Göttingen: Hogrefe.

Fahrenberg, J., Walschburger, P., Foerster, F., Myrtek, M. & Müller, W. (1979). *Psychophysiologische Aktivierungsforschung: Ein Beitrag zu den Grundlagen der multivariaten Emotions- und Stress-Theorie*. München: Minerva.

Faigman, D. L., Kaye, D. H., Saks, M. J. & Sanders, J. (1997). The legal relevance of scientific research on polygraph tests. In D. L. Faigman, D. H. Kaye, M. J. Saks & J. Sanders (Eds.), *Modern scientific evidence: The law and science of expert testimony* (Vol. 1, pp. 554–564). St. Paul, MN: West Publishing Company.

Faller, K. C. (1997). The polygraph, its use in cases of alleged sexual abuse: An exploratory study. *Child Abuse & Neglect*, *21*, 993–1008.

Farwell, L. A. & Donchin, E. (1991). The truth will out: Interrogative polygraphy ("lie detection") with event-related brain potentials. *Psychophysiology*, *28*, 531–547.

Fiedler, K. (1999). Gutachterliche Stellungnahme zur wissenschaftlichen Grundlage der Lügendetektion mithilfe sogenannter Polygraphentests. *Praxis der Rechtspsychologie*, *9* (Sonderheft), 5–44.

Fowles, D. C., Christie, M. J., Edelberg, R., Grings, W. W., Lykken, D. T. & Venables, P. H. (1981). Publication recommendations for electrodermal measurements. *Psychophysiology*, *18*, 232–239.

Frister, H. (1994). Der Lügendetektor – Zulässiger Sachbeweis oder unzulässige Vernehmungsmethode? *Zeitschrift für die gesamte Strafrechtswissenschaft (ZStW)*, *106* (2), 303–331.

Fröhlich, W. D. (2000). *Wörterbuch Psychologie* (23. Aufl.). München: Deutscher Taschenbuch Verlag.

Furedy, J. J. (1985). Joint use of heart-rate and T-wave amplitude as non-invasive cardiac performance measures: A psychophysiological perspective. In J. F. Orlebeke, G. Mulder & L. J. P. van Doornen (Eds.), *Psychophysiology of cardiovascular control* (pp. 237–256). New York: Plenum Press.

Furedy, J. J. (1986). Lie detection as psychophysiological differentiation: Some fine lines. In M. G. H. Coles, E. Donchin & S. W. Porges (Eds.), *Psychophysiology: Systems, processes, and applications – A handbook* (pp. 683–701). New York: Guilford Press.

Furedy, J. J. (1991). Alice-in-Wonderland terminological usage in, and communicational concerns about, that peculiarly American flight of technological fancy: The CQT polygraph. *Integrative Physiological and Behavioral Science*, *26*, 241–247.

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. (1996a). The North American polygraph and psychophysiology: Disinterested, uninterested, and interested perspectives. *International Journal of Psychophysiology*, *21*, 97–105.

Furedy, J. J. (1996b). Some elementary distinctions among, and comments concerning, the 'control' question 'test' (CQT) polygrapher's many problems: A reply to Honts, Kircher and Raskin. *International Journal of Psychophysiology*, *22*, 53–59.

Furedy, J. J. & Ben-Shakhar, G. (1991). The roles of deception, intention to deceive, and motivation to avoid detection in the psychophysiological detection of guilty knowledge. *Psychophysiology*, *28*, 163–171.

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., Gigliotti, F. & Ben-Shakhar, G. (1994). Electrodermal differentiation of deception: The effect of choice versus no choice of deceptive items. *International Journal of Psychophysiology*, *18*, 13–22.

Furedy, J. J. & Heslegrave, R. J. (1991a). 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.

Furedy, J. J. & Heslegrave, R. J. (1991b). Some elaborations on the specific-effects orientation's application to North American CQT polygraphs. In J. R. Jennings, P. K. Ackles & M. G. H. Coles (Eds.), *Advances in psychophysiology* (Vol. 4, pp. 233–245). London: Jessica Kingsley Publishers.

Furedy, J. J. & Liss, J. (1986). Countering confessions induced by the polygraph: Of confessionals and psychological rubber hoses. *Criminal Law Quarterly*, *29*, 91–114.

Furedy, J. J., Posner, R. T. & Vincent, A. (1991). Electrodermal differentiation of deception: Perceived accuracy and perceived memorial content manipulations. *International Journal of Psychophysiology, 11,* 91–97.

Furedy, J. J., Szabo, A. & Péronnet, F. (1996). Effects of psychological and physiological challenges on heart rate, T-wave amplitude, and pulse-transit time. *International Journal of Psychophysiology, 22,* 173–183.

Geddes, L. A. & Newberg, D. C. (1977). Cuff pressure oscillations in the measurement of relative blood pressure. *Psychophysiology, 14,* 198–202.

Geppert, K. (1999). Nochmals: Zur Unzulässigkeit einer Beweiserhebung mit Hilfe eines „Lügendetektors“ (Polygraphen). *Juristische Ausbildung (Jura), 21* (7), Jura-Kartei: StPO § 136 a/11.

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.

Ginton, A., Daie, N., Elaad, E. & Ben-Shakhar, G. (1982). A method for evaluating the use of the polygraph in a real-life situation. *Journal of Applied Psychology, 67,* 131–137.

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.

Graham, F. K. & Clifton, R. K. (1966). Heart-rate change as a component of the orienting response. *Psychological Bulletin, 65,* 305–320.

Graham, F. K. & Hackley, S. A. (1991). Passive and active attention to input. In J. R. Jennings & M. G. H. Coles (Eds.), *Handbook of cognitive psychophysiology: Central and autonomic nervous system approaches* (pp. 251–356). Chichester: Wiley.

Greenhouse, S. W. & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24,* 95–112.

Greuel, L. (1998). Der Einsatz des Polygraphen in der Behandlung und Überwachung hoch rückfallgefährdeter Sexualstraftäter. *Praxis der Rechtspsychologie, 8,* 54–70.

Greuel, L., Offe, S., Fabian, A., Wetzels, P., Fabian, T., Offe, H. & Stadler, M. (1998). *Glaubhaftigkeit der Zeugenaussage*. Weinheim: Psychologie Verlags Union.

Grossman, P. (1992). Respiratory and cardiac rhythms as windows to central and autonomic biobehavioral regulation: Selection of window frames, keeping the panes clean and viewing the neural topography. *Biological Psychology, 34,* 131–161.

Groves, P. M. & Thompson, R. F. (1970). Habituation: A dual-process theory. *Psychological Review, 77,* 419–450.

Groves, P. M. & Thompson, R. F. (1973). A dual-process theory of habituation: Neural mechanisms. In H. V. S. Peeke & M. J. Herz (Eds.), *Habituation, Vol. II Physiological substrates* (pp. 175–205). New York: Academic Press.

Gudjonsson, G. H. (1988). How to defeat the polygraph tests. In A. Gale (Ed.), *The polygraph test: Lies, truth and science* (pp. 126–136). London: Sage Publications.

Gudjonsson, G. H. (1999). *The psychology of interrogations, confessions and testimony*. Chichester: Wiley.

Gustafson, L. A. & Orne, M. T. (1963). Effects of heightened motivation on the detection of deception. *Journal of Applied Psychology, 47*, 408–411.

Gustafson, L. A. & Orne, M. T. (1964). The effects of task and method of stimulus presentation on the detection of deception. *Journal of Applied Psychology*, *48*, 383–387.

Gustafson, L. A. & Orne, M. T. (1965a). Effects of perceived role and role success on the detection of deception. *Journal of Applied Psychology, 49*, 412–417.

Gustafson, L. A. & Orne, M. T. (1965b). The effects of verbal responses on the laboratory detection of deception. *Psychophysiology, 2*, 10–13.

Hamm, R. (1999). Monokeltests und Menschenwürde. *Neue Juristische Wochenschrift (NJW), 52* (13), 922–923.

Heckel, R. V., Brokaw, J. R., Salzberg, H. C. & Wiggins, S. L. (1962). Polygraphic variations in reactivity between delusional, non-delusional, and control groups in a "crime" situation. *Journal of Criminal Law, Criminology and Police Science*, *53*, 380–383.

Heidenreich, K. (1984). Die Verwendung standardisierter Tests. In E. Roth (Hrsg.), *Sozialwissenschaftliche Methoden* (S. 399–416). München: Oldenbourg.

Hemsley, G., Heslegrave, R. J. & Furedy, J. J. (1980). Can deception be detected when stimulus familiarity is controlled? [Abstract]. *Psychophysiology, 17*, 286–287.

Heslegrave, R. J. (1982). An examination of the psychological mechanisms underlying deception [Abstract]. *Psychophysiology, 19*, 323.

Heslegrave, R. J. & Furedy, J. J. (1984). *Using cardiac responses to differentiate arousal during deception*. Unpublished manuscript. (zitiert nach Furedy, 1985, S. 252)

Höfling, S. (1998). Die Klärung des sexuellen Mißbrauchs durch den Polygraphen (Lügendetektor). Internationale Praxis, wissenschaftlicher Stand und juristische Entwicklung. *Politische Studien, 49* (Heft 359), 30–41.

Holstein, W. (1990). Technik und Methodik bei Wahrheits-Tests. *Kriminalistik*, *44* (3), 155–158.

Honts, C. R. (1991). The emperor's new clothes: Application of polygraph tests in the American workplace. *Forensic Reports, 4*, 91–116.

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. *The Journal of General Psychology, 123*, 309–324.

Honts, C. R., Devitt, M. K., Winbush, M. & Kircher, J. C. (1996). Mental and physical countermeasures reduce the accuracy of the concealed knowledge test. *Psychophysiology*, 33, 84–92.

Honts, C. R., Kircher, J. C. & Raskin, D. C. (1995). Polygrapher's dilemma or psychologist's chimaera: A reply to Furedy's logico-ethical considerations for psychophysiological practitioners and researchers. *International Journal of Psychophysiology, 20*, 199–207.

Honts, C. R. & Perry, M. V. (1992). Polygraph admissibility: Changes and challenges. *Law and Human Behavior, 16*, 357–379.

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. (1992). Effectiveness of control questions answered "yes": Dispelling a polygraph myth. *Forensic Reports, 5*, 265–272.

Honts, C. R., Raskin, D. C. & Kircher, J. C. (1994). Mental and physical countermeasures reduce the accuracy of polygraph tests. *Journal of Applied Psychology, 79*, 252–259.

Horneman, C. J. & O'Gorman, J. G. (1985). Detectability in the card test as a function of the subject's verbal response. *Psychophysiology, 22*, 330–333.

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.

Horvath, F. (1977). The effect of selected variables on interpretation of polygraph records. *Journal of Applied Psychology, 62*, 127–136.

Horvath, F. (1984). Detecting deception in eyewitness cases: Problems and prospects in the use of the polygraph. In G. L. Wells & E. F. Loftus (Eds.), *Eyewitness testimony* (pp. 214–255). Cambridge: Cambridge University Press.

Horvath, F. (1988). The utility of control questions and the effects of two control question types in field polygraph techniques. *Journal of Police Science and Administration, 16*, 198–209.

Iacono, W. G. (1991). Can we determine the accuracy of polygraph tests? In J. R. Jennings, P. K. Ackles & M. G. H. Coles (Eds.), *Advances in psychophysiology* (Vol. 4, pp. 201–207). London: Jessica Kingsley Publishers.

Iacono, W. G. (2000). The detection of deception. In J. T. Cacioppo, L. G. Tassinary & G. G. Berntson (Eds.), *Handbook of psychophysiology* (2nd ed., pp. 772–793). Cambridge: Cambridge University Press.

Iacono, W. G., Boisvenu, G. A. & Fleming, J. A. (1984). Effects of diazepam and methylphenidate on the electrodermal detection of guilty knowledge. *Journal of Applied Psychology, 69*, 289–299.

Iacono, W. G. & Lykken, D. T. (1997). The scientific status of research on polygraph techniques: The case against polygraph tests. In D. L. Faigman, D. H. Kaye, M. J. Saks & J. Sanders (Eds.), *Modern scientific evidence: The law and science of expert testimony* (Vol. 1, pp. 582–618). St. Paul, MN: West Publishing Company.

Iacono, W. G. & Patrick, C. J. (1987). What psychologists should know about lie detection. In I. B. Weiner & A. K. Hess (Eds.), *Handbook of forensic psychology* (pp. 460–489). New York: Wiley.

Iacono, W. G. & Patrick, C. J. (1988). Assessing deception: Polygraph techniques. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (pp. 205–233). New York: Guilford Press.

Jänig, W. (1987). Vegetatives Nervensystem. In R. F. Schmidt (Hrsg.), *Grundriß der Neurophysiologie* (6. Aufl., S. 221–274). Berlin: Springer.

Janisse, M. P. & Bradley, M. T. (1980). Deception, information and the pupillary response. *Perceptual and Motor Skills*, *50*, 748–750.

Jaworski, R. (1990). Der Lügendetektor auf dem Prüfstand. *Kriminalistik, 44* (3), 123–130.

Jaworski, R. (2000). Nochmals: Der Polygraph als Beweismittel. *Kriminalistik, 54* (1), 23–26.

Jennings, J. R., Nebes, R. D. & Yovetich, N. A. (1990). Aging increases the energetic demands of episodic memory: A cardiovascular analysis. *Journal of Experimental Psychology: General, 119*, 77–91.

Jennings, J. R., Tahmoush, A. J. & Redmond, D. P. (1980). Non-invasive measurement of peripheral vascular activity. In I. Martin & P. H. Venables (Eds.), *Techniques in psychophysiology* (pp. 69–137). Chichester: Wiley.

Jennings, J. R., van der Molen, M. W. & Brock, K. (1997). Mnemonic search, but not arithmetic transformation, is associated with psychophysiological inhibition. *Journal of Experimental Psychology: Human Perception and Performance, 23*, 154–167.

Johnson, L. C. (1963). Some attributes of spontaneous autonomic activity. *Journal of Comparative and Physiological Psychology*, *56*, 415–422.

Kallus, K. W. & Janke, W. (1988). Klassenzuordnung. In R. S. Jäger (Hrsg.), *Psychologische Diagnostik* (pp. 131–145). Weinheim: Psychologie Verlags Union.

Kargl, W. & Kirsch, S. (2000). Zur Zulässigkeit eines untauglichen Beweismittels im Strafverfahren – BGHSt 44, 308. *Juristische Schulung (JuS), 40* (6), 537–542.

Kassin, S. M. (1997). The psychology of confession evidence. *American Psychologist*, *52*, 221–233.

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.

Keeler, L. (1933). Scientific methods of crime detection with the polygraph. *Kansas Bar Association Journal*, *2*, 22–31. (zitiert nach Lykken, 1960, S. 260)

Kircher, J. C. & Raskin, D. C. (1981). Computerized decision-making in physiological detection of deception [Abstract]. *Psychophysiology*, *18*, 204–205.

Kircher, J. C. & Raskin, D. C. (1982). Is there a “specific lie pattern” of autonomic responses? [Abstract]. *Psychophysiology*, *19*, 569.

Kircher, J. C. & Raskin, D. C. (1988). Human versus computerized evaluations of polygraph data in a laboratory setting. *Journal of Applied Psychology*, *73*, 291–302.

Kleinmuntz, B. & Szucko, J. J. (1984). A field study of the fallibility of polygraphic lie detection. *Nature*, *308*, 449–450.

Klimke, O. (1981). Der Polygraphentest im Strafverfahren. *Neue Zeitschrift für Strafrecht (NStZ), 1* (11), 433–434.

Knögel, [Vorname nicht genannt]. (1954). Der Lügendetektor. *Deutsche Richterzeitung (DRiZ), 32* (11), 234–236.

Koepchen, H. P. (1982). Zentralnervöse und reflektorische Steuerung der Herzfrequenz. In B. Brisse & F. Bender (Hrsg.), *Autonome Innervation des Herzens: Medikamentöse Therapie bradykarder Rhythmusstörungen* (S. 66–85). Darmstadt: Steinhoff.

Koers, G., Gaillard, A. W. K. & Mulder, G. (1997). Evoked heart rate and blood pressure in an S1-S2 paradigm. *Biological Psychology*, *46*, 247–274.

Köhnken, G. (1990). *Glaubwürdigkeit: Untersuchungen zu einem psychologischen Konstrukt*. München: Psychologie Verlags Union.

Kugelmass, S. & Lieblich, I. (1966). Effects of realistic stress and procedural interference in experimental lie detection. *Journal of Applied Psychology*, *50*, 211–216.

Kugelmass, S., Lieblich, I. & Bergman, Z. (1967). The role of "lying" in psychophysiological detection. *Psychophysiology*, *3*, 312–315.

Lacey, J. I. (1967). Somatic response patterning and stress: Some revisions of activation theory. In M. H. Appley & R. Trumbull (Eds.), *Psychological stress: Issues in research* (pp. 14–37). New York: Appleton Century Crofts.

Lacey, J. I. & Lacey, B. C. (1958). Verification and extension of the principle of autonomic response-stereotypy. *American Journal of Psychology*, *71*, 50–73.

Lacey, B. C. & Lacey J. I. (1974). Studies of heart rate and other bodily processes in sensorimotor behavior. In P. A. Obrist, A. H. Black, J. Brener & L. V. DiCara (Eds.), *Cardiovascular Psychophysiology* (pp. 538–564). Chicago, IL: Aldine.

Lalumiere, M. L. & Quinsey, V. L. (1991a). Polygraph testing of child molesters: Are we ready? Part one of two. *Violence Update*, *1* (11), 3–11.

Lalumiere, M. L. & Quinsey, V. L. (1991b). Polygraph testing of child molesters: Are we ready? Part two of two. *Violence Update*, *1* (12), 6–7.

Larsen, P. B., Schneiderman, N. & Pasin, R. D. (1986). Physiological bases of cardiovascular psychophysiology. In M. G. H. Coles, E. Donchin & S. W. Porges (Eds.), *Psychophysiology: Systems, processes, and applications – A handbook* (pp. 122–165). New York: Guilford Press.

Levy, M. N. & Martin, P. J. (1979). Neural control of the heart. In R. M. Berne, N. Sperelakis & S. R. Geiger (Eds.), *Handbook of physiology: Section 2 The cardiovascular system, Vol. 1 The heart* (pp. 581–620). Bethesda, MD: American Physiological Society.

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.

Lieblich, I., Naftali, G., Shmueli, J. & Kugelmass, S. (1974). Efficiency of GSR detection of information with repeated presentation of series of stimuli in two motivational states. *Journal of Applied Psychology*, *59*, 113–115.

Lienert, G. A. & Raatz, U. (1994). *Testaufbau und Testanalyse* (5. Aufl.). Weinheim: Psychologie Verlags Union.

Lindsay, P. H. & Norman, D. A. (1972). *Human information processing: An introduction to psychology*. New York: Academic Press.

Loftus, G. (1982). Scientific and legal aspects on the use of the polygraph in trials: A panel discussion. In A. Trankell (Ed.), *Reconstructing the past: The role of psychologists in criminal trials* (pp. 373–389). Stockholm: Norstedt.

Lorig, T. S. & Schwartz, G. E. (1990). The pulmonary system. In J. T. Cacioppo & L. G. Tassinary (Eds.), *Principles of psychophysiology: Physical, social, and inferential elements* (pp. 580–598). Cambridge: Cambridge University Press.

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, 44,* 258–262.

Lykken, D. T. (1974). Psychology and the lie detector industry. *American Psychologist, 29,* 725–739.

Lykken, D. T. (1975). A reply to Abrams. *American Psychologist, 30,* 711–712.

Lykken, D. T. (1978). The psychopath and the lie detector. *Psychophysiology, 15,* 137–142.

Lykken, D. T. (1979). The detection of deception. *Psychological Bulletin, 86,* 47–53.

Lykken, D. T. (1981). *A tremor in the blood: Uses and abuses of the lie detector* (1st ed.). New York: McGraw-Hill.

Lykken, D. T. (1988). The case against polygraph testing. In A. Gale (Ed.), *The polygraph test: Lies, truth and science* (pp. 111–125). London: Sage Publications.

Lykken, D. T. (1991a). The lie detector controversy: An alternate solution. In J. R. Jennings, P. K. Ackles & M. G. H. Coles (Eds.), *Advances in psychophysiology* (Vol. 4, pp. 209–214). London: Jessica Kingsley Publishers.

Lykken, D. T. (1991b). 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.

Lynn, R. (1966). *Attention, arousal and the orientation reaction*. Oxford: Pergamon Press.

Marston, W. M. (1938). *The lie detector test*. New York: Smith.

Martin, S. P. (1999). Lügendetektor als ungeeignetes Beweismittel. *Juristische Schulung (JuS), 39* (7), 714–715.

Matte, J. A. (1996). *Forensic psychophysiology using the polygraph. Scientific truth verification – lie detection*. Williamsville, NY: J. A. M. Publications.

McCanne, T. R. & Lyons, G. M. (1990). Decelerative changes in heart rate are associated with performance on tasks that assess intelligence. *International Journal of Psychophysiology, 8,* 235–248.

McCrae, R. R. & Costa, P. T., Jr (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52,* 81–90.

McLean, P. D. (1969). Induced arousal and time of recall as determinants of paired-associate recall. *British Journal of Psychology, 60,* 57–62.

Munro, L. L., Dawson, M. E., Schell, A. M. & Sakai, L. M. (1987). Electrodermal lability and rapid vigilance decrement in a degraded stimulus continuous performance task. *Journal of Psychophysiology, 1*, 249–257.

Myrtek, M., Hilgenberg, B., Brügner, G. & Müller, W. (1997). Influence of sex, college major, and chronic study stress on psychophysiological reactivity and behavior: Results of ambulatory monitoring in students. *Journal of Psychophysiology, 11*, 124–137.

Näätänen, R. (1979). Orienting and evoked potentials. In H. D. Kimmel, E. H. van Olst & J. F. Orlebeke (Eds.), *The orienting reflex in humans* (pp. 61–75). Hillsdale, NJ: Lawrence Erlbaum Associates.

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology*. San Francisco, CA: Freeman.

Oberlandesgericht Bamberg. (1995). Beschluß vom 14. März 1995 – 7 WF 122/94. *Neue Juristische Wochenschrift (NJW), 48* (25), 1684–1685.

Offe, H. & Offe, S. (1999, September). *Das BGH-Urteil zum Polygraphen: Eine Herausforderung an die Psychologie*. Vortrag gehalten auf der 8. Arbeitstagung der Fachgruppe Rechtspsychologie der Deutschen Gesellschaft für Psychologie e. V., Nürnberg.

Office of Technology Assessment. (1983). *Scientific validity of polygraph testing: A research review and evaluation – A technical memorandum* (OTA-TM-H-15). Washingon, DC: U. S. Congress, Office of Technology Assessment.

O'Gorman, J. G. (1979). The orienting reflex: Novelty or significance detector? *Psychophysiology, 16*, 253–262.

O'Gorman, J. G. (1981). Novel uses for a novelty response [Besprechung des Buches *The orienting reflex in humans*]. *Contemporary Psychology, 26*, 787–788

O'Gorman, J. G. (1983). Individual differences in the orienting response. In D. A. T. Siddle (Ed.), *Orienting and habituation: Perspectives in human research* (pp. 431–448). Chichester: Wiley.

Olsen, D. E., Harris, J. C., Capps, M. H. & Ansley, N. (1997). Computerized polygraph scoring system. *Journal of Forensic Sciences, 42*, 61–70.

Orne, M. T., Thackray, R. I. & Paskewitz, D. A. (1972). On the detection of deception: A model for the study of physiological effects of psychological stimuli. In N. S. Greenfield & R. A. Sternbach (Eds.), *Handbook of psychophysiology* (pp. 743–785). New York: Holt, Rinehart & Winston.

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.

Otten, L. J., Gaillard, A. W. K. & Wientjes, C. J. E. (1995). The relation between event-related brain potential, heart rate, and blood pressure responses in an S_1-S_2 paradigm. *Biological Psychology, 39*, 81–102.

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.

Petermann, F. & Noack, H. (1984). Nicht-reaktive Meßverfahren. In E. Roth (Hrsg.), *Sozialwissenschaftliche Methoden* (S. 450–470). München: Oldenbourg.

Peters, K. (1975). Eine Antwort auf Undeutsch: Die Verwertbarkeit unwillkürlicher Ausdruckserscheinungen bei der Aussagenwürdigung. *Zeitschrift für die gesamte Strafrechtswissenschaft (ZStW), 87*, 663–679.

Podlesny, J. A. (1995). *A lack of operable case facts restricts applicability of the guilty knowledge deception detection method in FBI criminal investigations* [FBI Technical Report]. Quantico, VA: FBI. (zitiert nach Lykken, 1998, S. 305)

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.

Podlesny, J. A. & Truslow, C. M. (1993). Validity of an expanded-issue (modified general question) polygraph technique in a simulated distributed-crime-roles context. *Journal of Applied Psychology, 78*, 788–797.

Porges, S. W. & Byrne, E. A. (1992). Research methods for measurement of heart rate and respiration. *Biological Psychology, 34*, 93–130.

Raskin, D. C. (1978). Scientific assessment of the accuracy of detection of deception: A reply to Lykken. *Psychophysiology, 15*, 143–147.

Raskin, D. C. (1979). Orienting and defensive 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, NJ: Lawrence Erlbaum Associates.

Raskin, D. C. (1981). Science, competence, and polygraph techniques. *Criminal Defense, 8* (3), 11–18.

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. (1987). Methodological issues in estimating polygraph accuracy in field applications. *Canadian Journal of Behavioural Science, 19*, 389–404.

Raskin, D. C. (1988). Does science support polygraph testing? In A. Gale (Ed.), *The polygraph test: Lies, truth and science* (pp. 96–110). London: Sage Publications.

Raskin, D. C. (1989). Polygraph techniques for the detection of deception. In D. C. Raskin (Ed.), *Psychological methods in criminal investigation and evidence* (pp. 247–296). New York: Springer.

Raskin, D. C. & Hare, R. D. (1978). Psychopathy and detection of deception in a prison population. *Psychophysiology, 15*, 126–136.

Raskin, D. C., Honts, C. R. & Kircher, J. C. (1997). The scientific status of research on polygraph techniques: The case for polygraph tests. In D. L. Faigman, D. H. Kaye, M. J. Saks & J. Sanders (Eds.), *Modern scientific evidence: The law and science of expert testimony* (Vol. 1, pp. 565–582). St. Paul, MN: West Publishing Company.

Raskin D. C. & Kircher J. C. (1991). 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., Kircher, J. C., Horowitz, S. W. & Honts, C. R. (1988). Recent laboratory and field research on polygraph techniques. In J. C. Yuille (Ed.), *Credibility assessment* (pp. 1–24). Dordrecht: Kluwer.

Raskin, D. C. & Podlesny, J. A. (1979). Truth and deception: A reply to Lykken. *Psychological Bulletin, 86*, 54–59.

Raskin, D. C. & Steller, M. (1989). Assessing credibility of allegations of child sexual abuse: Polygraph examinations and statement analysis. In H. Wegener, F. Lösel & J. Haisch (Eds.), *Criminal behavior and the justice system. Psychological perspectives* (pp. 290–302). New York: Springer-Verlag.

Reed, S. (1994). A new psychophysiological detection of deception examination for security screening [Abstract]. *Psychophysiology*, *31*, S80.

Reid, J. E. (1947). A revised questioning technique in lie-detection tests. *Journal of Criminal Law and Criminology*, *37*, 542–547.

Reid, J. E. & Inbau, F. E. (1966). *Truth and deception. The polygraph ("lie-detector") technique* (1st ed.). Baltimore, MD: Williams and Wilkins.

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. (1997). *Zur psychophysiologischen Differenzierung von Täuschung und Aufrichtigkeit: Die Einflüsse des Wahrheitsgehalts, der elektrodermalen Labilität und des Antwortmodus auf Hautleitfähigkeitsreaktionen und das subjektive Befinden*. Unveröffentlichte Diplomarbeit, Johannes Gutenberg-Universität, Mainz.

Rill, H.-G. & Vossel, G. (1998). Psychophysiologische Täterschaftsbeurteilung („Lügendetektion", „Polygraphie"): Eine kritische Analyse aus psychophysiologischer und psychodiagnostischer Sicht. *Neue Zeitschrift für Strafrecht (NStZ)*, *18* (10), 481–486.

Rüddel, H. & Curio, I. (Eds.). (1991). *Non-invasive continuous blood pressure measurement*. Frankfurt/Main: Lang.

Salzgeber, J. & Stadler, M. (1997). Programm zur Behandlung von Sexualstraftätern – Vorstellung eines in den USA erprobten Interventionsprogrammes. *Politische Studien*, *48* (Sonderheft 2), 141–146.

Salzgeber, J., Stadler, M. & Vehrs, W. (1997). Die psychophysiologische Aussagebegutachtung im Rahmen des Familiengerichtsverfahrens. *Praxis der Rechtspsychologie, 7*, 213–221.

Sampson, J. R. (1969). Further study of encoding and arousal factors in free recall of verbal and visual material. *Psychonomic Science, 16*, 221–222.

Saxe, L. (1991). Science and the CQT polygraph: A theoretical critique. *Integrative Physiological and Behavioral Science*, *26*, 223–231.

Saxe, L. & Cross, T. P. (1991). Scientific evaluation of psychological technologies: Commentary on 'The forensic use of the polygraph'. In J. R. Jennings, P. K. Ackles & M. G. H. Coles (Eds.), *Advances in psychophysiology* (Vol. 4, pp. 225–230). London: Jessica Kingsley Publishers.

Schandry, R. (1998). *Lehrbuch Psychophysiologie: Körperliche Indikatoren psychischen Geschehens* (Studienausgabe). Weinheim: Psychologie Verlags Union.

Schell, A. M., Dawson, M. E. & Filion, D. L. (1988). Psychophysiological correlates of electrodermal lability. *Psychophysiology, 25,* 619–632.

Schumacher, A. (1993). *Aufmerksamkeit und kognitive Bewältigungsstrategien im Kontext einer psychophysiologischen Aussagebeurteilung*. Unveröffentlichte Dissertation, Johannes Gutenberg-Universität, Mainz.

Schwabe, J. (1979). Rechtsprobleme des „Lügendetektors". *Neue Juristische Wochenschrift (NJW), 32* (12), 576–582.

Schwabe, J. (1982). Der „Lügendetektor" vor dem Bundesverfassungsgericht. *Neue Juristische Wochenschrift (NJW), 35* (8), 367–368.

Schwartz, P. J. & Weiss, T. (1983). T-wave amplitude as an index of cardiac sympathetic activity: A misleading concept. *Psychophysiology, 20*, 696–701.

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.

Siddle, D. A. T., Remington, B., Kuiack, M. & Haines, E. (1983). Stimulus omission and dishabituation of the skin conductance response. *Psychophysiology, 20,* 136–145.

Simons, R. F. (1988). Event-related slow brain potentials: A perspective from ANS psychophysiology. In P. K. Ackles, J. R. Jennings & M. G. H. Coles (Eds.), *Advances in psychophysiology* (Vol. 3, pp. 223–267). Greenwich: JAI Press.

Simons, R. F. (1989). 'A rose by any other name': A comment on Vossel and Zimmer. *Journal of Psychophysiology, 3*, 125–127.

Sokolov, E. N. (1963). *Perception and the conditioned reflex*. Oxford: Pergamon Press.

Sokolov, E. N. (1966). Orienting reflex as information regulator. In A. Leontyev, A. Luriya & A. Smirnov (Eds.), *Psychological research in the U.S.S.R.* (Vol. 1, pp. 334–360). Moscow: Progress Publishers.

Sokolov, E. N. (1969). The modeling properties of the nervous system. In M. Cole & I. Maltzman (Eds.), *A handbook of contemporary Soviet psychology* (pp. 671–704). New York: Basic Books.

Steinke, W. (1987). Lügendetektor zugunsten des Beschuldigten? *Monatsschrift für Deutsches Recht (MDR), 41* (7), 535–537.

Steller, M. (1983a). Psychophysiologische Aussagebeurteilung – Zum Stand der wissenschaftlichen „Lügendetektion". *Psychologische Beiträge, 25*, 459–493.

Steller, M. (1983b). Validität und forensische Verwendbarkeit von Methoden der psychophysiologischen Aussagebeurteilung („Lügendetektion"). In G. Lüer (Hrsg.), *Bericht über den 33. Kongreß der Deutschen Gesellschaft für Psychologie in Mainz, 1982* (Band 2, S. 887–894). Göttingen: Hogrefe.

Steller, M. (1987). *Psychophysiologische Aussagebeurteilung*. Göttingen: Hogrefe.

Steller, M. (1997). Psychophysiologische Täterschaftsermittlung („Lügendetektion“, „Polygraphie“). In M. Steller & R. Volbert (Hrsg.), *Psychologie im Strafverfahren* (S. 89–104). Bern: Huber.

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. & Dahle, K.-P. (1999). Wissenschaftliches Gutachten: Grundlagen, Methoden und Anwendungsprobleme psychophysiologischer Aussage- bzw. Täterschaftsbeurteilung („Polygraphie“, „Lügendetektion“). *Praxis der Rechtspsychologie, 9* (Sonderheft), 127–204.

Steller, M., Haenert, P. & Eiselt, W. (1987). Extraversion and the detection of information. *Journal of Research in Personality, 21*, 334–342.

Steller, M. & Volbert, R. (1997). Glaubwürdigkeitsbegutachtung. In M. Steller & R. Volbert (Hrsg.), *Psychologie im Strafverfahren* (S. 12–39). Bern: Huber.

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.

Stern, R. M. & Sison, C. E. E. (1990). Response patterning. In J. T. Cacioppo & L. G. Tassinary (Eds.), *Principles of psychophysiology: Physical, social, and inferential elements* (pp. 193–215). Cambridge: Cambridge University Press.

Szucko, J. J. & Kleinmuntz, B. (1981). Statistical versus clinical lie detection. *American Psychologist, 36*, 488–496.

Tent, L. (1967). Psychologische Tatbestandsdiagnostik (Spurensymptomatologie, Lügendetektion). In U. Undeutsch (Hrsg.), *Handbuch der Psychologie: Band 11 Forensische Psychologie* (S. 185–259). Göttingen: Hogrefe.

Thackray, R. I. & Orne, M. T. (1968). A comparison of physiological indices in detection of deception. *Psychophysiology, 4*, 329–339.

Thornton, P. (1988). Lie detection and civil liberties in the UK. In A. Gale (Ed.), *The polygraph test: Lies, truth and science* (pp. 150–158). London: Sage Publications.

Timm, H. W. (1982). Effect of altered outcome expectancies stemming from placebo and feedback treatments on the validity of the guilty knowledge technique. *Journal of Applied Psychology, 67*, 391–400.

Turpin, G. (1989). An adequate test of the habituation of the cardiac decelerative response component of the orienting reflex: Necessary conditions and sufficient evidence. A comment on Vossel and Zimmer. *Journal of Psychophysiology, 3*, 129–140.

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. (1975). Die Verwertbarkeit unwillkürlicher Ausdruckserscheinungen bei der Aussagenwürdigung. Eine Anfrage von psychologischer Seite. *Zeitschrift für die gesamte Strafrechtswissenschaft (ZStW), 87*, 650–662.

Undeutsch, U. (1979). Die Leistungsfähigkeit der heutigen Methoden der psychophysiologischen Täterschaftsermittlung. *Monatsschrift für Kriminologie und Strafrechtsreform (MschrKrim), 62* (4), 228–241.

Undeutsch, U. (1983). Vernehmung und non-verbale Information. In E. Kube, H. U. Störzer & S. Brugger (Hrsg.), *BKA-Forschungsreihe: Band 16 Wissenschaftliche Kriminalistik. Grundlagen und Perspektiven, Teilband 1 Systematik und Bestandsaufnahme* (S. 389–418). Wiesbaden: Bundeskriminalamt.

Undeutsch, U. (1996). Die Untersuchung mit dem Polygraphen („Lügendetektor“) – eine wissenschaftliche Methode zum Nachweis der Unschuld. *Zeitschrift für das gesamte Familienrecht (FamRZ), 43 (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.

Undeutsch, U. & Klein, G. (1999). Wissenschaftliches Gutachten zum Beweiswert physiopsychologischer Untersuchungen. *Praxis der Rechtspsychologie, 9* (Sonderheft), 45–126.

van der Veen, F. M., van der Molen, M. W. & Jennings, J. R. (2000). Selective inhibition is indexed by heart rate slowing. *Psychophysiology, 37*, 607–613.

van Olst, E. H., Heemstra, M. L. & ten Kortenaar, T. (1979). Stimulus significance and the orienting reaction. In H. D. Kimmel, E. H. van Olst & J. F. Orlebeke (Eds.), *The orienting reflex in humans* (pp. 521–547). Hillsdale, NJ: Lawrence Erlbaum Associates.

Velden, M. (1994). *Psychophysiologie: Eine kritische Einführung*. München: Quintessenz.

Velden, M. (1999). New aspects when depicting heart rate and blood pressure over time? Comment on Koers et al. *Journal of Psychophysiology, 13*, 92–94.

Velden, M. & Graham, F. K. (1988). Depicting heart rate over real time: Two procedures that are mathematically identical. *Journal of Psychophysiology, 2*, 291–292.

Velden, M., Karemaker, J. M., Wölk, C. & Schneider, R. (1990). Inferring vagal effects on the heart from changes in cardiac cycle length: Implications for cycle time-dependency. *International Journal of Psychophysiology, 10*, 85–93.

Velden, M. & Vossel, G. (1985). How can skin conductance responses increase over trials while skin resistance responses decrease? *Physiological Psychology, 13*, 291–295.

Velden, M. & Wölk, C. (1987). Depicting cardiac activity over real time: A proposal for standardization. *Journal of Psychophysiology, 1*, 173–175.

Venables, P. H. & Christie, M. J. (1980). Electrodermal activity. In I. Martin & P. H. Venables (Eds.), *Techniques in psychophysiology* (pp. 3–67). Chichester: Wiley.

Vincent, A. & Furedy, J. J. (1992). Electrodermal differentiation of deception: Potentially confounding and influencing factors. *International Journal of Psychophysiology, 13*, 129–136.

Volckart, B. (1998). Das Verwertungsverbot für Lügendetektortests. *Recht & Psychiatrie, 16* (3), 138–144.

Vorberg, D. & Blankenberger, S. (1999). Die Auswahl statistischer Tests und Maße. *Psychologische Rundschau*, *50*, 157–164.

Vossel, G. (1990). *Elektrodermale Labilität: Ein Beitrag zur differentiellen Psychophysiologie*. Göttingen: Hogrefe.

Vossel, G., Gödert, H. W. & Rill, H.-G. (2001). „Lügendetektion“: Zufällig, zuverlässig? Methoden der forensischen Psychophysiologie. *Psychoscope, 22* (4), 6–9.

Vossel, G. & Rossmann, R. (1984). Electrodermal habituation speed and visual monitoring performance. *Psychophysiology, 21*, 97–100.

Vossel, G. & Zimmer, H. (1989a). Heart rate deceleration as an index of the orienting response? *Journal of Psychophysiology*, *3*, 111–124.

Vossel, G. & Zimmer, H. (1989b). ‘Roses have thorns and silver fountains mud’: A reply to Simons and Turpin. *Journal of Psychophysiology*, *3*, 141–146.

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.

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. & Orne M. T. (1981). Cognitive, social, and personality processes in the physiological detection of deception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 14, pp. 61–106). New York: Academic Press.

Waid, W. M., Orne, E. C., Cook, M. R. & Orne, M. T. (1978). Effects of attention, as indexed by subsequent memory, on electrodermal detection of information. *Journal of Applied Psychology, 63*, 728–733.

Waid, W. M., Orne, E. C. & Orne, M. T. (1981). Selective memory for social information, alertness, and physiological arousal in the detection of deception. *Journal of Applied Psychology*, *66*, 224–232.

Waid, W. M., Orne, M. T. & Wilson, S. K. (1979). Effects of level of socialization on electrodermal detection of deception. *Psychophysiology*, *16*, 15–22.

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. (1976). *Zur Beschreibung von Aktivierungsprozessen. Eine Methodenstudie zur psychophysiologischen Diagnostik*. Unveröffentlichte Dissertation, Albert-Ludwigs-Universität, Freiburg.

Walter, G. F. & Porges, S. W. (1976). Heart rate and respiratory responses as a function of task difficulty: The use of discriminant analysis in the selection of psychologically sensitive physiological responses. *Psychophysiology*, *13*, 563–571.

Wells, G. L. & Loftus, E. F. (Eds.). (1984). *Eyewitness testimony*. Cambridge: Cambridge University Press.

Wilder, J. (1931). Das „Ausgangswert-Gesetz“, ein unbeachtetes biologisches Gesetz und seine Bedeutung für Forschung und Praxis. *Zeitschrift für Neurologie*, *137*, 317–338.

Williams, V. L. (1995). Response to Cross and Saxe's "A critique of the validity of polygraph testing in child sexual abuse cases". *Journal of Child Sexual Abuse*, *4* (3), 55–71.

Wingard, J. A. & Maltzman, I. (1980). Interest as a predeterminer of the GSR index of the orienting reflex. *Acta Psychologica*, *46*, 153–160.

Wölk, C. & Velden, M. (1989). Revision of the baroreceptor hypothesis on the basis of a new cardiac cycle effect. In N. W. Bond & D. A. T. Siddle (Eds.), *Psychobiology: Issues and applications* (pp. 371–379). Amsterdam: Elsevier.

Wölk, C., Velden, M., Zimmermann, U. & Krug, S. (1989). The interrelation between phasic blood pressure and heart rate changes in the context of the 'baroreceptor hypothesis'. *Journal of Psychophysiology*, *3*, 397–402.

Yamamura, T. & Miyata, Y. (1990). Development of the polygraph technique in Japan for detection of deception. *Forensic Science International*, *44*, 257–271.

Yankee, W. J. (1995). The current status of research in forensic psychophysiology and its application in the psychophysiological detection of deception. *Journal of Forensic Sciences*, *40*, 63–68

Zallmanzig, A. (1999). Das „Lügendetektorurteil" des BGH – doch ein Recht eines Beschuldigten auf polygraphische Untersuchung. *JuraThek* [Internet-Zeitschrift]. Verfügbar unter: http://www.jurathek.de/tom/urteile/besprech_bgh_luegen.html [2000-07-24].

Zimmer, H., Vossel, G. & Fröhlich, W. D. (1989). Antizipatorische und ereignis-bezogene Veränderungen der Schlagfrequenz des Herzens als Indikatoren der Aufmerksamkeitsregulation. *Archiv für Psychologie*, *141*, 251–272.

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:

Abriß der Rechtsprechung zur „Lügendetektion“ in Deutschland

In Deutschland hat der **Bundesgerichtshof** (BGH) bereits 1954 die Verwendung von Polygraphen im Strafverfahren und in den Vorermittlungen mit Verweis auf die verfassungsrechtlich geschützte Menschenwürde des Angeklagten abgelehnt (Urteil vom 16. 2. 1954 – 1 StR 578/53; zum Argument der Menschenwürde in der rechtswissenschaftlichen Polygraphie-Diskussion vgl. Kargl & Kirsch, 2000, S. 539ff.). Ferner wurden „Lügendetektortests“ unter die verbotenen Vernehmungsmethoden gemäß § 136 a Strafprozeßordnung (StPO) subsumiert, da die Freiheit der Willensentschließung und Willensbetätigung des Untersuchten beeinträchtigt werde (BGH, 1954, S. 332f.). Der Senat begründete seine Entscheidung wie folgt: Unter der Voraussetzung, daß die Grundannahmen des Verfahrens zuträfen, könne der Proband keine der Untersuchungsfragen willentlich beantworten oder übergehen, ohne daß gleichzeitig unwillkürliche Körperreaktionen auftreten würden, die der Untersucher mittels des Polygraphen messe und zur Beurteilung der Glaubwürdigkeit heranziehe. Folglich sei die Freiheit des Angeklagten, bewußt zu entscheiden, inwiefern er sich zum Tatvorwurf äußern wolle, nicht gewährleistet. Damit wurde der „Lügendetektor“ strafprozessual für unzulässig erklärt, analog zu anderen Maßnahmen der Willensbeeinträchtigung, die im § 136 a StPO exemplarisch aufgelistet sind. Dabei handelt es sich z. B. um Mißhandlung, Ermüdung, Verabreichung von Mitteln, Quälerei, Täuschung und Hypnose. Das Beweiserhebungs- und Verwertungsverbot galt unabhängig von der Einwilligung des Untersuchten und hing „nicht von der Brauchbarkeit des Polygraphen zur Aufklärung von Straftaten ab und auch nicht von der Richtigkeit und Verläßlichkeit der wissenschaftlichen Erwägungen, auf denen er beruht“ (BGH, 1954, S. 333).

Im Grundsatz wurde das BGH-Urteil später vom **Bundesverfassungsgericht** (BVerfG) bestätigt (Beschluß vom 18. 8. 1981 – 2 BvR 166/81), das in einer derartigen „Durchleuchtung“ des Beschuldigten mittels des „Lügendetektors“ einen Eingriff in das per Grundgesetz (GG, Artikel 2 Absatz 1 in Verbindung mit Artikel 1 Absatz 1) gewahrte Persönlichkeitsrecht sah (BVerfG, 1981, S. 446f.). Ohne an dieser Stelle im Detail darauf eingehen zu wollen, sei vermerkt, daß beide Urteile in juristischen Fachkreisen umstritten waren und z. T. heftig kritisiert wurden (z. B. Amelung, 1982, S. 38ff.; Klimke, 1981, S. 434; Knögel, 1954, S. 235f.; Schwabe, 1979, S. 578ff., 1982, S. 367f.). Die Diskussion kreiste v. a. um die Frage, ob man einem (eventuell zu Unrecht) Beschuldigten unter Berufung auf seine Menschenwürde und Persönlichkeitsrechte prinzipiell die Möglichkeit verwehren darf, mittels eines freiwillig durchgeführten Tests einen Unschuldsnachweis zu erbringen (vgl. zusammenfassend Eisenberg, 1993, S. 122ff.; Steller, 1987, S. 156ff.). Zudem blieb streitig, inwiefern die Verstärkung und Aufzeichnung physiologischer Reaktionen durch Meßgeräte einen qualitativen Unterschied dar-

stellt zu einer nicht apparativen Registrierung offensichtlicher körperlicher Begleiterscheinungen des Aussagenden (z. B. Erröten, nervöse Bewegungen, verkrampfte Haltung). Laut BGH (1954, S. 335) sind solche „offen hervortretenden Ausdrucksbewegungen“ durchaus der richterlichen Beweiswürdigung zugänglich und dürfen vom Gericht mit „Vorsicht, Zurückhaltung und Menschenkenntnis“ bei der Glaubwürdigkeitsbeurteilung berücksichtigt werden (vgl. auch die Kontroverse zwischen Undeutsch, 1975, S. 658f., und Peters, 1975, S. 669ff.).

Jahrzehntelang hatte die deutsche Judikatur an dem höchstrichterlichen Verwertungsverbot festgehalten (zusammenfassend Berning, 1992, S. 223–229). Anfang der neunziger Jahre zeichnete sich hier jedoch eine **Trendwende** ab. Diese betraf zunächst familien- und vormundschaftsgerichtliche Streitigkeiten (vgl. Undeutsch, 1996, S. 331, 1997, S. 303ff.). Meist handelte es sich um Umgangs- und Sorgerechtsfälle, in denen Eltern, die des sexuellen Mißbrauchs an ihren Kindern bezichtigt wurden, entlastende Testergebnisse als Unschuldsnachweis in die Verfahren einbrachten. Dabei wurden „Lügendetektor“-Untersuchungen von mehreren Zivilgerichten als objektive Beweismittel gewürdigt (z. B. Oberlandesgericht Bamberg, 1995, S. 1684). Dieser plötzliche „Boom“ stand in einem krassen Mißverhältnis zum damaligen Stand der grundwissenschaftlichen Auseinandersetzung mit der „Lügendetektion“ in Deutschland. Bis auf wenige Ausnahmen in Form von Übersichtsarbeiten (z. B. Berning, 1992; Steller, 1987; Tent, 1967; Undeutsch, 1983), Kurzdarstellungen in psychophysiologischen Monographien oder Lehrbüchern (z. B. Boucsein, 1988, S. 422–430; Schandry, 1998, S. 321–325) und vereinzelten empirischen Studien (z. B. Curio & Scholz, 1991; Schumacher, 1993) fand die Thematik in der hiesigen psychologischen bzw. psychophysiologischen Fachliteratur kaum Resonanz.

Parallel zum zivilrechtlichen Bereich gab es Bemühungen, die psychophysiologische Aussagebegutachtung auch im **Strafrecht** zu etablieren. Die Entwicklung mündete schließlich darin, daß erstmals ein deutsches Strafgericht explizit die „Lügendetektion“ als Entlastungsnachweis anerkannte (Amtsgericht Demmin/Zweigstelle Malchin, Urteil vom 7. 9. 1998 – 94 Ls 182/98) und sich der BGH erneut mit der Notwendigkeit konfrontiert sah, eine Grundsatzentscheidung zu fällen. In dem entsprechenden Urteil vom 17. Dezember 1998 lehnte der 1. Strafsenat des BGH die strafrechtliche Verwertung psychophysiologischer Untersuchungen mit dem Polygraphen („Lügendetektor“) ab. Diesmal jedoch nicht aus verfassungsrechtlichen, sondern aus methodischen Gründen, da die gängigen Verfahren zu einem völlig ungeeigneten Beweismittel im Sinne des § 244 Absatz 3 StPO führen würden (BGH, 2000, S. 308).

Anhang B:

Deskriptive Statistik der Anzahl elektrodermaler Spontanfluktuationen (NSRs) während der fünfminütigen Ruhemessung

Tabelle 28. Deskriptive Statistik der NSR-Anzahl: Mittelwert (*M*), Standardfehler (*SE*), Standardabweichung (*SD*), Bereich (*B*, Minimum – Maximum), Schiefe (*S*) und Exzeß (*E*), nach Testart (DLT vs. GAT) und EL (Stabil vs. Labil) getrennt und gesamt

		Elektrodermale Labilität	
Testart		Stabil	Labil
DLT	*M*	14.85	45.10
(*n* = 40 Pbn)	*SE*	2.33	2.75
	SD	10.42	12.32
	B	0 – 28	30 – 80
	S	-0.26	1.53
	E	-1.63	2.66
GAT	*M*	11.80	48.85
(*n* = 40 Pbn)	*SE*	2.17	3.09
	SD	9.70	13.82
	B	0 – 28	30 – 76
	S	0.32	0.43
	E	-1.57	-0.88
		Stabile und Labile zusammen	
Gesamt	*M*	30.15	
(*N* = 80 Pbn)	*SE*	2.29	
	SD	20.51	
	B	0 – 80	
	S	0.32	
	E	-0.57	

Anmerkungen. *n* = Teilstichprobe, *N* = Gesamtstichprobe.

<u>Anhang C</u>:

Schriftliche Versuchsmaterialien – Information, Einverständniserklärung, Instruktionen, Gedächtnistests, Rating- und Protokollbögen

C.1: Information und Einverständniserklärung
C.2: Instruktion 1a: Anweisung zum Scheinverbrechen (Tatbedingung Ring)
C.3: Instruktion 1b: Anweisung zum Scheinverbrechen (Tatbedingung Kette)
C.4: Instruktion 2: Anweisung zum Rollenverhalten
C.5: Instruktion 3: Vorbereitung auf psychophysiologische Aussagebeurteilung
C.6: Instruktion 4: Anweisung für Ruhemessung
C.7: Instruktion 5: Anweisung für Stimulationstest
C.8: Instruktion 6a: Anweisung für DLT
C.9: Instruktion 6b: Anweisung für GAT
C.10: Rating 1a: Reaktionsstärkerating für DLT (Auszug)
C.11: Rating 1b: Reaktionsstärkerating für GAT (Auszug)
C.12: Rating 2a: Bedeutsamkeitsrating für DLT (Auszug)
C.13: Rating 2b: Bedeutsamkeitsrating für GAT (Auszug)
C.14: Gedächtnistest 1: Wiedererinnern mit Hinweisreizen (Auszug)
C.15: Gedächtnistest 2: Wiedererkennen (Auszug)
C.16: Postexperimentelle Ratings: Motivation und subjektive Treffsicherheit
C.17: Protokollbogen für Nachbefragung
C.18: Protokollbogen zur Erfassung der soziodemographischen Daten
C.19: Protokollbogen für Ablauf DLT
C.20: Protokollbogen für Ablauf GAT

<u>Anhang C.1</u>: Information und Einverständniserklärung

JOHANNES GUTENBERG-UNIVERSITÄT MAINZ
Psychologisches Institut
Abteilung Allgemeine Experimentelle Psychologie
Interdisziplinäre Forschungsgruppe Forensische Psychophysiologie

Dipl.-Psych. Hans-Georg Rill

Staudingerweg 9 . D-55099 Mainz . Telefon 06131 / 39-2795 . Fax 39-2480 . E-mail rill@psych.Uni-Mainz.de

<u>Experiment zur Lügendetektion</u>

<u>Zu Ihrer Information</u>

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.

Im Rahmen des Experiments wird eine **Belohnung von DM 20,-** in Aussicht gestellt. Die Vergabe dieser Belohnung kann nur an jene Versuchsteilnehmer erfolgen, die mit dem **Lügendetektortest** als „**unschuldig**“ eingestuft werden. Es besteht kein grundsätzlicher Anspruch auf eine Entlohnung.

Alle Versuchsteilnehmer werden am Ende des Experiments über dessen Hintergründe **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.

<u>Einverständniserklärung</u>

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 lediglich der Anspruch auf die Belohnung von DM 20,-. Ferner erkläre ich mich mit den oben genannten Bedingungen **einverstanden** und verpflichte mich, bis zum Ende der gesamten Versuchsreihe (voraussichtlich Herbst 1999) keine Detailinformationen über das Experiment weiterzugeben.

Nachname: _____ Vorname: _____

Mainz, den _____ Unterschrift: _____

Im Anschluß an das Experiment erhalten Sie eine Kopie dieser Einverständniserklärung.

<u>Anhang C.2</u>: Instruktion 1a: Anweisung zum Scheinverbrechen (Tatbedingung Ring)

Instruktion 1a (Ring)

Sehr geehrter Versuchsteilnehmer,

vielen Dank, daß Sie sich bereit erklärt haben, an dieser wissenschaftlichen Studie zur **Lügendetektion** teilzunehmen! Der Versuch dauert ca. **2 Stunden**.

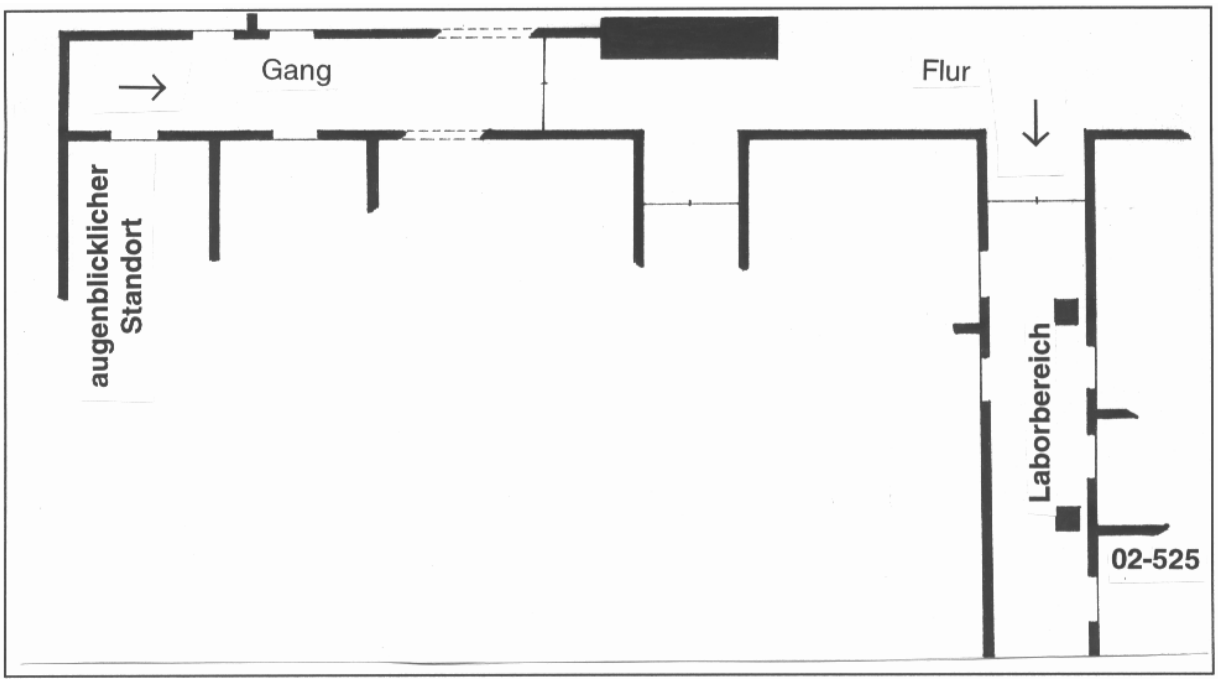
In diesem Experiment gibt es **2 Gruppen**:
Täter (Schuldige) und **Nicht-Täter (Unschuldige)**.
Sie wurden nach dem Zufallsprinzip den **Tätern** zugeteilt.

Im folgenden sollen Sie einen **simulierten Schmuckdiebstahl** begehen und anschließend einen **Lügendetektortest** absolvieren.

Bitte begeben Sie sich nach dem Lesen dieser Anweisung in den **Raum 02-525**.
Der **Raum 02-525** dient als Arbeitszimmer für Diplomanden und Tutoren. Er befindet sich ebenfalls im **2. Stockwerk** dieses Gebäudes, und zwar im **Laborbereich**.
Die genaue Lage können Sie dem **Plan** unten auf diesem Blatt entnehmen.
Gehen Sie bitte auf direktem Weg dorthin!

An der Glastür zum Eingang des Laborbereichs ist ein **Schild** angebracht mit der Aufschrift: „Abteilung Allgemeine Experimentelle Psychologie - Laborbereich - **Kein Durchgang**“. Ignorieren Sie dieses Verbot. Legen Sie sich jedoch eine **plausible Ausrede** zurecht, falls Sie von Mitarbeitern der Abteilung nach dem Grund Ihrer Anwesenheit gefragt werden. Auf keinen Fall dürfen Sie erwähnen, daß Sie an einem psychologischen Experiment teilnehmen.

Falls Sie sich als Versuchsteilnehmer zu erkennen geben, müssen wir Sie leider vom Experiment ausschließen.



Bitte umblättern!

Bitte lesen Sie die folgenden Anweisungen aufmerksam durch. Halten Sie sich bei der Durchführung des Diebstahls genau an die vorgegebene Reihenfolge. Achten Sie auf alle Einzelheiten!

1. Betreten Sie den Raum 02-525 und schließen Sie die Tür! In dem Raum steht ein **Schreibtisch** mit mehreren Schubladen. Zwei dieser **Schubladen** sind mit Nummern versehen.
2. Gehen Sie zum Schreibtisch.
3. Öffnen Sie die **Schublade Nr. 3**. Darin liegt neben einem **Hefter** ein durchsichtiger **Plastikkasten**. In dem Plastikkasten befindet sich eine **Halskette** mit einem **gelben Stein**. An der Kette ist ein Etikett mit dem Buchstaben „**A**“ befestigt.
4. **Prägen Sie sich den Inhalt der Schublade Nr. 3 gut ein, ohne die Gegenstände zu berühren**.
5. Schließen Sie die Schublade Nr. 3 wieder.
6. Öffnen Sie dann die **Schublade Nr. 6**. Darin liegt neben einem **Locher** eine durchsichtige **Glasschale.** In der Glasschale befindet sich ein **Ring** mit einem **blauen Stein**. Am Ring ist ein Etikett mit dem Buchstaben „**E**“ befestigt.
7. **Prägen Sie sich den Inhalt der Schublade Nr. 6 gut ein**.
8. **Öffnen** Sie die Glasschale.
9. **Nehmen Sie den Ring** und verstecken Sie ihn in einer Ihrer **Hosentaschen**, so daß man ihn von außen nicht erkennen kann.
10. Schließen Sie die Schublade und **verlassen Sie den Raum**.
11. Schließen Sie die Tür und **kommen Sie wieder hierher zurück**.

Bitte lesen Sie jetzt die Punkte 1 bis 11 nochmals durch und prägen Sie sich den Ablauf gut ein.

Falls Sie Fragen haben, wenden Sie sich bitte an den Versuchsleiter.

<u>Anhang C.3</u>: Instruktion 1b: Anweisung zum Scheinverbrechen (Tatbedingung Kette)

Instruktion 1b (Kette)

Sehr geehrter Versuchsteilnehmer,

vielen Dank, daß Sie sich bereit erklärt haben, an dieser wissenschaftlichen Studie zur **Lügendetektion** teilzunehmen! Der Versuch dauert ca. **2 Stunden**.

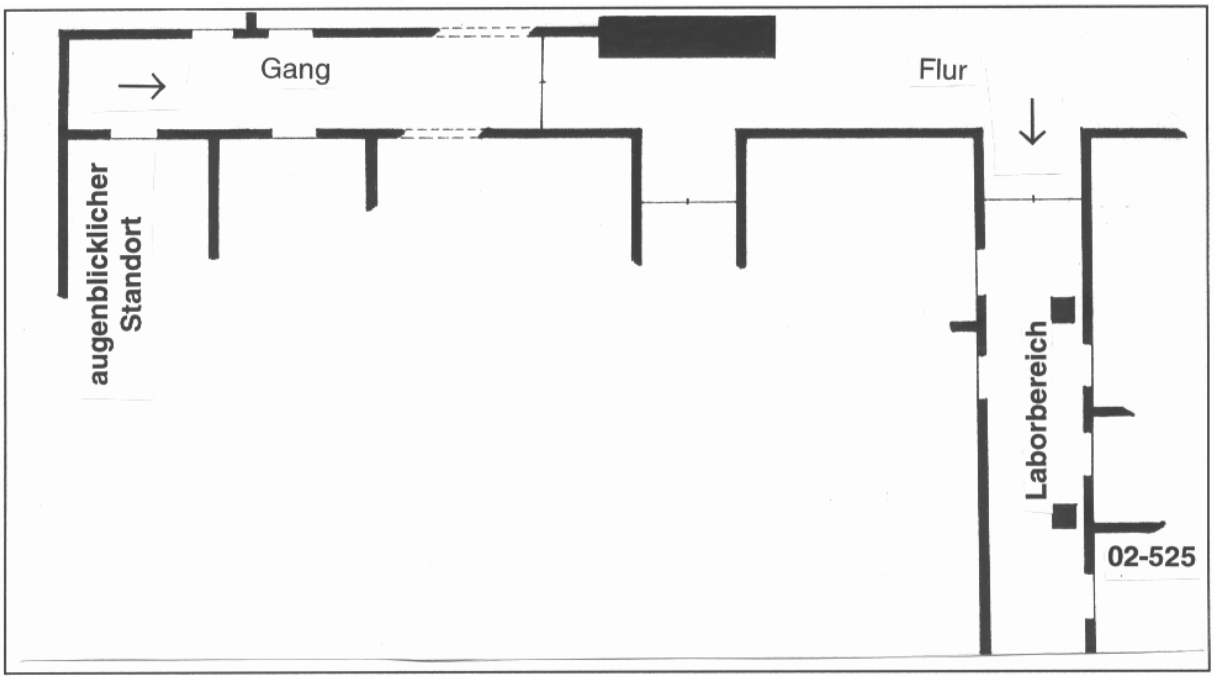
In diesem Experiment gibt es **2 Gruppen**:
Täter (Schuldige) und **Nicht-Täter (Unschuldige)**.
Sie wurden nach dem Zufallsprinzip den **Tätern** zugeteilt.

Im folgenden sollen Sie einen **simulierten Schmuckdiebstahl** begehen und anschließend einen **Lügendetektortest** absolvieren.

Bitte begeben Sie sich nach dem Lesen dieser Anweisung in den **Raum 02-525**.
Der **Raum 02-525** dient als Arbeitszimmer für Diplomanden und Tutoren. Er befindet sich ebenfalls im **2. Stockwerk** dieses Gebäudes, und zwar im **Laborbereich**.
Die genaue Lage können Sie dem **Plan** unten auf diesem Blatt entnehmen.
Gehen Sie bitte auf direktem Weg dorthin!

An der Glastür zum Eingang des Laborbereichs ist ein **Schild** angebracht mit der Aufschrift: „Abteilung Allgemeine Experimentelle Psychologie - Laborbereich - **Kein Durchgang**“. Ignorieren Sie dieses Verbot. Legen Sie sich jedoch eine **plausible Ausrede** zurecht, falls Sie von Mitarbeitern der Abteilung nach dem Grund Ihrer Anwesenheit gefragt werden. Auf keinen Fall dürfen Sie erwähnen, daß Sie an einem psychologischen Experiment teilnehmen.

Falls Sie sich als Versuchsteilnehmer zu erkennen geben, müssen wir Sie leider vom Experiment ausschließen.



Bitte umblättern!

Bitte lesen Sie die folgenden Anweisungen aufmerksam durch. Halten Sie sich bei der Durchführung des Diebstahls genau an die vorgegebene Reihenfolge. Achten Sie auf alle Einzelheiten!

1. Betreten Sie den Raum 02-525 und schließen Sie die Tür! In dem Raum steht ein **Schreibtisch** mit mehreren Schubladen. Zwei dieser **Schubladen** sind mit Nummern versehen.
2. Gehen Sie zum Schreibtisch.
3. Öffnen Sie die **Schublade Nr. 6**. Darin liegt neben einem **Locher** eine durchsichtige **Glasschale.** In der Glasschale befindet sich ein **Ring** mit einem **blauen Stein**. Am Ring ist ein Etikett mit dem Buchstaben „**E**“ befestigt.
4. **Prägen Sie sich den Inhalt der Schublade Nr. 6 gut ein, ohne die Gegenstände zu berühren**.
5. Schließen Sie die Schublade Nr. 6 wieder.
6. Öffnen Sie dann die **Schublade Nr. 3**. Darin liegt neben einem **Hefter** ein durchsichtiger **Plastikkasten**. In dem Plastikkasten befindet sich eine **Halskette** mit einem **gelben Stein**. An der Kette ist ein Etikett mit dem Buchstaben „**A**“ befestigt.
7. **Prägen Sie sich den Inhalt der Schublade Nr. 3 gut ein**.
8. **Öffnen** Sie den Plastikkasten.
9. **Nehmen Sie die Kette** und verstecken Sie sie in einer Ihrer **Hosentaschen**, so daß man sie von außen nicht erkennen kann.
10. Schließen Sie die Schublade und **verlassen Sie den Raum**.
11. Schließen Sie die Tür und **kommen Sie wieder hierher zurück**.

Bitte lesen Sie jetzt die Punkte 1 bis 11 nochmals durch und prägen Sie sich den Ablauf gut ein.

Falls Sie Fragen haben, wenden Sie sich bitte an den Versuchsleiter.

<u>Anhang C.4</u>: Instruktion 2: Anweisung zum Rollenverhalten

Instruktion 2

Bitte stellen Sie sich nun folgende Geschichte vor:

Sie und andere Personen werden des Schmuckdiebstahls verdächtigt. **Sie streiten den Diebstahl entschieden ab**. Darum bietet man Ihnen an, die Ermittlungen gegen Sie einzustellen, falls Sie einen **Lügendetektortest** bestehen. Sie willigen ein.

Der **Untersucher**, der die Lügendetektion durchführt, weiß nicht, zu welcher Gruppe Sie gehören („schuldig“ oder „unschuldig“?). Außerdem ist er nicht darüber informiert, welches der beiden Schmuckstücke jeweils gestohlen wurde. Er wird versuchen, mit einem Lügendetektortest festzustellen, ob Sie den Diebstahl begangen haben, und wenn ja, welches Schmuckstück Sie entwendet haben.

- **Falls Sie unschuldig sind**, müssen Sie während des Lügendetektortests nur Ihre Unschuld beteuern.

- **Falls Sie den Schmuckdiebstahl begangen haben**, sollen Sie die Tat unter allen Umständen abstreiten. Verhalten Sie sich stets wie ein Unschuldiger. Achten Sie darauf, daß Sie sich nicht selbst verraten. **Sie dürfen sich unter keinen Umständen als Täter zu erkennen geben.**

Wenn Sie den Lügendetektortest bestehen und Sie anhand der Untersuchung als **„unschuldig“ einschätzt** werden, erhalten Sie eine **Belohnung von DM 20,-**.

Wenn Sie mit dem Lügendetektortest als **„schuldig“ eingestuft** werden, erhalten Sie **keine Belohnung**.

Falls Sie Fragen haben, wenden Sie sich bitte an den Versuchsleiter.

<u>Anhang C.5</u>: Instruktion 3: Vorbereitung auf psychophysiologische Aussagebeurteilung

Instruktion 3

Im folgenden sollen Sie einen **Lügendetektortest** absolvieren.

Bei der **Lügendetektion** müssen Sie unterschiedliche Fragen beantworten. Das Ziel der Untersuchung besteht darin, anhand Ihrer körperlichen Reaktionen den Wahrheitsgehalt der Antworten zu bestimmen.

Die Lügendetektion beruht darauf, daß man beim Lügen aufgeregter ist als beim wahrheitsgemäßen Antworten. Diese Aufregung zeigt sich auch in körperlichen Veränderungen (z.B. Schwitzen, Herzklopfen, Veränderung der Atmung).
Die Reaktionen werden mit speziellen Geräten gemessen.
Jede Lüge führt zu körperlichen Veränderungen.

Wissenschaftliche Studien in den USA zeigen für die Lügendetektion **sehr hohe Trefferquoten**. D.h., die Wahrscheinlichkeit ist sehr hoch, daß festgestellt wird, ob Sie zu den Tätern oder den Unschuldigen gehören. Die Studien deuten aber auch darauf hin, daß besonders **intelligente** Personen, die ihre **Gefühlsregungen gut kontrollieren** können, in der Lage sind, den Lügendetektor zu **überlisten**.

Um Ihre körperlichen Reaktionen messen zu können, werden nun zunächst einige **Meßfühler** angelegt. Es handelt sich dabei um:

1. drei Elektroden am Oberkörper zur Messung der Herztätigkeit,
2. zwei Elektroden an der Hand zur Messung der Schweißdrüsenaktivität,
3. einen Gurt am Oberkörper zur Messung der Atmung,
4. zwei Meßfühler an den Fingern zur Erfassung der Durchblutung und Temperatur.

Bitte entfernen Sie eventuell an diesen Stellen vorhandenen Schmuck. Nach dem Anbringen der Elektroden nehmen Sie bitte in der Kabine Platz. Dort werden Sie an den Lügendetektor angeschlossen. Danach folgen zunächst eine Ruhemessung und zwei weitere Voruntersuchungen, bevor der eigentliche Lügendetektortest beginnt.

Bitte melden Sie sich beim Untersucher, wenn Sie diese Anweisung durchgelesen haben.

Falls Sie Fragen haben, wenden Sie sich bitte an den Untersucher.

Anhang C.6: Instruktion 4: Anweisung für Ruhemessung

Instruktion 4

Vor der eigentlichen Untersuchung mit dem Lügendetektor erfolgt eine **6-minütige Ruhemessung**. Diese wird benötigt, um Ihre körperliche Aktivität während einer Ruhephase zu erfassen und um die Meßinstrumente genau einzustellen.

Die gesamte Messung ist absolut schmerzfrei und ungefährlich.

Bitte setzen Sie sich **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** etc.

Die **Kamera** in der Kabine dient lediglich dazu, mögliche Meßfehler aufgrund von Bewegungen erkennen zu können.

Beginn und Ende der Ruhemessung werden Ihnen mitgeteilt. Nach der Ruhemessung erhalten Sie die genauen Anweisungen zur Untersuchung mit dem Lügendetektor.

Bitte melden Sie sich beim Untersucher, wenn Sie diese Anweisung durchgelesen haben.

Falls Sie Fragen haben, wenden Sie sich bitte an den Untersucher.

<u>Anhang C.7</u>: Instruktion 5: Anweisung für Stimulationstest

Instruktion 5

Als nächstes wird überprüft, ob Sie für den anschließenden Lügendetektortest **geeignet** sind. Dazu muß der Untersucher messen, wie Sie auf **einfache Fragen** reagieren, wenn Sie **wahrheitsgemäß** oder **wahrheitswidrig antworten**.

Im folgenden sollen Sie eine beliebige **Zahl zwischen 11 und 16 wählen** und diese dem **Untersucher mitteilen**.

Anschließend werden Ihnen folgende **Fragen nach der Zahl** gestellt:

Welche Zahl haben Sie gewählt?
- **War es die 11?**
- **die 12?**
- **die 13?**
- **die 14?**
- **die 15?**
- **die 16?**

Diese Fragen werden **computergesteuert** auf dem **Bildschirm** eingeblendet und gleichzeitig über **Lautsprecher** dargeboten. Dabei werden Ihre körperlichen **Reaktionen** gemessen. Die körperlichen Reaktionen **dauern eine gewisse Zeit** an. Um sie vollständig messen zu können, liegen zwischen den einzelnen Fragen **Pausen** von ca. 20 Sekunden Länge. Außerdem ist es wichtig, daß Sie erst **nach der Ausblendung der Frage** auf dem Bildschirm antworten.

1. Beantworten Sie jede Frage laut und deutlich mit „**Nein**“.
2. Antworten Sie erst, nachdem die Frage **auf dem Bildschirm ausgeblendet** wurde. Sagen Sie sonst nichts.
3. Versuchen Sie, möglichst **glaubwürdig** zu erscheinen.
4. **Bewegen Sie sich möglichst wenig**. Bleiben Sie während der gesamten Untersuchung **entspannt**.

Wenn Sie stets mit „**Nein**“ antworten, müssen Sie auf eine der Fragen **lügen**. Anhand Ihrer **Reaktionen auf diese Lüge** werden die Meßinstrumente eingestellt. Außerdem wird bestimmt, ob Sie für die Untersuchung mit dem Lügendetektor **geeignet** sind. Wenn Sie dabei **zu schwach** reagieren, müssen wir Sie leider vom Experiment ausschließen.

Bitte wählen Sie nun eine Zahl zwischen 11 und 16 und teilen Sie diese dem Untersucher mit.

Bitte melden Sie sich beim Untersucher, wenn Sie diese Anweisung durchgelesen haben.

Falls Sie Fragen haben, wenden Sie sich bitte an den Untersucher.

<u>Anhang C.8</u>: Instruktion 6a: Anweisung für DLT

Instruktion 6a (DLT)

Das Psychologische Institut der Universität Mainz hat **ein neuartiges Programm zur Lügendetektion** entwickelt. Mit diesem Experiment wollen wir die **Treffsicherheit des Programms** überprüfen.

Im folgenden werden Ihnen **15 verschiedene Fragen** gestellt, die sich in **4 Gruppen** aufteilen lassen:

1. Belanglose Fragen, die sich auf Ihren momentanen Aufenthaltsort beziehen und nichts mit dem Diebstahl und Ihrer Glaubwürdigkeit zu tun haben:

- Befinden Sie sich momentan in Mainz?
- Befinden Sie sich momentan an der Universität?
- Befinden Sie sich momentan am Psychologischen Institut?

➔ **Beantworten** Sie diese Fragen stets **wahrheitsgemäß** mit „**Ja**“.
Diese Fragen sind für die Beurteilung Ihrer Glaubwürdigkeit **unwichtig**. Sie dienen lediglich dazu, Ihre **Reaktionen auf belanglose Fragen** zu bestimmen.

2. Fragen, die sich direkt auf den Schmuckdiebstahl beziehen:

- Haben Sie die Glasschale geöffnet?
- Haben Sie den Plastikkasten geöffnet?
- Haben Sie den Ring gestohlen?
- Haben Sie die Kette gestohlen?
- Tragen Sie den gestohlenen Ring bei sich?
- Tragen Sie die gestohlene Kette bei sich?

➔ **Beantworten** Sie diese Fragen stets mit „**Nein**“. Streiten Sie den Diebstahl ab!

3. Vergleichsfragen, um Ihre typischen Reaktionen beim Lügen bestimmen zu können:

- Handeln Sie manchmal illegal?
- Sind Sie manchmal unehrlich?
- Sagen Sie manchmal die Unwahrheit?

➔ **Beantworten** Sie diese Fragen stets **wahrheitswidrig** mit „**Nein**“!
Denken Sie bitte bei der Beantwortung **an konkrete Situationen**, in denen Sie illegal gehandelt haben, unehrlich waren bzw. die Unwahrheit gesagt haben.
Achten Sie dabei auf Ihre **emotionalen Reaktionen**.

Wir gehen davon aus, daß fast jeder Mensch im Laufe seines Lebens derartige Missetaten begeht. Dennoch sollen Sie diese Vergleichsfragen **wahrheitswidrig** mit „**Nein**“ beantworten, damit festgestellt werden kann, wie Ihr Körper beim Lügen reagiert.

4. Vergleichsfragen, um Ihre typischen Reaktionen bei wahrheitsgemäßen Antworten bestimmen zu können:

- Handeln Sie stets legal?
- Sind Sie immer ehrlich?
- Sagen Sie immer die Wahrheit?

➔ **Beantworten** Sie diese Fragen stets **wahrheitsgemäß** mit „**Nein**“!
Denken Sie bitte bei der Beantwortung **an konkrete Situationen**, in denen Sie illegal gehandelt haben, unehrlich waren bzw. die Unwahrheit gesagt haben.
Achten Sie dabei auf Ihre **emotionalen Reaktionen**.

Sie sollen diese Fragen **wahrheitsgemäß** mit „**Nein**“ beantworten, damit festgestellt werden kann, wie Ihr Körper bei wahrheitsgemäßen Antworten reagiert.

Nur wenn Sie auf die Vergleichsfragen ganz **bewußt lügen bzw. die Wahrheit sagen**, resultieren **angemessene Reaktionen**. Diese werden mit den Reaktionen auf die Fragen nach dem Diebstahl **verglichen**. Dadurch kann Ihre **Glaubwürdigkeit** bestimmt werden.

Falls Sie keine angemessenen Reaktionen auf die Vergleichsfragen zeigen, besteht ein hohes **Risiko**, daß Sie als „**schuldig**“ eingestuft werden. Dann würden Sie die **Belohnung** von DM 20,- **verlieren**.

Diese 15 Fragen werden nun **computergesteuert** auf dem Bildschirm eingeblendet und gleichzeitig über **Lautsprecher** dargeboten. Zwischen den einzelnen Fragen liegen jeweils **Pausen** von ca. 20 Sekunden. In einem **zweiten Durchgang** werden die gleichen Fragen nochmals in veränderter Reihenfolge präsentiert.

Bitte lesen Sie die folgenden Anweisungen aufmerksam durch:

1. Beantworten Sie alle Fragen **laut und deutlich**.
2. Antworten Sie erst, nachdem die Frage **auf dem Bildschirm ausgeblendet** wurde. Sagen Sie sonst nichts.
3. Beantworten Sie die Fragen nach Ihrem momentanen **Aufenthaltsort** stets mit „**Ja**“.
4. Beantworten Sie alle anderen Fragen, also die nach dem **Diebstahl** und die **Vergleichsfragen**, stets mit „**Nein**“.
5. Versuchen Sie, möglichst **glaubwürdig** zu erscheinen.
6. Bleiben Sie während der gesamten Untersuchung **entspannt**. **Bewegen Sie sich möglichst wenig**.

Eventuelle Versuche, die Messung zu stören, können entdeckt werden. In diesem Fall werden Sie als „**schuldig**“ eingestuft und **verlieren die Belohnung**.

Zu Beginn werden **5 Fragen zum Üben** dargeboten.
Die entsprechenden Fragen und Antworten sind:

- **Sitzen Sie auf einem Stuhl? → Ja!**
- **Haben Sie die Schublade Nr. 3 geöffnet? → Nein!**
- **Haben Sie die Schublade Nr. 6 geöffnet? → Nein!**
- **Sind Sie immer aufrichtig? → Nein!**
- **Sind Sie manchmal unaufrichtig? → Nein!**

Bitte befolgen Sie die Regeln!

Bitte melden Sie sich beim Untersucher, wenn Sie diese Anweisung durchgelesen haben.

Falls Sie noch Fragen haben, wenden Sie sich bitte an den Untersucher.

<u>Anhang C.9</u>: Instruktion 6b: Anweisung für GAT

Instruktion 6b (GAT)

Das Psychologische Institut der Universität Mainz hat ein **neuartiges Programm zur Lügendetektion** entwickelt. Mit diesem Experiment wollen wir die **Treffsicherheit des Programms** überprüfen.

Im folgenden werden Ihnen **mehrere Fragen** nach Details des Schmuckdiebstahls gestellt. Die Fragen beziehen sich auf folgende **Details**:

1. In welchem **Raum** haben Sie den Schmuckdiebstahl begangen?
2. Welches **Schmuckstück** haben Sie gestohlen?
3. Aus welcher **Schublade** haben Sie den Schmuck gestohlen?
4. In welchem **Gefäß** befand sich das von Ihnen gestohlene Schmuckstück?
5. Welcher **Gegenstand** befand sich noch in der Schublade, aus der Sie den Schmuck gestohlen haben?
6. Welche **Farbe** hat der Stein des von Ihnen gestohlenen Schmuckstücks?
7. Wie lautet der **Buchstabe** auf dem Etikett des von Ihnen gestohlenen Schmuckstücks?

Zu jedem Detail werden sechs Fragen gestellt, z.B.:

In welchem Raum haben Sie den Schmuckdiebstahl begangen?
- **War es Raum 02-522?**
- **Raum 02-528?**
- **Raum 02-525?**
- **Raum 02-523?**
- **Raum 02-527?**
- **Raum 02-521?**

Diese Fragen werden **computergesteuert** auf dem **Bildschirm** eingeblendet und gleichzeitig über **Lautsprecher** dargeboten. Zwischen den einzelnen Fragen liegen jeweils **Pausen** von ca. 20 Sekunden.

Bitte lesen Sie die folgenden Anweisungen aufmerksam durch:

1. Beantworten Sie jede Frage **laut und deutlich** mit „**Nein**“. Streiten Sie den Diebstahl ab.
2. Antworten Sie erst, nachdem die Frage **auf dem Bildschirm ausgeblendet** wurde. Sagen Sie sonst nichts.
3. Versuchen Sie, möglichst **glaubwürdig** zu erscheinen.
4. Bleiben Sie während der gesamten Untersuchung **entspannt**. **Bewegen Sie sich möglichst wenig**.

Wenn Sie **unschuldig** sind, sagen Sie bei allen Fragen **die Wahrheit**.
Wenn Sie zu den **Tätern** gehören, müssen Sie jeweils auf **eine der Fragen lügen**.

Eventuelle Versuche, die Messung zu stören, können entdeckt werden. In diesem Fall werden Sie als „**schuldig**“ eingestuft und **verlieren die Belohnung**.

Zu Beginn werden **mehrere Fragen zum Üben** dargeboten. Es handelt sich um das oben genannte Beispiel. Die entsprechenden Fragen und Antworten sind:

In welchem Raum haben Sie den Schmuckdiebstahl begangen?

- **War es Raum 02-522? → Nein!**
- **Raum 02-528? → Nein!**
- **Raum 02-525? → Nein!**
- **Raum 02-523? → Nein!**
- **Raum 02-527? → Nein!**
- **Raum 02-521? → Nein!**

Bitte befolgen Sie die Regeln!

Bitte melden Sie sich beim Untersucher, wenn Sie diese Anweisung durchgelesen haben.

Falls Sie noch Fragen haben, wenden Sie sich bitte an den Untersucher.

<u>Anhang C.10</u>: Rating 1a: Reaktionsstärkerating für DLT (Auszug)

Rating 1a (DLT) **Vp:_____**

Die folgenden Fragebögen werden **nicht zur Beurteilung Ihrer Täterschaft herangezogen**. Der Untersucher, der den Lügendetektortest durchgeführt hat, erhält keinen Einblick in die Ergebnisse. Sie können also frei antworten, **ohne Ihre Täterschaft verbergen** zu müssen.

Auf diesem Bogen sind nochmals **alle Fragen** des Lügendetektortests abgedruckt.

Bitte versuchen Sie nun abzuschätzen, **wie stark Sie <u>Ihrer Meinung nach</u> auf die Fragen reagiert haben**.

Sie haben die Möglichkeit, zwischen **7 Abstufungen** zu wählen. Bitte **kreuzen** Sie bei jeder Frage diejenige **Intensitätsstufe** an, welche die Stärke Ihrer körperlichen Reaktionen auf die Fragen am besten beschreibt, z.B.:

Auf diese Frage habe ich ...

gar nicht reagiert	kaum reagiert	eher schwach reagiert	mäßig reagiert	deutlich reagiert	stark reagiert	sehr stark reagiert
1	2	3	4	X 5	6	7

Es gibt keine richtigen oder falschen Antworten. Sie sollen eine **persönliche Einschätzung** Ihrer Empfindungen abgeben.
Bitte achten Sie darauf, daß Sie **keine Frage auslassen**.

Falls Sie Fragen haben, wenden Sie sich bitte an den Versuchsleiter.

1. **Befinden Sie sich momentan in Mainz?**

Auf diese Frage habe ich ...

gar nicht reagiert	kaum reagiert	eher schwach reagiert	mäßig reagiert	deutlich reagiert	stark reagiert	sehr stark reagiert
1	2	3	4	5	6	7

2. **Haben Sie den Ring gestohlen?**

Auf diese Frage habe ich ...

gar nicht reagiert	kaum reagiert	eher schwach reagiert	mäßig reagiert	deutlich reagiert	stark reagiert	sehr stark reagiert
1	2	3	4	5	6	7

3. **Sind Sie immer ehrlich?**

Auf diese Frage habe ich ...

gar nicht reagiert	kaum reagiert	eher schwach reagiert	mäßig reagiert	deutlich reagiert	stark reagiert	sehr stark reagiert
1	2	3	4	5	6	7

Bitte umblättern!

<u>Anhang C.11</u>: Rating 1b: Reaktionsstärkerating für GAT (Auszug)

Rating 1b (GAT) **Vp:_____**

Die folgenden Fragebögen werden **nicht zur Beurteilung Ihrer Täterschaft herangezogen**. Der Untersucher, der den Lügendetektortest durchgeführt hat, erhält keinen Einblick in die Ergebnisse. Sie können also frei antworten, **ohne Ihre Täterschaft verbergen** zu müssen.

Auf diesem Bogen sind nochmals **alle Fragen** des Lügendetektortests abgedruckt.

Bitte versuchen Sie nun abzuschätzen, **wie stark Sie <u>Ihrer Meinung nach</u> auf die Fragen reagiert haben**.

Sie haben die Möglichkeit, zwischen **7 Abstufungen** zu wählen. Bitte **kreuzen** Sie bei jeder Frage diejenige **Intensitätsstufe** an, welche die Stärke Ihrer körperlichen Reaktionen auf die Fragen am besten beschreibt, z.B.:

Auf diese Frage habe ich ...

gar nicht reagiert	kaum reagiert	eher schwach reagiert	mäßig reagiert	deutlich reagiert	stark reagiert	sehr stark reagiert
1	2	3	4	X 5	6	7

Es gibt keine richtigen oder falschen Antworten. Sie sollen eine **persönliche Einschätzung** Ihrer Empfindungen abgeben.
Bitte achten Sie darauf, daß Sie **keine Frage auslassen**.

Falls Sie Fragen haben, wenden Sie sich bitte an den Versuchsleiter.

1. Welches Schmuckstück haben Sie gestohlen?

- **War es eine Krawattennadel?**

Auf diese Frage habe ich ...

gar nicht reagiert	kaum reagiert	eher schwach reagiert	mäßig reagiert	deutlich reagiert	stark reagiert	sehr stark reagiert
1	2	3	4	5	6	7

- **eine Uhr?**

Auf diese Frage habe ich ...

gar nicht reagiert	kaum reagiert	eher schwach reagiert	mäßig reagiert	deutlich reagiert	stark reagiert	sehr stark reagiert
1	2	3	4	5	6	7

- **eine Brosche?**

Auf diese Frage habe ich ...

gar nicht reagiert	kaum reagiert	eher schwach reagiert	mäßig reagiert	deutlich reagiert	stark reagiert	sehr stark reagiert
1	2	3	4	5	6	7

Bitte umblättern!

<u>Anhang C.12</u>: Rating 2a: Bedeutsamkeitsrating für DLT (Auszug)

Rating 2a (DLT) **Vp:_____**

Auf diesem Bogen sind nochmals **alle Fragen** des Lügendetektortests abgedruckt.

Sie sollen nun einschätzen, **wie wichtig die einzelnen Fragen für das Ergebnis des Lügendetektortests, d.h. für die Beurteilung Ihrer Glaubwürdigkeit bzw. Täterschaft waren**. Sie haben die Möglichkeit, zwischen **7 Abstufungen** zu wählen:

Diese Frage war für die Beurteilung meiner Glaubwürdigkeit ...

gar nicht wichtig	kaum wichtig	eher wenig wichtig	mäßig wichtig	eher wichtig	sehr wichtig	äußerst wichtig
1	2	X 3	4	5	6	7

Es gibt keine richtigen oder falschen Antworten. Sie sollen **Ihre persönliche Einschätzung** abgeben. Bitte achten Sie darauf, daß Sie **keine Frage auslassen**.

Falls Sie Fragen haben, wenden Sie sich bitte an den Versuchsleiter.

1. **Befinden Sie sich momentan in Mainz?**

Diese Frage war für die Beurteilung meiner Glaubwürdigkeit ...

gar nicht wichtig	kaum wichtig	eher wenig wichtig	mäßig wichtig	eher wichtig	sehr wichtig	äußerst wichtig
1	2	3	4	5	6	7

2. **Haben Sie den Ring gestohlen?**

Diese Frage war für die Beurteilung meiner Glaubwürdigkeit ...

gar nicht wichtig	kaum wichtig	eher wenig wichtig	mäßig wichtig	eher wichtig	sehr wichtig	äußerst wichtig
1	2	3	4	5	6	7

3. **Sind Sie immer ehrlich?**

Diese Frage war für die Beurteilung meiner Glaubwürdigkeit ...

gar nicht wichtig	kaum wichtig	eher wenig wichtig	mäßig wichtig	eher wichtig	sehr wichtig	äußerst wichtig
1	2	3	4	5	6	7

4. **Haben Sie die Kette gestohlen?**

Diese Frage war für die Beurteilung meiner Glaubwürdigkeit ...

gar nicht wichtig	kaum wichtig	eher wenig wichtig	mäßig wichtig	eher wichtig	sehr wichtig	äußerst wichtig
1	2	3	4	5	6	7

5. **Sind Sie manchmal unehrlich?**

Diese Frage war für die Beurteilung meiner Glaubwürdigkeit ...

gar nicht wichtig	kaum wichtig	eher wenig wichtig	mäßig wichtig	eher wichtig	sehr wichtig	äußerst wichtig
1	2	3	4	5	6	7

Bitte umblättern!

<u>Anhang C.13</u>: Rating 2b: Bedeutsamkeitsrating für GAT (Auszug)

Rating 2b (GAT) **Vp:_____**

Auf diesem Bogen sind nochmals **alle Fragen** des Lügendetektortests abgedruckt.

Sie sollen nun einschätzen, **wie wichtig die einzelnen Fragen für das Ergebnis des Lügendetektortests, d.h. für die Beurteilung Ihrer Glaubwürdigkeit bzw. Täterschaft waren**. Sie haben die Möglichkeit, zwischen **7 Abstufungen** zu wählen:

Diese Frage war für die Beurteilung meiner Glaubwürdigkeit ...

gar nicht wichtig	kaum wichtig	eher wenig wichtig	mäßig wichtig	eher wichtig	sehr wichtig	äußerst wichtig
1	2	X 3	4	5	6	7

Es gibt keine richtigen oder falschen Antworten. Sie sollen **Ihre persönliche Einschätzung** abgeben. Bitte achten Sie darauf, daß Sie **keine Frage auslassen**.

Falls Sie Fragen haben, wenden Sie sich bitte an den Versuchsleiter.

1. Welches Schmuckstück haben Sie gestohlen?

- **War es eine Krawattennadel?**

Diese Frage war für die Beurteilung meiner Glaubwürdigkeit ...

gar nicht wichtig	kaum wichtig	eher wenig wichtig	mäßig wichtig	eher wichtig	sehr wichtig	äußerst wichtig
1	2	3	4	5	6	7

- **eine Uhr?**

Diese Frage war für die Beurteilung meiner Glaubwürdigkeit ...

gar nicht wichtig	kaum wichtig	eher wenig wichtig	mäßig wichtig	eher wichtig	sehr wichtig	äußerst wichtig
1	2	3	4	5	6	7

- **eine Brosche?**

Diese Frage war für die Beurteilung meiner Glaubwürdigkeit ...

gar nicht wichtig	kaum wichtig	eher wenig wichtig	mäßig wichtig	eher wichtig	sehr wichtig	äußerst wichtig
1	2	3	4	5	6	7

- **ein Ring?**

Diese Frage war für die Beurteilung meiner Glaubwürdigkeit ...

gar nicht wichtig	kaum wichtig	eher wenig wichtig	mäßig wichtig	eher wichtig	sehr wichtig	äußerst wichtig
1	2	3	4	5	6	7

Bitte umblättern!

<u>Anhang C.14</u>: Gedächtnistest 1: Wiedererinnern mit Hinweisreizen (Auszug)

Gedächtnistest 1 **Vp:_____**

Mit den folgenden **Gedächtnistests** wollen wir überprüfen, inwiefern Sie sich an wichtige Einzelheiten des Schmuckdiebstahls erinnern können. Dazu werden Ihnen nun einige Fragen zum Diebstahl gestellt. Bitte geben Sie die **zutreffenden** Antworten.

Der **Untersucher**, der den Lügendetektortest durchgeführt hat, erhält **keinen Einblick** in die Ergebnisse der Gedächtnistests. D.h., in diesen Fragebögen können Sie sich ohne weiteres **als Täter zu erkennen geben**.

Bitte beantworten Sie die folgenden Fragen schriftlich, z.B.:

In welchem Raum fand der Schmuckdiebstahl statt?

Antwort: Raum 02-525

Wenn Sie sich **nicht** mehr an das entsprechende Detail **erinnern können**, dann überlegen Sie sich bitte eine Antwort, die Ihnen am **ehesten zutreffend** erscheint. Wenn Ihnen dabei nichts einfällt, dann machen Sie bitte einen **Strich** in die Lücke.

Falls Sie Fragen haben, wenden Sie sich bitte an den Versuchsleiter.

1. **Welches Schmuckstück haben Sie gestohlen?**

Antwort: _____

2. **Welches Schmuckstück haben Sie liegengelassen?**

Antwort: _____

3. **Welche Nummer stand auf der Schublade, in der sich das gestohlene Schmuckstück befand?**

Antwort: _____

4. **Welche Nummer stand auf der Schublade, in der sich das liegengelassene Schmuckstück befand?**

Antwort: _____

5. **In welchem Gefäß befand sich das gestohlene Schmuckstück?**

Antwort: _____ **Bitte umblättern!**

<u>Anhang C.15</u>: Gedächtnistest 2: Wiedererkennen (Auszug)

Gedächtnistest 2 **Vp:_____**

Bei den folgenden Fragen sollen Sie die **zutreffenden Antworten** aus mehreren Möglichkeiten **aussuchen** und **ankreuzen**. Auch bei diesem Test erhält der Lügendetektor-Untersucher keine Informationen über die Ergebnisse.

Zu jeder Frage gibt es nur **eine richtige Antwort**, z.B.:

In welchem Raum fand der Schmuckdiebstahl statt?
❍ Raum 02-522
❍ Raum 02-528
☒ Raum 02-525
❍ Raum 02-523
❍ Raum 02-527
❍ Raum 02-521

Wenn Sie sich **nicht** mehr an das entsprechende Detail **erinnern** können, dann versuchen Sie die Antwort zu **erraten**. Dazu wählen Sie bitte eine **Antwort**, die Ihnen am **wahrscheinlichsten** bzw. **plausibelsten** erscheint.

1. **Welches Schmuckstück haben Sie gestohlen?**
❍ eine Krawattennadel
❍ eine Uhr
❍ eine Brosche
❍ ein Ring
❍ ein Armreif
❍ eine Kette

2. **Welches Schmuckstück haben Sie liegengelassen?**
❍ eine Krawattennadel
❍ eine Uhr
❍ eine Brosche
❍ ein Ring
❍ ein Armreif
❍ eine Kette

3. **Welche Nummer stand auf der Schublade, in der sich das gestohlene Schmuckstück befand?**
❍ Nr. 5
❍ Nr. 3
❍ Nr. 4
❍ Nr. 1
❍ Nr. 6
❍ Nr. 2

Bitte umblättern!

<u>Anhang C.16</u>: Postexperimentelle Ratings: Motivation und subjektive Treffsicherheit

Postexperimentelle Ratings **Vp:_____**

Bitte beantworten Sie nun noch einige allgemeine Fragen zum Experiment. Dabei gibt es keine richtigen oder falschen Antworten. **Wichtig ist, daß Sie Ihre persönlichen Einschätzungen und Meinungen abgeben**. Hierzu sollen Sie entsprechend der Fragestellung jeweils ein Feld ankreuzen.

1. **Wie hoch war Ihre Motivation, dem Untersucher die Lügendetektion zu erschweren?**

Ich war ...

gar nicht motiviert	kaum motiviert	etwas motiviert	ziemlich motiviert	deutlich motiviert	stark motiviert	äußerst motiviert
1	2	3	4	5	6	7

... dem Untersucher die Lügendetektion zu erschweren.

2. **Wie gut ist es Ihrer Einschätzung nach dem Untersucher gelungen, nur anhand Ihrer körperlichen Reaktionen wahrheitsgemäße Antworten und Lügen zu unterscheiden?**

Meiner Einschätzung nach ist es dem Untersucher ...

nie gelungen	selten gelungen	manchmal gelungen	wiederholt gelungen	öfters gelungen	meistens gelungen	immer gelungen
1	2	3	4	5	6	7

... wahrheitsgemäße Antworten und Lügen zu unterscheiden.

Bitte melden Sie sich beim Versuchsleiter, wenn Sie mit der Beantwortung dieses Fragebogens fertig sind.

<u>Anhang C.17</u>: Protokollbogen für Nachbefragung

Nachbefragung **Vp:_____**

3. **Haben Sie irgendeine Strategie, Taktik oder Technik angewendet, um dem Untersucher die Lügendetektion zu erschweren? (z.B.: Entspannungstechniken, gedankliche Ablenkung, Herbeiführen körperlicher Reaktionen o.ä.)**

nein	ja
0	1

<u>wenn</u> **ja**, <u>welche</u>?

4. **Haben Sie beim Lügen und Wahrheitsagen unterschiedliche Empfindungen, Gefühle, körperliche Reaktionen o.ä. verspürt?**

nein	ja
0	1

<u>wenn</u> **ja**, <u>welche</u>?

5. **Haben Sie Anmerkungen, Kommentare, Kritik und Anregungen zum Experiment? (z.B.: Ist Ihnen etwas angenehm oder unangenehm aufgefallen?)**

<u>Anhang C.18</u>: Protokollbogen zur Erfassung der soziodemographischen Daten

Vp: _____ **Bed.:** _____ **Test:** _____ **Diagnose:** _____

Datum: _____ **ca. Zeit:** _____ **VL 1 (T/M):** _____

Alter: _____ Jahre

Studienfach / Beruf: _____ Semesterzahl: _____

Sehschwächen (farbenblind?): _____ Korrektur (Brille o.ä.): _____

dominante Hand: ❍ Rechtshänder ❍ Linkshänder

Erfahrung mit psychol. Exp. (Häufigkeit d. Teilnahme): _____

Art der Anwerbung (Aushang, Ansprechen etc.): _____

Vorinformationen hins. Experiment: _____

Befinden in der Kabine (Platzangst o.ä.): _____

ggf. Muttersprache: _____

Temp. / Luftfeucht.; Beginn: _____ / _____; Ende: _____ / _____

Anmerkungen: _____

Ruhemessung, Anzahl NSRs: _____

Zahlentest, gewählte Zahl: _____

1.	**Welche Zahl haben Sie gewählt?**	
	War es	die 11?
2.		die 12?
3.		die 13?
4.		die 14?
5.		die 15?
6.		die 16?

<u>Anhang C.19</u>: Protokollbogen für Ablauf DLT

1. UDLT

Vp:_____

	Frage	Antw.		Bemerkungen
1.	Sitzen Sie auf einem Stuhl?	j		
2.	Haben Sie die Schublade Nr. 6 geöffnet?	n		
3.	Sind Sie immer aufrichtig?	n		
4.	Haben Sie die Schublade Nr. 3 geöffnet?	n		
5.	Sind Sie manchmal unaufrichtig?	n		

2. DLT

1. Durchgang

	Frage	Antw.		Bemerkungen
1.	Befinden Sie sich momentan in Mainz?	j		
2.	Haben Sie den Ring gestohlen?	n		
3.	Sind Sie immer ehrlich?	n		
4.	Haben Sie die Kette gestohlen?	n		
5.	Sind Sie manchmal unehrlich?	n		
6.	Befinden Sie sich momentan an der Universität?	j		
7.	Haben Sie den Plastikkasten geöffnet?	n		
8.	Handeln Sie manchmal illegal?	n		
9.	Haben Sie die Glasschale geöffnet?	n		
10.	Handeln Sie stets legal?	n		
11.	Befinden Sie sich momentan am Psycholog. Institut?	j		
12.	Sagen Sie manchmal die Unwahrheit?	n		
13.	Tragen Sie die gestohlene Kette bei sich?	n		
14.	Sagen Sie immer die Wahrheit?	n		
15.	Tragen Sie den gestohlenen Ring bei sich?	n		

2. Durchgang

	Frage	Antw.		Bemerkungen
16.	Befinden Sie sich momentan an der Universität?	j		
17.	Haben Sie die Kette gestohlen?	n		
18.	Handeln Sie stets legal?	n		
19.	Haben Sie den Ring gestohlen?	n		
20.	Handeln Sie manchmal illegal?	n		
21.	Befinden Sie sich momentan am Psycholog. Institut?	j		
22.	Sagen Sie immer die Wahrheit?	n		
23.	Haben Sie die Glasschale geöffnet?	n		
24.	Sagen Sie manchmal die Unwahrheit?	n		
25.	Haben Sie den Plastikkasten geöffnet?	n		
26.	Befinden Sie sich momentan in Mainz?	j		
27.	Sind Sie manchmal unehrlich?	n		
28.	Tragen Sie den gestohlenen Ring bei sich?	n		
29.	Sind Sie immer ehrlich?	n		
30.	Tragen Sie die gestohlene Kette bei sich?	n		

<u>Anhang C.20</u>: Protokollbogen für Ablauf GAT

1. UGAT **Vp:_____**

1.	**In welchem Raum haben Sie den Schmuckdiebstahl begangen?**	
	War es	Raum 02-522?
2.		Raum 02-528?
3.		Raum 02-525?
4.		Raum 02-523?
5.		Raum 02-527?
6.		Raum 02-521?

2. GAT

1.	**Welches Schmuckstück haben Sie gestohlen?**	
	War es	eine Krawattennadel?
2.		eine Uhr?
3.		eine Brosche?
4.		ein Ring?*
5.		ein Armreif?
6.		eine Kette?#
7.	**Aus welcher Schublade haben Sie den Schmuck gestohlen?**	
	War es	die Schublade Nr. 5?
8.		die Schublade Nr. 3?#
9.		die Schublade Nr. 4?
10.		die Schublade Nr. 1?
11.		die Schublade Nr. 6?*
12.		die Schublade Nr. 2?
13.	**In welchem Gefäß befand sich das von Ihnen gestohlene Schmuckstück?**	
	Befand es sich	in einem Porzellanbehälter?
14.		in einer Pappschachtel?
15.		in einem Plastikkasten?#
16.		in einer Holzschatulle?
17.		in einer Metalldose?
18.		in einer Glasschale?*
19.	**Welcher Gegenstand befand sich noch in der Schublade, aus der Sie den Schmuck gestohlen haben?**	
	War es	ein Bleistift?
20.		ein Füller?
21.		ein Locher?*
22.		ein Hefter?#
23.		ein Lineal?
24.		eine Schere?
25.	**Welche Farbe hat der Stein des von Ihnen gestohlenen Schmuckstücks?**	
	Ist der Stein	schwarz?
26.		gelb?#
27.		weiß?
28.		blau?*
29.		grün?
30.		rot?
31.	**Wie lautet der Buchstabe auf dem Etikett des von Ihnen gestohlenen Schmuckstücks?**	
	Ist es	ein Y?
32.		ein E?*
33.		ein U?
34.		ein I?
35.		ein A?#
36.		ein O?

Anm: * = relevantes Item der Bedingung Ring # = relevantes Item der Bedingung Kette

<u>Anhang D</u>: Bilder des Büroraums 02-525 (Tatort der Scheinverbrechen)

Abbildung 39. Schreibtisch mit geöffneten Schubladen.

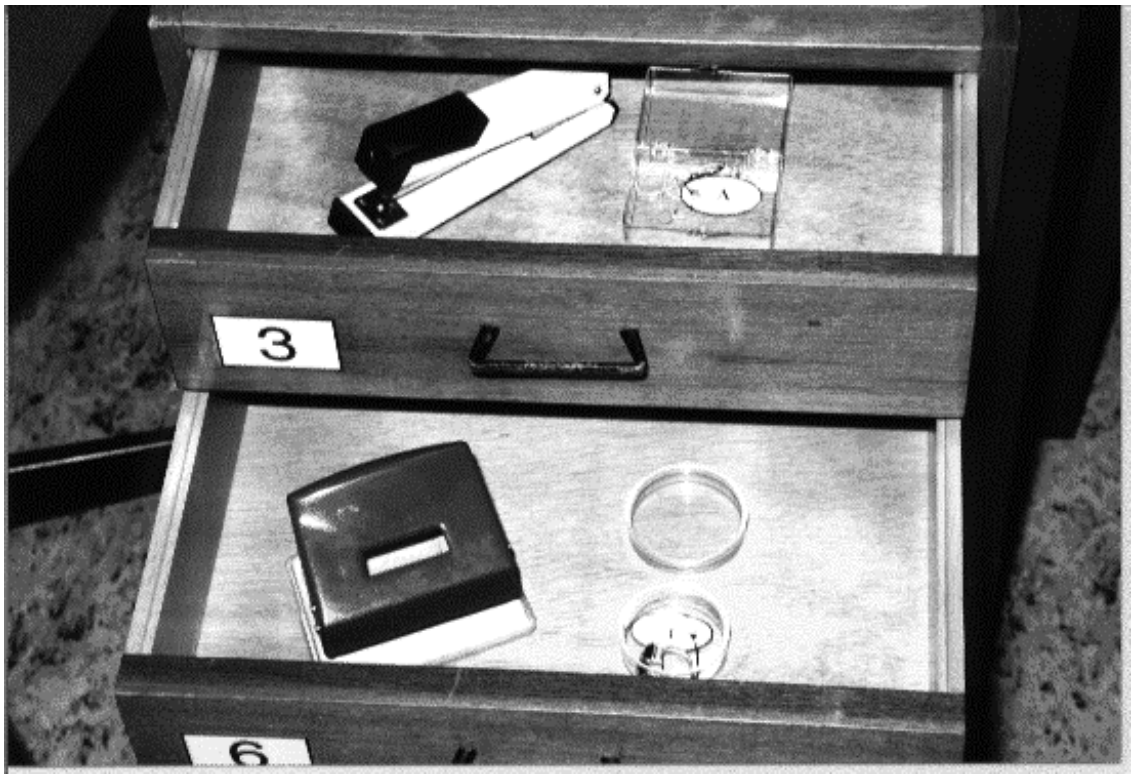


Abbildung 40. Schreibtischschubladen (Nr. 3 und Nr. 6) mit Inhalt: Schmuckgegenstände (Kette und Ring, mit gelbem vs. blauem Stein und Etiketten „A“ vs. „E“), Aufbewahrungsgefäße (Plastikkasten und Glasschale), Büroutensilien (Hefter und Locher).

Anhang E:

Vergleich der Häufigkeitsverteilungen der untransformierten und logarithmierten SCR-Amplituden, getrennt für die drei Auswertungen

SCRs nach Einblendung der Fragen bzw. Items (Latenzfenster 1 – 3 Sekunden)

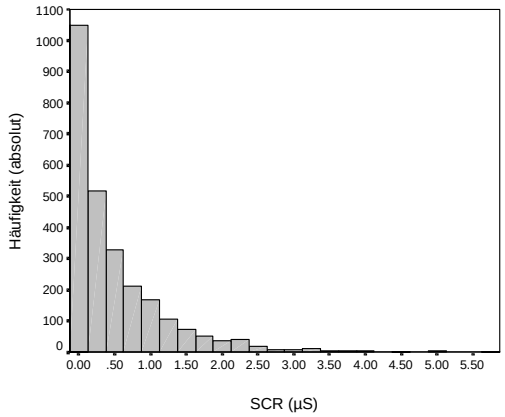


Abbildung 41. SCRs nach Einblendung der Fragen bzw. Items (Latenz 1 – 3 Sekunden): Histogramm der untransformierten Amplitudenwerte in Intervallschritten von 0.25 µS.

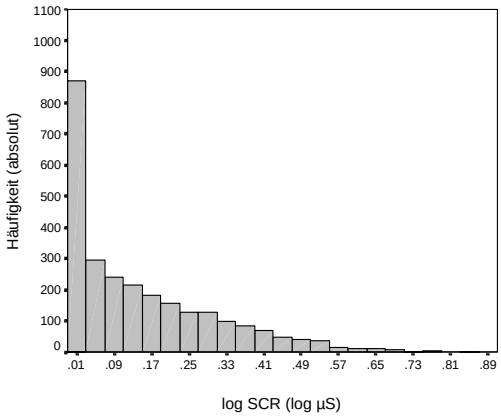


Abbildung 42. SCRs nach Einblendung der Fragen bzw. Items (Latenz 1 – 3 Sekunden): Histogramm der logarithmierten Amplitudenwerte in Intervallschritten von 0.04 log µS.

SCRs nach Einblendung der Fragen bzw. Items (Latenzfenster 1 – 10 Sekunden)

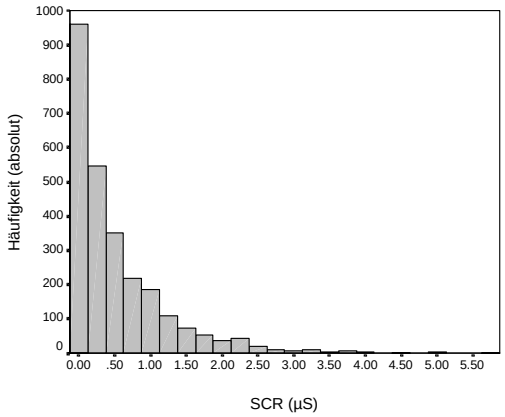


Abbildung 43. SCRs nach Einblendung der Fragen bzw. Items (Latenz 1 – 10 Sekunden): Histogramm der untransformierten Amplitudenwerte in Intervallschritten von 0.25 µS.

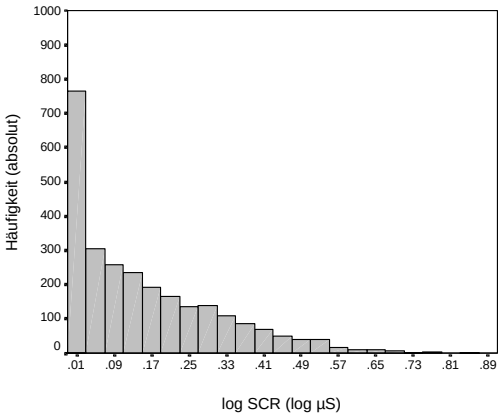


Abbildung 44. SCRs nach Einblendung der Fragen bzw. Items (Latenz 1 – 10 Sekunden): Histogramm der logarithmierten Amplitudenwerte in Intervallschritten von 0.04 log µS.

SCRs nach Ausblendung der Fragen bzw. Items (Latenzfenster 1 – 3 Sekunden)

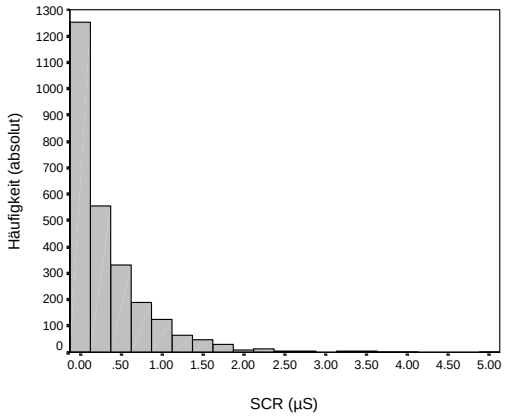


Abbildung 45. SCRs nach Ausblendung der Fragen bzw. Items (Latenz 1 – 3 Sekunden): Histogramm der untransformierten Amplitudenwerte in Intervallschritten von 0.25 µS.

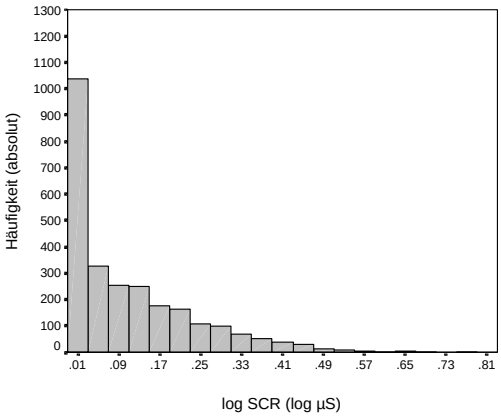


Abbildung 46. SCRs nach Ausblendung der Fragen bzw. Items (Latenz 1 – 3 Sekunden): Histogramm der logarithmierten Amplitudenwerte in Intervallschritten von 0.04 log µS.

Anhang F:

Post-hoc-Analysen der Habituation für die logarithmierten SCRs nach Einblendung (Latenzzeit 1 – 3 Sekunden), getrennt für DLT und GAT

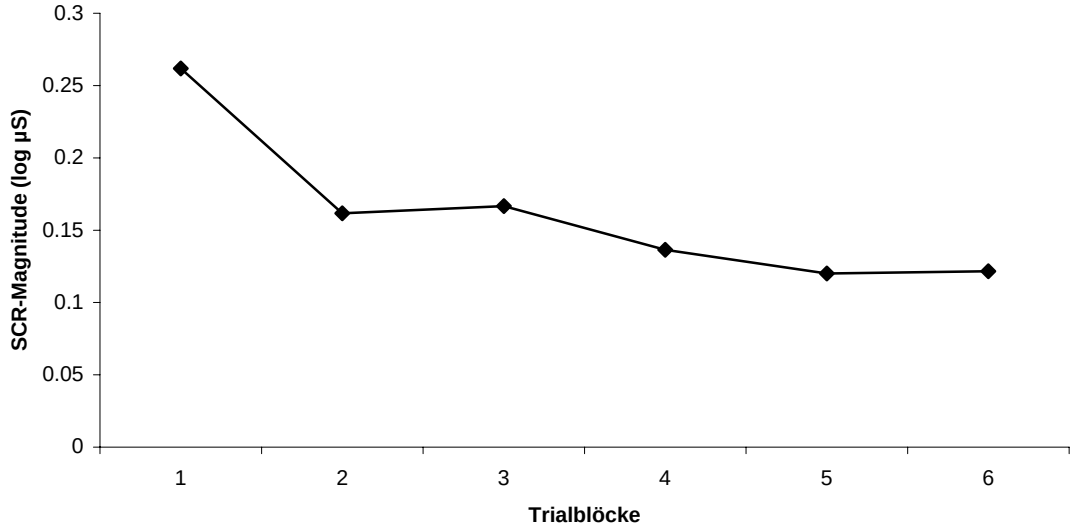


Abbildung 47. DLT – SCRs nach Einblendung der Fragen, Trialblockeffekt: $F(5,195) = 38.90$, $p < .001$, $\varepsilon = .550$, $\eta^2 = .50$; Mittelwerte der Blöcke 1 – 6 in log µS und Standardabweichungen (in Klammern): 0.2618 (0.1279), 0.1617 (0.1222), 0.1666 (0.1156), 0.1364 (0.0990), 0.1201 (0.0931), 0.1216 (0.0968); Trendanalyse: lineare Komponente: $F(1,39) = 66.27$, $p < .001$, $\eta^2 = .63$, quadratische Komponente: $F(1,39) = 22.43$, $p < .001$, $\eta^2 = .37$.

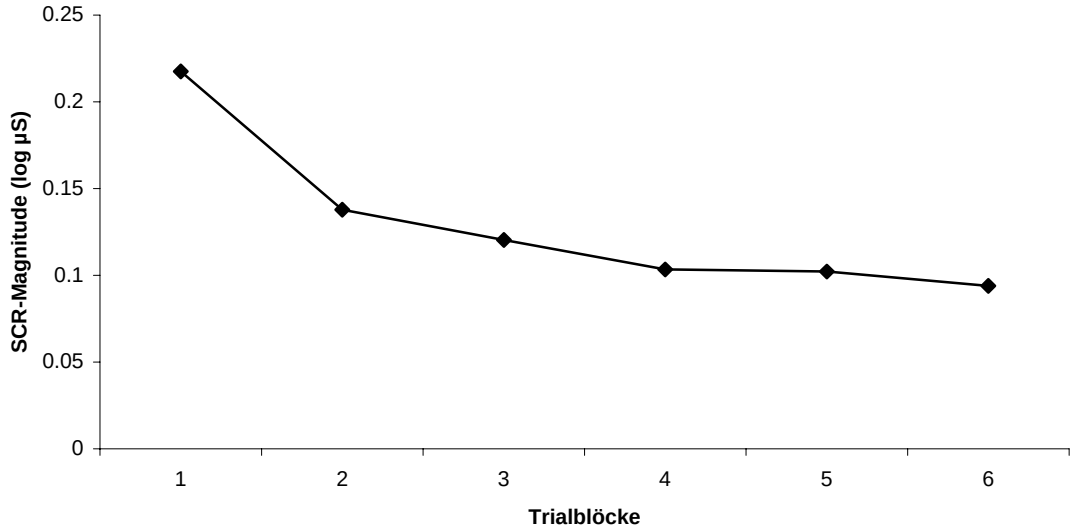


Abbildung 48. GAT – SCRs nach Einblendung der Items, Trialblockeffekt: $F(5,195) = 37.96$, $p < .001$, $\varepsilon = .696$, $\eta^2 = .49$; Mittelwerte der Blöcke 1 – 6 in log µS und Standardabweichungen (in Klammern): 0.2175 (0.1415), 0.1379 (0.1221), 0.1204 (0.1189), 0.1034 (0.1144), 0.1022 (0.1106), 0.0939 (0.0979); Trendanalyse: lineare Komponente: $F(1,39) = 87.86$, $p < .001$, $\eta^2 = .69$, quadratische Komponente: $F(1,39) = 38.78$, $p < .001$, $\eta^2 = .50$.

Abstract

Dieses Experiment untersuchte die Effekte der unterschiedlichen Fragen- bzw. Itemtypen (Stimuli), des Wahrheitsgehalts der Antworten und der elektrodermalen Labilität auf die Hautleitfähigkeitsreaktionen (SCR) und die phasische Herzschlagfrequenz (HR) für zwei relativ neue Befragungstechniken der forensischen Psychophysiologie („Lügendetektion“): Directed Lie Test (DLT) und Guilty Actions Test (GAT).

Achtzig Männer begingen einen simulierten Schmuckdiebstahl. Jeweils die Hälfte entwendete entweder einen Ring oder eine Kette. Während dieser Tat wurden jedoch alle Probanden mit den kritischen Details beider Scheinverbrechen konfrontiert. Anschließend absolvierten sie entweder einen DLT oder einen GAT. Die relevanten Stimuli der Tests bezogen sich auf beide Scheinverbrechen und wurden – intraindividuell variiert – wahrheitswidrig und wahrheitsgemäß verneint. Darüber hinaus umfaßte der DLT inhaltlich parallelisierte Paare von Kontrollfragen. Die Probanden wurden instruiert, die jeweiligen Kontrollfragen eines Paars wahrheitswidrig versus wahrheitsgemäß zu verneinen. Die Testverfahren beinhalteten außerdem nicht tatbezogene, irrelevante Stimuli, die wahrheitsgemäß beantwortet wurden.

Für beide Befragungstechniken fand man Reaktionsunterschiede zwischen den Stimulustypen, insbesondere stärkere SCR-Magnituden und eine niedrigere HR auf die wahrheitswidrig verneinten relevanten Stimuli. Bei den Kontrollfragen des DLT zeigten sich jedoch keine signifikanten Effekte des Wahrheitsgehalts. Die elektrodermale Labilität hatte keinen bedeutsamen Einfluß auf die Reaktionsunterschiede. Die Ergebnisse wurden vor allem anhand psychophysiologischer Theorien der Aufmerksamkeit, Konflikte und Informationsverarbeitung interpretiert.